nature portfolio

Peer Review File

Comprehensive Echocardiogram Evaluation with View Primed Vision Language AI

Corresponding Author: Dr David Ouyang

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Attachments originally included by the reviewers as part of their assessment can be found at the end of this file.

To help resolve the divergence of opinion, referee #2 was consulted and found the authors' response to reviewer #1 satisfactory.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
The manuscript introduces EchoPrime, a multi-task model for echocardiography that claims state-of-the-art performance across 17 binary and 2 continuous tasks, utilizing a contrastive encoder and anatomical attention module. I have some comments following the points below:

1 - Clinical Utility and Workflow Integration: EchoPrime's precision improves with additional views and videos (Figure 2A), indicating potential value as a diagnostic support tool. However, this approach relies on high-quality, multi-view echocardiographic data, which requires skilled sonographers or physicians. This dependency may limit its feasibility in resource-constrained settings. The model's focus on classification tasks (rather than regression tasks beyond EF estimation) could also restrict its usefulness for comprehensive echocardiography, which depends on detailed measurements and morphological assessments essential for patient care. If EchoPrime is intended for high-pre-test probability populations, its added value is unclear since comprehensive studies and specialist reviews are already standard. Alternatively, if the goal is to "clear" low-risk patients, its negative predictive value (NPV) should be compared to simpler and more accessible tools. For instance, Elias et al.'s ECG-based model detects aortic valve regurgitation with an AUC of 0.84 (JACC, 2022), close to EchoPrime's AUC of 0.88. Given that ECG is less resource-intensive and more widely available, the small AUC improvement raises questions about EchoPrime's cost-effectiveness.

The authors should clarify EchoPrime's NPV compared to existing tools in low- and intermediate-pre-test populations, its advantages in high-pre-test settings where comprehensive studies are already performed, and how it accommodates the detailed assessments needed in echocardiography. These points are essential to evaluate its clinical value and potential integration into workflows.

2 - Observed Discrepancy in Outputs: Testing the model on my MacBook Pro (macOS Apple M3 Max, 36 GB) with a cardiac ultrasound not provided in the GitHub repository, I found a discrepancy between the EF reported in the generated text (55%) and the EF derived from output logits (61%) using the provided Jupyter Notebook. This inconsistency raises questions about which value should be considered accurate and whether outputs are standardized across formats. It would be helpful for the authors to explain the rationale behind this discrepancy, such as whether it reflects differences in post-processing methods or thresholds used for generating outputs.

Report: 'Left Ventricle: Normal left ventricular size Normal left ventricular systolic function. LV Ejection Fraction is 55 % Normal left ventricular wall thickness. Mild diastolic dysfunction. There is reversal of the E to A ratio and/or prolonged deceleration time consistent with impaired left ventricular relaxation. [SEP]Resting Segmental Wall Motion Analysis: Total wall motion score is 1.00. There are no regional wall motion abnormalities [SEP]Right Ventricle: Normal right ventricular size and systolic function. Normal right ventricular size. [SEP]Left Atrium: Normal left atrial size and morphology. The left

atrium is normal in size. [SEP]Right Atrium: Normal right atrial size and morphology. [SEP]Atrial Septum: The interatrial septum is normal in appearance. [SEP]Mitral Valve: The mitral valve demonstrates normal leaflet morphology. The mitral valve demonstrates normal function with trace physiologic regurgitation. Mild MV leaflet calcification. [SEP]Aortic Valve: Normal appearance and function of the aortic valve. No significant aortic stenosis or insufficiency. [SEP]Tricuspid Valve: Normal appearance of the tricuspid valve. Est RV/RA pressure gradient is 12 mmHg There is trivial tricuspid regurgitation. [SEP]Pulmonic Valve: Pulmonic valve not well visualized. [SEP]Pericardium: Normal pericardium with no pericardial effusion. [SEP]Aorta: Normal aortic root. [SEP]IVC: The IVC diameter is 1.34 mm [SEP]Pulmonary Artery: Normal pulmonary artery size. [SEP]Pulmonary Veins: Pulmonary veins are normal in appearance and pulse Doppler interrogation shows normal systolic predominant flow. [SEP]'

Feature logits
{'impella': np.float64(0.0),
'ejection_fraction': np.float64(61.11363636363637),
'wall_motion_hypokinesis': np.float64(0.04),
'pacemaker': np.float64(0.0),
'rv_systolic_function_depressed': np.float64(0.0),
'right_ventricle_dilation': np.float64(0.0),
'left_atrium_dilation': np.float64(0.0),
'right_atrium_dilation': np.float64(0.0),
'atrial_septum_hypertrophy': np.float64(0.0),
'mitraclip': np.float64(0.0),
'mitral_annular_calcification': np.float64(0.0),
'mitral_stenosis': np.float64(0.0),
'mitral_regurgitation': np.float64(0.0),
'tavr': np.float64(0.0),
'bicuspid_aov_morphology': np.float64(0.0),
'aortic_stenosis': np.float64(0.0),
'aortic_regurgitation': np.float64(0.0),
'tricuspid_stenosis': np.float64(0.0),
'tricuspid_valve_regurgitation': np.float64(0.0),
'pulmonic_valve_regurgitation': np.float64(0.0),
'pericardial_effusion': np.float64(0.0),
'aortic_root_dilation': np.float64(0.0),
'dilated_ivc': np.float64(0.02),
'pulmonary_artery_pressure_continuous': np.float64(26.0),
'elevated_left_atrial_pressure': np.float64(0.0)}

3 - While the manuscript does not explicitly discuss pressure estimations such as right atrial pressure (RAP), testing the software using the provided Jupyter Notebook generated a text report that included an estimated RAP value (12 mmHg, see above). This raises questions about whether such estimations are validated or reliable. Are these pressure values directly derived from model predictions, and if so, have they been validated?

4 - Benchmarking Against Comprehensive Tools: The manuscript provides a comparison with EchoCLIP and BioMedCLIP. However, EchoPrime's focus on binary classifications raises questions about its clinical utility relative to commercially available tools like US2AI (https://us2.ai/product/), which deliver fully automated, multi-view quantitative echocardiography reports. US2AI includes continuous outputs including LV/RV morphology, and diastolic function, which are critical for standard echocardiographic reporting. While the lack of access to proprietary systems like US2AI is understandable, at least a discussion of how EchoPrime compares to or complements such tools would help articulate its standalone value and potential niche in clinical workflows.

5 - Statistical and Clinical Significance: The manuscript highlights small percentage improvements in performance over prior models, such as a 2% increase in AUC for tricuspid regurgitation. However, it is unclear whether these improvements are statistically significant or clinically meaningful. Without p-values or confidence intervals (at least I did not find them), the reported gains could reflect expected variability rather than true advancements.

6 - Performance in Challenging Imaging Conditions: the manuscript does not address EchoPrime's performance under suboptimal imaging conditions, such as poor acoustic windows, which are common in real-world practice. Understanding its robustness in these scenarios is crucial for assessing its clinical applicability. I believe including an analysis of performance stratified by image quality would enhance the model's credibility for diverse clinical settings.

7 - Implementation and User-Friendliness: running the software on a MacBook using a CPU backend required adjustments. To improve user-friendliness and ensure reproducibility, the README should specify exact requirements for setting up the virtual environment. Providing a Dockerized version would ensure cross-platform compatibility and consistency. Currently, it is unclear whether the exact Python modules and dependencies used align with those intended by the authors, potentially impacting reproducibility and performance.

8 - Non-Echocardiographic Disease Predictions (Minor): the manuscript refers to STEMI and cardiac amyloidosis as "non-echocardiographic primary diseases," which may not fully capture their diagnostic pathways. Echocardiography is commonly used to diagnose cardiac amyloidosis, offering structural and functional insights, and may play a significant role

in assessing cardiac function in STEMI patients. Clarifying the term “non-echocardiographic traits” would prevent potential misunderstandings and better align the claims with established clinical practices.

(Remarks on code availability)
Please consider my comments above.

Referee #2

(Remarks to the Author)
Congratulations to the authors as you continue your progressive refinement of the use of AI in echocardiography. Expanding from your recent foundational interpretive model EchoCLIP, which was trained on one million echo video clips, you have now used over 12 million echo clips (essentially all echocardiograms done at Cedars Sinai over a decade, along with their de-identified echo reports) to train a successor foundation model (EchoPrime), which has many advantages over its predecessor. Among the advances in modeling is contrastive learning to form a joint image-text representation, a more sophisticated view classifier that allows an anatomic attention model to understand which of the 58 characterized views are most likely to show the finding of current interest, and use of the full video clip for assessment rather than individual frames. The model was tested on hold-out sets of 4044 total echo studies from both Cedars Sinai as well as Stanford. EchoPrime claims the ability to predict the presence of hundreds of pathologies encountered over the years but was tested on 19 fairly common features encapsuling chamber size and function and valve stenosis and regurgitation as well as implanted devices (e.g., prosthetic valves, MitraClip, and pacers). Two parameters (LV ejection fraction and RV systolic pressure) were compared to the clinical report by linear regression (r^2 and mean average error), while the other 17 were tested with AUROC for binary variables, significantly outperforming EchoCLIP, the more generalized foundation model BioMedCLIP, and several task specific models.

One important echo capability not well demonstrated was detection and quantitation of regional wall motion abnormalities. It is unclear if EchoPrime is not capable of doing this natively or if it just wasn’t part of the test suite but should be mentioned in either case. This becomes quite relevant in their auxiliary metric of detecting STEMI (which probably should be spelled out at first mention for the medically uninitiated). It is unclear just what you mean by STEMI; is it an acute infarct or does a chronic regional wall motion abnormality count? Does the model understand the connection between wall motion and coronary territories? I’m also confused by predicate for the STEMI (and amyloidosis), where you state that these are diagnoses “not typically diagnosed with echocardiography,” as I would consider both of these to be “bread and butter” echo diagnoses. It also appears that you only considered the Ap4 view for STEMI, which would miss many RCA and LCx infarcts. Similarly, your focus on amyloid uses only the PLAx view, which would miss many of the key findings in amyloid (mid-septal stripe and apical strain sparing, for example).
This model clearly is an important advance in AI-guided echo interpretation and, after appropriate revision, quite worthy of publication. One possible limitation is that the model does not actually write the interpretation language but, as I understand it, rather finds the best fitting verbiage from the 275K reports generated at Cedars Sinai over the last ten years (admittedly, these will cover most every nuance!).

Another critique is the choice of batch size = 32. The key point of contrastive learning is to bring like embeddings (text reports and their corresponding videos, i.e., the positive example) closer together while pushing negative pairs far apart. This necessitates a sufficiently large batch size to create enough differentiation within the latent embedding space. You trained on only 2 A6000's, raising the possibility of a limitation of computing power, which would warrant some further discussion and ideally ablation studies showing the impact of different batch sizes.

A potential critique of multiple instance learning is the risk that your inference pipeline is over-engineered, carrying strong inductive bias with a classifier and the weighting objective. Have you experimented at all with learning the anatomical priors outside of a two-step system, i.e. a learned pooling of embeddings to create the study and anatomical embeddings?

Retrieval Augmented Interpretation (page 14, par 2): I'm not sure why this was described uniquely in comparison to Retrieval Augmented Generation. If Retrieval Augmented Interpretation has been described elsewhere this should be cited in this section. Otherwise, Retrieval Augmented Interpretation seems to describe how CLIP models are used as zero shot __ (where __ is a regression task, classifier, or 'report generator'). Probe text is encoded using the text encoder and a similarity metric is used to retrieve the most relevant text based on the similarity between the video and text encodings. So, a description of the 'historical' reports that are retrieved is necessary here. Are these reports all from the training set or a different set of reports? In addition, you average 50 retrieved reports to generate the interpretation. Evidence for k=50 reports being optimal? Also, what is meant by averaging since this is at the text level?

Along the same lines there is no discussion as to why mViT was chosen as the video encoder and no ablations for other video encoding architectures. Additionally, the same can be said about using PubMedBERT for the text encoder. In fact, in the previous iteration, EchoCLIP dedicated significant discussion space to describing a unique tokenization technique that took advantage of the templated structure of Echo reports. This was shown to be better than standard Byte Pair Encoding with the CLIP text encoder. I think the switch to PubMedBERT and Wordpiece encoding warrants deeper discussion and some comparison studies with the EchoCLIP text encoder and the text encoder used in EchoPrime.

I would note that there is break in the text continuity between page 5 and 6. It is unclear if there is any lost text here (I suspect not), but it should be corrected.

The authors have provided model code with all initialization parameters at their github. While I have not implemented this for

my own testing, it does appear that it is in standard form and should be readily implemented with the appropriate subroutine libraries, which they specified. In addition, you provide a video example of EchoPrime in operation, interpreting a relatively normal echo. This overall is effective in showing model operation, but I am puzzled by the lack of any spectral Doppler frames in the example data. It is simply impossible to estimate RV pressure, diastolic function, or valve stenosis without Doppler, so this should be redone with the appropriate needed images. It should be noted that you have not provided the echo training data, but this seems reasonable. Not only would this likely exceed a petabyte in size, but despite all best efforts to strip identifying information from the images, such PHI might still be present. Of note, you do provide a sample casein DICOM format, though again I only see video clips, no M-mode nor spectral Doppler. Unclear if this is the same case shown in the model operation example. You certainly need to include the single frame images and specify if this is the case shown in the MOV file.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
A. Summary of Key Results
The authors propose a novel vision-language model trained on 12 million samples to interpret echocardiography studies by performing view classification, mapping videos into a video-text latent space, and retrieving study interpretations. Experiments on both internal and external datasets demonstrate that EchoPrime outperforms both task-specific approaches and prior foundation models.

B. Originality and Significance
The authors introduce EchoPrime, which incorporates multi-view information to enhance the accuracy of echocardiography interpretations—an innovative approach. Additionally, they collected an extensive dataset of 12 million samples, highlighting the scale of their work. However, although the authors claim EchoPrime can assist physicians in the automated preliminary assessment of comprehensive echocardiography after rigorous clinical evaluation, no human evaluation has been conducted; the assessment relies solely on metrics such as AUC, limiting the study's significance.

While the dataset used is comprehensive, the analysis is purely retrospective. Prospective experiments would align more directly with clinical applications, providing a more robust validation of EchoPrime's effectiveness in real-world clinical settings. The lack of prospective analysis significantly limits the study's overall validity and impact.

C. Data & Methodology
Video Encoder: The authors state that their video encoder is trained on a mViT checkpoint; however, mViT dates back to 2021. Recent advancements such as BEVT [1] and VideoMAE [2] could enhance relevance and demonstrate a thorough exploration of state-of-the-art methods.

Text Encoder: EchoPrime supports over 500 tokens and utilizes WordPiece encoding within a BERT architecture. The authors state, "This marks a significant improvement over previous medical foundation models like BioMedCLIP and EchoCLIP, which process only single-view, individual images and handle text up to 77 tokens." However, no ablation studies on token length are provided, which are necessary to validate the benefit of increasing the token limit. Comparisons with other models alone do not substantiate that this design choice is significantly advantageous.

Method: The evaluation of EchoPrime lacks comprehensiveness. While the authors selected two foundation models, BioMedCLIP and EchoCLIP, additional comparisons, such as with PubMedCLIP [3], PMC-CLIP [4], and ROCO [5], would provide a broader context.

Data: For the view classification task, the origin of the training dataset is unclear. Is this dataset part of the 12,124,168 echocardiography videos from CSMC as listed in Table 1? If not, please specify the data source and provide statistical details for this specific dataset.

Presentation: The metric represented in Figure 2b is not specified. Including a description in the figure caption would clarify this. Furthermore, providing figures for data statistics would enhance understanding of the dataset's distribution and characteristics.

[1] R. Wang et al., "BEVT: BERT Pretraining of Video Transformers," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 14713-14723, doi: 10.1109/CVPR52688.2022.01432.
[2] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. ArXiv, abs/2203.12602.
[3] Eslami, Sedigheh, Christoph Meinel, and Gerard De Melo. "PubMedCLIP: How Much Does CLIP Benefit Visual Question Answering in the Medical Domain?" Findings of the Association for Computational Linguistics: EACL 2023. 2023.
[4] Lin, Weixiong, et al. "PMC-CLIP: Contrastive Language-Image Pre-Training Using Biomedical Documents." arXiv preprint arXiv:2303.07240 (2023).
[5] Pelka, Obioma, et al. "Radiology Objects in Context (ROCO): A Multimodal Image Dataset." Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint

International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3. Springer International Publishing, 2018.

D. Use of Statistics and Treatment of Uncertainties
The authors claim that EchoPrime matches or exceeds the performance of task-specific models, but they have not employed statistical methods to verify whether EchoPrime's results differ significantly from those of task-specific models. The lack of statistical measures in the paper is a major limitation. Including confidence intervals for the results is strongly recommended to demonstrate the robustness of the performance comparisons.

E. Conclusions: Robustness, Validity, Reliability
While the model demonstrates promising results, the robustness of these findings is limited by the absence of human evaluation, lack of prospective analysis, and exclusive reliance on quantitative metrics such as AUC. Additionally, the absence of statistical measures (e.g., confidence intervals) weakens the reliability of the conclusions. Inclusion of additional comparison methods such as PubMedCLIP [3], PMC-CLIP [4], and ROCO [5], or at least reference to them, would improve the comprehensiveness and validity of the evaluation.

F. Suggested Improvements
1. To enhance scientific rigor and comprehensiveness, additional experiments and comparisons with other state-of-the-art methods would solidify the model's relative performance.
2. Conducting ablation studies on token length used in the text encoder would empirically demonstrate the value of supporting over 500 tokens, clarifying whether this design choice significantly improves performance.
3. Providing details on the source and structure of the view classification training dataset, particularly if it differs from the primary dataset, is essential to support the reproducibility of the work.
4. Incorporating statistical analysis would offer a more reliable benchmark for EchoPrime's capabilities in automated echocardiography interpretation.

(Remarks on code availability)
It is not very clear about the checkpoint of the model in the Github repo.

Version 2:

Reviewer comments:

Referee #1

(Remarks to the Author)
Dear Authors,

I appreciate your efforts in revising the manuscript and your detailed rebuttal. Here are my point-by-point comments that, to my view, are not fully addressed.

1 - In your rebuttal, you stated that comparing EchoPrime to an ECG-based AI model (AI-ECG) constitutes an "apples to oranges" comparison. My intention was not to equate ECG and echocardiography purely on technical grounds, but to emphasize that both might be used as screening or triage tools. ECG is often the first-line test in nearly every clinical setting, from outpatient clinics to the ICU, and an AI-based approach with high negative predictive value (NPV) can rule out certain pathologies. If EchoPrime's binary classification tasks can achieve similarly high NPV, I was suggesting it could be valuable in a similar triage context. My question remains: if EchoPrime is not intended for screening in a low pre-test probability population, then what distinct advantage does it offer in high pretest-probability (in-hospital) settings, especially if there is already a comprehensive in-person exam?

2 - I note that you have provided Mean Absolute Error (MAE) and R^2 for multiple regression tasks, but these two metrics, as they stand (which units in the MAEs? Clinical relevance?), do not fully capture the model's performance. How do we know how it behaves throughout the whole range? I suggest to address this question by also using Bland–Altman plots to analyse possible biases and agreement across the full range of measurements. For example, an MAE of ~7 mmHg for pulmonary pressures could be clinically significant or not, depending on the range of values and whether the errors skew low or high. Without Bland–Altman or explicit discussion of bias over the entire distribution, interpreting such differences remains difficult. I would also suggest to include the units to which the MAEs report to, otherwise in the metrics regarding to morphology, it could mean cms, or mm, or anything else, which makes it difficult to assess the magnitude of the findings. Additionally, you mention internal held-out sets (by the second supplementary table 7 that is provided in your rebuttal letter), yet it's unclear how consistent the regression performance is across other external cohorts. It would help to see these continuous metrics (beyond EF) over the other multiple sites explicitly. If you have performed regression analyses externally (e.g., for pulmonary pressures, aortic gradients, etc.), sharing those results—ideally with Bland–Altman plots or at least a more granular view—would greatly clarify how EchoPrime's continuous outputs hold up in diverse settings.

2b. In your rebuttal, you explain that the 55% EF is an "intermediate" output (closest single retrieved report), whereas 61% is the final aggregated EF from the top 50 retrieved reports. While that clarifies your pipeline, if both outputs are visible in the code or user-facing tool, users may be unsure which number is the "true" result. If users are to rely on EchoPrime, one would expect only the final aggregated EF (61%)—or some single definitive number—be displayed to avoid confusion. It would be

helpful if you could clarify how such intermediate results are handled in any final production version, and whether the user-facing EF is always the aggregated figure.

3. You argue EchoPrime is primarily for assisting cardiologists and sonographers by automatically generating preliminary reports—particularly in a higher pretest probability setting. My question is how you envision these automatically regressed measurements being used when some R^2 values are relatively low, the model's error margins remain unclear (units, clinical significance, how the model behaves throughout the whole range and in external cohorts). With this in mind, are users expected to trust or override these measurements? If they must be manually re-checked anyway, what exactly is the efficiency gain? Also, if the model excels at classification (binary predictions), it might have value in ruling out significant abnormalities in a broader population. Yet you state explicitly it is not for "clearing" low-risk patients. This leaves me uncertain how you see these two sets of outputs (classifications vs. continuous measurements) integrating into a real workflow.

4. I noted that US2AI can provide robust quantitative measurements—often recognized as the more challenging aspect of echo interpretation. In the rebuttal, you characterize US2AI as essentially segmentation-focused with minimal classification. However, many of these commercial tools produce key quantitative outputs (dimensions, strain, EF, volumes, etc.) that can then be thresholded or automatically labeled. If US2AI and similar solutions reliably measure the same metrics EchoPrime is attempting to regress, I believe it is still relevant to discuss whether EchoPrime's performance is near, below, or above that commercial standard—even if direct side-by-side comparisons are not feasible. Indeed, while US2AI may be a "black box," that does not preclude testing both EchoPrime and US2AI on the same de-identified dataset (with ground-truth labels) to compare final metrics such as EF, gradients, and volumes.

Furthermore, you highlight EchoPrime's ability to produce qualitative classification findings, whereas US2AI allegedly does not. However, in an actual echo lab, measured parameters (e.g., EF, volume, valve gradient) are regularly binned or thresholded to yield clinical categories ("mild," "moderate," etc.). Thus, the distinction between "quantitative segmentation" vs. "qualitative classification" can be blurred in real-world practice. In this sense, more clarity on how EchoPrime's performance on these key quantitative parameters compares to or complements existing commercial tools would help readers appreciate the added value.

5. In your rebuttal letter, you mention that EchoPrime "streamlines workflow" and "reduces inter-physician variability." These are potentially large claims. Yet I see no user-study data or operational metrics documenting time-savings, nor do I see statistics on variance reduction between cardiologists when aided by EchoPrime. In many echo labs, we already have standardized "template" reports with placeholders for key measurements. Merely producing "1000 possible statements" from a retrieval-based model may not, on its own, guarantee a more efficient or consistent process—particularly if measured values still require manual over-read or if some statements are clinically extraneous. My suggestion is, if you have any preliminary user feedback, pilot data, or even a small prospective trial that suggests the model truly improves turnaround time or consistency, that would be hugely beneficial to strengthen your claims of workflow improvement.

6. While I appreciate that the retrieval-based approach can generate a wide array of textual phrases, I believe most clinicians are far more concerned with core structured findings (e.g., EF, diastolic function, valve gradients, presence/absence of regurgitation). Without robust external validation of these measured or classified parameters—or a demonstrated added value of these 1000 phrases—the mere variety of textual statements is unlikely to impress. Many clinical settings already use semi-structured "auto-text" with blank fields. Thus, clarifying the true novelty in actual practice would strengthen the paper.

7. In your rebuttal to Reviewer #3, you provided a sample echocardiography report in which EchoPrime outputs statements describing "Mild diastolic dysfunction" (e.g., mentioning E-to-A reversal, prolonged deceleration time). However, it is not clear whether diastolic function was explicitly benchmarked among the model's validated outputs. If not, how should the user interpret these automatically generated statements on diastolic dysfunction? Do you have internal metrics or pilot data indicating that EchoPrime can reliably detect and characterize diastolic abnormalities, or are these statements purely illustrative at this stage? By clarifying this, readers can better assess the reliability of EchoPrime's commentary on diastolic parameters and understand whether these reported findings have been validated in a rigorous clinical or quantitative manner.

I want to emphasize that I do see potential in EchoPrime, particularly given the scale of training data and the comprehensiveness of tasks. My comments aim to ensure clarity and clinical credibility.
Thank you for the opportunity to review and for your time in addressing these concerns.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
Congratulations for a significantly improved manuscript. You have expanded your testing cohort to include two more clinical sites (including one from Taiwan with somewhat different echo standards than American academic labs). You have added several new metrics for both categorical and parametric variables, which shows both the versatility and limitation of your

approach. One limitation that is not evident on a casual reading of the paper is that you did not include in your model training any M-mode or spectral Doppler (PW & CW), which are essential to modern echocardiography (OK, maybe M-mode is a bit of a throw-back, but you can't quantify aortic stenosis without Doppler). This results in the fairly dismal results in Supplementary Table 7 for tasks like transaortic gradient, TR gradient, etc. This limitation needs to be stated more explicitly in the paper and its limitations with a discussion of how spectral Doppler might be included in the future.

In addition to these general comments, I have a few specific issues for you to consider.

Line 168: 512 tokens are indeed a significant improvement over the 77 in EchoCLIP and BioMedCLIP, but it remains a limitation when representing complex reports that may have more than 400 words and would contain the most detailed (and perhaps most useful to the model) RAI information. Is there a way to expand to a greater token length than the current 512.

Line 172-174: A typo that should be corrected.

Line 174-177: If I am reading this correctly, these are quite impressive results. I'm particularly struck that you could uniquely (at least top-10) find an essentially normal report (and echo going the other direction), as there must be 10's of thousands of normal studies that look similar to each other. Did the 97-98% success extend to the many nearly identical normal echoes?

Thank you for confirming that you have not used spectral Doppler or M-mode frames in the training or testing phases. You argue that much of their information can be obtained from 2-D clips (for M-mode) and color Doppler clips (for spectral Doppler data, PW more than CW, given the impossibility of dealiasing the color Doppler for high-speed jets). However, this limitation does deserve explicit mention in the methods and limitations. It does help explain why your accuracy for RVSP was so much worse than EF, for example. Do you have plans to integrate these modalities into your model in future iterations?

There seems to be an error with the description of the contrastive learning set up. You say all clip-report pairs are from different studies. This would lead to non-informative training as all of the pairs would be different or 0 in identity matrix. To learn joint embeddings like this some of the pairs should be matching (ie, from same study) which is the typical set up for this. They need to clarify as this might just be an error.

The self-supervised way that you produce anatomical attention is actually very clever. Once you have trained an image and text model that share the same embedding space, you can then look at individual sections of the text and see if which views are most "important" for predicting that particular section. You might expand a bit on this in the discussion, which really is quite brief given all the innovative techniques introduced here and doesn't really put EchoPrime into context regarding previous attempts. In addition, as you discuss how this might be applied in resource limited settings, a reference to AI methods for echo guidance would be appropriate.

I am still a bit confused as to how the individual level predictions were made and evaluated by EchoPrime. I understand that 50 most similar reports from training set were retrieved and it seems that the embeddings (or "features" – what features specifically?) from these reports were then averaged. How did you go from this to individual level predictions or even an entire report generation (also not clear which of these they actually did). Essentially, what is the actual output of this model?

I do have concerns with the "retrieval augmented interpretation" approach and how this might work on rare findings that might not be highly represented in the training dataset. If the findings are not well represented, then you won't be able to retrieve enough relevant reports for averaging the features to generate the correct output. For example, can your model recognize something like (for example) Ebstein's anomaly or congenitally corrected TGA? Also, as novel echo methods evolve, will this need to be completely retrained rather than just fine tuned. The whole concept of a "foundational model" is that it can be used and adapted for many downstream tasks, including fine-tuning of the model. The problem I see here is that the "retrieval augmented interpretation" framework does not allow for an obvious approach to fine tuning, as retrieval is happening with the original training approach. If you could detail how this could be used downstream for fine-tuning, it would be helpful.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
The authors did not address several concerns I raised in the first round. I do not agree with their responses, which rely on discussion rather than real experiments or clinical evaluations conducted by human physicians.

Although the authors claim that EchoPrime can assist physicians in the automated preliminary assessment of comprehensive echocardiography following rigorous clinical validation, no human evaluation has been performed. The current assessment relies solely on quantitative metrics such as AUC, which limits the study's clinical relevance. Despite the dataset being comprehensive, the analysis remains entirely retrospective. Prospective experiments would provide stronger evidence of EchoPrime's applicability in real clinical settings and would offer more convincing validation of its utility. The absence of such analysis significantly undermines the study's validity and potential impact.

Video Encoder: The authors note that their video encoder is based on a mViT checkpoint. However, mViT was introduced in 2021, and more recent methods such as BEVT [1] and VideoMAE [2] offer notable improvements. Incorporating or comparing against these recent models would strengthen the technical contribution and demonstrate a more thorough engagement with current developments.

(Remarks on code availability)

Version 4:

Reviewer comments:

Referee #1

(Remarks to the Author)
I appreciate the substantial revisions, particularly the additional continuous-metric analyses:

a) for aortic-root dimensions the updated site-specific MAE and R^2 values are now reported (Supplementary Table 8);
b) for diastolic function you have added an explicit AUROC row (Table 2);
c) the new Bland-Altman plots (Supplementary Fig. 5) are helpful, but visual inspection reveals residual proportional bias and heteroscedasticity in several tasks (e.g., pulmonary-artery pressure, trans-aortic velocity/gradient, TR gradient, IVC diameter). Differences become increasingly negative at higher true values (under-prediction) and the error spread widens, particularly for trans-aortic gradient. In addition, centre-specific colour-coding suggests distribution shifts across sites (e.g., IVC diameter). Please quantify these effects (slope ± 95 % CI of difference vs. mean and limits-of-agreement by value decile), report whether slopes differ from zero, and clarify whether any statistical recalibration (overall or site-specific) has been applied or is planned. If proportional bias is significant, a simple linear-correction table or isotonic recalibration would let readers judge the clinical impact.

My remaining questions centre on the incremental value of EchoPrime in light of PanEcho (JAMA 2025):

1. Benchmarking vs. PanEcho: Please provide a direct side-by-side comparison (classification AUROC, regression MAE/R^2, external-site count, compute) so readers can judge whether EchoPrime offers measurable performance gains or complementary tasks.

2. Free-text generation as an additive value:
Current echo workflows already use structured “pre-text” templates into which sonographers or fellows drop quantitative values. The manuscript argues that only ~¼–⅓ of a report is numeric and the remainder is narrative, but the added clinical utility of automatically generated prose remains hypothetical:
– Could you quantify editing time, keystrokes or click reduction for EchoPrime-drafted reports versus a standard template populated with EchoPrime (or EchoNet-Measurements) numbers? Even a small pilot (e.g., first 100 cases) would help substantiate the claim.
– Please clarify what safeguards exist against hallucination. Given the acknowledged difficulty of validating more than 1 000 narrative phrases, how will users be alerted if the model outputs text that is not factually supported? For example, could common phenotypic descriptors—such as “granular sparkling,” “features suggestive of ARVC,” or “hyper-trabeculation” appear in the report? If so, what mechanisms ensure that such statements are appropriate for the current study and not hallucinated? Are these phrases constrained by explicit rule sets or diagnostic-criteria checks, or subjected to expert review before sign-off? Please provide any phrase-level validation metrics (e.g., precision and recall against blinded cardiologist adjudication) and explain how the interface flags sentences that lack sufficient evidentiary support. Clarifying these safeguards would help demonstrate the added value of EchoPrime’s free-text output over a standard template.

3. I acknowledge the planned randomised clinical trial slated to read out in 2026. Until at least preliminary results are available, I would encourage tempering workflow-efficiency claims or explicitly positioning them as future work.

In short, while the revision addresses several of my prior measurement requests, the new Bland-Altman plots suggest that non-trivial proportional bias and heteroscedasticity persist in key regression tasks. Consequently, both the calibration of the continuous outputs and the distinctive advantages of the free-text generation still need empirical validation. Demonstrating either (a) superior diagnostic performance relative to PanEcho, or (b) proven reductions in reporting time/error through prospective or pilot data, would strengthen the manuscript.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
In this second revision to your foundational model of echocardiography interpretation, the authors have addressed many of my concerns about the manuscript. Overall, the manuscript does a really nice job of translating a key principle of AI research

into the clinical domain: namely that some high dimensional space can be mapped to a low dimension (learned) manifold. In this case, an echocardiographic study (consisting of many videos each with many pixels, i.e. very many dimensions!) can be mapped to a lower vector space that can be clustered. You further take this one step further by doing report retrieval to basically translate that vector space back into clinician speak (i.e. embedding vector space into clinical natural language).

You better acknowledge the fundamental limitation that this current approach does not utilize any standard echo measurements from still images (Doppler, M-mode, or 2D still images). A pitfall of doing this is that any retrieved measurements are basically just guesses by the foundational algorithm. Patients that are similar to “xyz” will generally have measurements close to a certain distribution. However, this “guessing” is really only as good as the model’s ability to learn given the data at hand (which lacks actual measurements).

Remarkably, in your rebuttal to me (but not in the manuscript), you note an on-line pre-print “EchoNet-Measurements”, https://www.medrxiv.org/content/10.1101/2025.03.18.25324215v1, in which you appear to overcome this limitation. You demonstrate accuracy in a host of 1-D, 2-D, and Doppler measurements similar to Tromp et al (Lancet Dig Health 2022). Unfortunately, this capability is not integrated into Echo-Prime in the current manuscript, which thus remains a limitation to the current report. There is no question that the authors have developed a significant advance in AI interpretation of echocardiograms, and with the multiple revisions and broader validation, this has become a high impact publication. However, the lack of integration of quantitative measurements into the echo reports remains a significant limitation. I fully appreciate the challenge of performing this in the context of the current manuscript, given word count limitations and the divergent methods of the two approaches. I would be supportive of the current approach alone, but you should emphasize more the need to integrate measurements into the automated reports in the future. If it is your policy to cite preprint publications, I would suggest doing that to emphasize the importance of this capability. I suspect it won’t be long till we see a manuscript from your group that will combine quantitative measurements with this video-encoding foundational model, which will be a truly remarkable achievement.

Other than this broad observation, I believe you have adequately addressed my concerns from the prior revision.

(Remarks on code availability)

Referee #3

(Remarks to the Author)
I have carefully reviewed the comments from all reviewers as well as the authors’ rebuttal. The absence of prospective clinical validation substantially limits the clinical credibility of EchoPrime. Without real-world data demonstrating clinical efficacy and physician acceptance, the model's practical utility remains speculative.

Moreover, the manuscript lacks empirical evidence of clinician interaction with EchoPrime. Demonstrating actual clinician engagement, workflow integration, and efficiency gains is essential, especially for a tool intended to function in real clinical settings. Even preliminary user studies or pilot trials would considerably strengthen the authors’ claims.

More critically, the proposed vision-language model is not built upon any established large or small language model backbone (Llama 3.2, InternVL, QWen, etc). Given the absence of general-purpose language pretraining, its generative language capabilities are inherently constrained. This limitation raises concerns regarding its ability to generalize across diverse reporting tasks. Furthermore, the model’s performance should be directly compared with recent state-of-the-art alternatives, such as GPT-4o, Gemini 2.5 Pro, and domain-specific models like MedGemma, Llava-med. In light of these considerations, I no longer believe the proposed AI approach represents the best available solution for this task.

In summary, the technical approach underlying the AI method for the general healthcare domain lacks novelty and is, in fact, somewhat outdated at this point. I do not believe that any further rebuttal can adequately address this major concern.

(Remarks on code availability)
It is not well-maintained and has not been updated for an extended period of time.

Version 5:

Reviewer comments:

Referee #1

(Remarks to the Author)
This revision presents extensive benchmarking of EchoPrime across five institutional datasets, with consistent advantages in key regression tasks such as LVEF and RVSP. However, several foundational issues remain unresolved, and multiple original reviewer concerns, particularly regarding generalizability, calibration, and validation methodology, have not been adequately addressed in the rebuttal.

First, the use of Cedars-Sinai as a validation site is methodologically problematic. From prior publications and model descriptions, it is clear that Cedars was one of EchoPrime's primary training sources. Yet the manuscript presents Cedars results as external validation, where EchoPrime also performs best—winning 13 of 15 tasks against PanEcho, including the largest performance margins in LVEF and RVSP estimation. This undermines the claim of generalizability. The authors must provide a detailed breakdown of which sites contributed to model training and which were held out entirely for testing, alongside a clear statement on whether any patient overlap occurred across datasets. Without this, readers cannot interpret the Cedars performance as independent.

Second, while EchoPrime outperforms PanEcho on many tasks, the manuscript lacks formal statistical comparison. No paired significance tests (such as DeLong for AUROC or bootstrap methods for R^2 and MAE) are presented. Given that PanEcho outperforms EchoPrime on specific chamber size and morphology tasks (e.g., LA dilation at Stanford 0.81 vs 0.73, RA dilation 0.84 vs 0.77, BAV at Cedars 0.95 vs 0.83), this absence weakens the comparative analysis. The authors should pre-specify a primary endpoint (e.g., LVEF MAE), conduct paired tests across all shared metrics, and consider reporting macro-averaged performance across tasks to support any claim of superiority.

Third, the calibration of EchoPrime's regression outputs remains unquantified. Referee #1 requested Bland–Altman analyses to assess proportional bias and heteroscedasticity, including slope estimates, statistical testing, and decile-wise limits of agreement. Despite this, the authors provide only qualitative plots and do not report slope ±95% CI, p-values, or recalibration experiments. Visual inspection suggests clear underestimation at higher values and increased error spread for several outputs, including RVSP and TR gradient. This demands quantification. At minimum, the authors should report slope and LoA by value decile for key continuous variables and provide recalibrated performance using simple linear or isotonic correction to evaluate the clinical relevance of these biases.

Fourth, the manuscript's core claim that EchoPrime is "superior across most tasks" is overstated. The benchmarking data show mixed results: PanEcho matches or outperforms EchoPrime in several chamber size assessments, particularly at Stanford and CGMH, and RVSP is not reported at two of the five sites. A more accurate framing would acknowledge that EchoPrime performs comparably or better across most, but not all—tasks, with particularly strong gains in LVEF and RVSP, while PanEcho maintains advantages in certain structural phenotypes.

Fifth, the narrative generation feature of EchoPrime remains speculative. Although technically promising, there is no validation of the generated free-text reports, no check of potential hallucinated phrases, and no quantitative data on editing burden or time saved compared to template-based workflows. Given the acknowledged risk of inappropriate descriptors (e.g., "granular sparkling," "hyper-trabeculation," "features suggestive of ARVC"), some form of validation is essential. This could take the form of a pilot study measuring keystrokes, edit time, or phrase-level precision and recall against expert adjudication. In the absence of such data, the claims regarding clinical utility and workflow efficiency should be clearly presented as future directions, not as current strengths.

Finally, the discussion does not address the confounding impact of data scale. EchoPrime was trained on 12 million echo videos: ten times more than PanEcho. While this may partly explain the observed performance advantage, the scale differential introduces a bias that should be acknowledged when interpreting head-to-head comparisons. The model's superiority may reflect access to larger training resources as much as architectural innovation.

In summary, while EchoPrime is an impressive model with promising results, its evaluation design needs greater transparency, its calibration should be quantitatively assessed, and its claims moderated in light of performance trade-offs and unresolved reviewer concerns.

(Remarks on code availability)

Referee #2

(Remarks to the Author)
For this review, I was asked to focus primarily on the relative merits of EchoPrime vs PanEcho, which you benchmark thoroughly in the supplement.

Overall, I have no doubt that EchoPrime is the superior model both in terms of performance and novelty. The authors' rebuttal gives good points of comparison between the two (3 with EchoNet-Measurements) models on the first page of your rebuttal to Reviewer 1. I did not see this in the main manuscript or the supplement, but you should consider including it to better amplify the differences in architecture and output among the 3 models.

Overall, Supplemental Tables 2-6 show EchoPrime with superior performance in most tasks and especially outperforming in

tasks that reviewers have previously criticized it for poor performance. In particular, EchoPrime's performance for LVEF (perhaps the most fundamental parameter from an echo) is far superior to PanEcho (MAE 4.8 vs 7.5, p highly significant, though not stated). It is actually surprising how poorly PanEcho works for EF given that there are many AI algorithms that used similar supervised training with a CNN to get much better results (e.g., Asch et al, DOI: 10.1161/CIRCIMAGING.120.012293 and Tromp et al, doi.org/10.1038/s41467-022-34245-1). This holds true for virtually all parameters and diagnoses, and even in those few diagnoses for which the PanEcho AUROC point estimate is higher than EchoPrime's (e.g., RA dilation and bicuspid AV in the Cedars Sinai testset), the 95% C.I.'s overlap, showing no statistical significance. This becomes particularly striking for the many diagnoses that PanEcho did not train on, such as mitral annular calcification, MitraClip, and TAVR, but that EchoPrime, with its more holistic training (and on 10x data than PanEcho) was able to address, giving AUROCs for these diagnoses between 0.96 and 1.00.

EchoPrime is also more novel in its approach, whereas PanEcho is a standard ML model, using ~1 million echo videos to train a CNN by supervised training to perform specific tasks. EchoPrime does a nice job translating a key principle of AI research into the clinical domain: namely that some high dimensional space can be mapped to a low dimension (learned) manifold. In this case, an echocardiographic study (consisting of many videos each with many pixels, i.e. very many dimensions) can be mapped to a lower vector space that can be clustered. You take this one step further by doing report retrieval to basically translate that vector space back into clinician speak (i.e. embedding vector space into clinical natural language). One salutary consequence of this is that EchoPrime "thinks" like an echocardiographer, that is, it learns which views are likely to contribute to specific diagnoses, and which ones aren't (e.g., don't look at the apical 2-chamber view for any information on the RV).

My principal critique of EchoPrime all along has been that it does not extract specific quantitative information from the echo, in particular that it does not use PW or CW Doppler to characterize diastolic function, aortic valve gradient, right ventricular systolic pressure, etc. However, I don't think PanEcho makes any effective use of these data either. PanEcho's r^2 for RVSP is only 0.27, significantly less than EchoPrime's (admittedly marginal) 0.43. Given that you have already submitted EchoNET-Measurements in preprint form, which demonstrates your ability to extract these measurements from Doppler, M-mode, and still images with an accuracy similar to sonographers. It is entirely logical to combine EchoPrime with EchoNet-Measurements to further finetune the report generation.

My only suggestion would be to consider being a bit more expansive in discussing PanEcho in the main manuscript to compare and contrast with EchoPrime, perhaps putting the 3-model comparison from your rebuttal in the supplement and directing the readers' attention to specific findings in Supplemental Tables 2-6.

I look forward to hearing your EchoPrime and EchoNet-Measurements presentations at the ASE meeting.

James D. Thomas, MD, FASE, FACC, FESC

(Remarks on code availability)

Referee #3

(Remarks to the Author)
I have carefully reviewed the comments from all reviewers as well as the authors' rebuttal. In their response, the authors state that comparisons with third-party models are infeasible due to data privacy constraints. However, this justification is not entirely persuasive. Numerous state-of-the-art large language models (LLMs)—including LLaVA-3.2, LLaVA-Med, and Qwen-3—are openly available and amenable to fine-tuning in secure, local environments without violating data governance requirements. Within the field of multimodal report generation, fine-tuning LLMs on domain-specific datasets has become standard practice. Given that the authors possess a 12-million-sample dataset, the omission of benchmarking against fine-tuned open-source LLMs constitutes a significant gap. A direct comparison would provide a more rigorous and informative assessment of EchoPrime's relative performance.

From a methodological standpoint, the core architecture employed in EchoPrime (i.e., a convolutional video encoder paired with a BERT-based text module) does not reflect contemporary standards in multimodal learning. The absence of a foundation model backbone pretrained on large-scale instruction-following or generative tasks (e.g., via causal language modeling or reinforcement learning with human feedback) limits the model's expressiveness and its capacity for generalization across diverse clinical reporting tasks. Although the authors present an extensive retrospective evaluation, the underlying model design is technically dated and lacks the flexibility, reasoning capability, and transferability that characterize current-generation multimodal architectures.

Moreover, as emphasized by other reviewers, the absence of prospective clinical validation—whether through randomized trials, user studies, or real-world workflow integration—substantially diminishes the translational significance of the work. Claims regarding improvements in clinical efficiency and report quality remain hypothetical in the absence of empirical evidence. If we waved the need for clinical testing of this model, the contribution of the proposed AI solution is outdated, limited, and not SOTA, thus not promising and novel.

In its current form, the manuscript offers limited advancement in both AI algorithmic innovation and clinical applicability. Substantive revision is warranted, including comparative evaluation against modern domain-adapted LLMs and

incorporation of prospective or clinician-facing validation, in order to meet the threshold for publication in Nature.

(Remarks on code availability)

Referee #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports.

(Remarks on code availability)

Reviewer Comments

Referee #1 (Remarks to the Author):

The manuscript introduces EchoPrime, a multi-task model for echocardiography that claims state-of-the-art performance across 17 binary and 2 continuous tasks, utilizing a contrastive encoder and anatomical attention module. I have some comments following the points below:

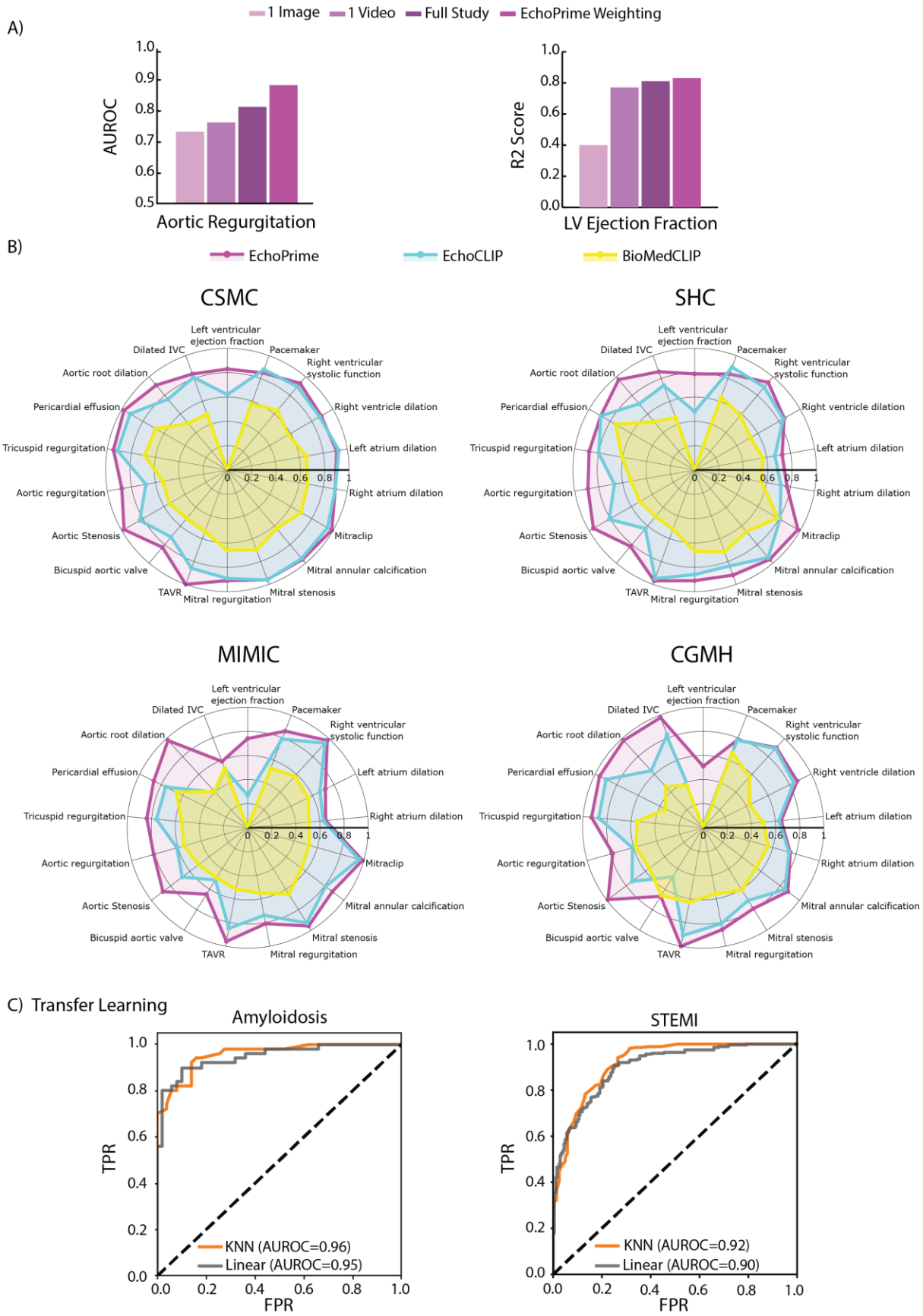
Reply: We thank the reviewer for their feedback and time in reviewing this manuscript. In contrast to prior work, which focus on either individual tasks or a limited set of binary and continuous tasks, EchoPrime generates the entire report, which comprises of 15 sections of over 1000 different statements that cardiologists frequently make in interpreting echocardiograms. The original manuscript benchmark metrics sought to highlight 19 relevant tasks for comparison with previous models, but EchoPrime is able to do many more tasks than just the 19 tasks mentioned.

These EchoPrime statements comprise the full range of common cardiovascular diseases and include both classification (often multiclass) and regression/continuous tasks. To clarify this point, in this revision, we demonstrate EchoPrime's performance on an increased range of classification tasks and comparisons with 11 regression tasks.

Furthermore, to emphasize the generalizability of EchoPrime's performance, we introduce two new external validation sites (including an international validation site) which replicate the strong performance across multiple settings, image acquisition protocols, and workflows.

In **Results** "**Automated echocardiogram interpretation without supervised learning**" section:

> *This approach allows a single foundation model to be applied directly to all echocardiography interpretation tasks, enabling the interpretation of hundreds of pathologies across 15 anatomical sections resulting in over 1000 different report statements that cardiologists frequently make in interpreting echocardiograms.*
>
> *We evaluated EchoPrime on a diverse set of benchmarks for cardiac structure and function across datasets from four different institutions (**Error! Reference source not found.**). On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging (Supplementary Tables 2-5), and without additional fine-tuning or training, EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6). EchoPrime had a mean AUC of 0.92 on CSMC, 0.89 on Stanford Healthcare (SHC), 0.86 on Beth Israel Deaconess Medical Center (MIMIC) and 0.86 on Chang Gung Memorial Hospital (CGMH) in Taiwan across 17 classification tasks and 11 regression tasks (Table 2, Supplementary Table 7).*


A)
1 Image
1 Video
Full Study
EchoPrime Weighting
AUROC
Aortic Regurgitation
R2 Score
LV Ejection Fraction
B)
EchoPrime
EchoCLIP
BioMedCLIP
CSMC
SHC
MIMIC
CGMH
Left ventricular ejection fraction
Pacemaker
Right ventricular systolic function
Right ventricle dilation
Left atrium dilation
Right atrium dilation
Mitraclip
Mitral annular calcification
Mitral stenosis
Mitral regurgitation
TAVR
Bicuspid aortic valve
Aortic Stenosis
Aortic regurgitation
Tricuspid regurgitation
Pericardial effusion
Aortic root dilation
Dilated IVC
C) Transfer Learning
Amyloidosis
STEMI
TPR
FPR
KNN (AUROC=0.96)
Linear (AUROC=0.95)
KNN (AUROC=0.92)
Linear (AUROC=0.90)

Supplementary Table 7 EchoPrime performance on regression tasks

Task	R^2	MAE
Ejection Fraction	0.83 (0.81-0.85)	4.79 (4.60-4.97)
Pulmonary Artery Pressure	0.43 (0.39-0.47)	7.30 (7.07-7.80)
Wall Motion Score	0.77 (0.74-0.80)	0.10 (0.10-0.11)
TAPSE	0.24 (0.05-0.49)	0.35 (0.28-0.45)
MV Area	0.34 (0.20-0.44)	0.81 (0.71-0.93)
Transaortic velocity	0.65 (0.58-0.72)	23.84 (20.38-27.82)
Transaortic gradient	0.36 (0.29-0.42)	6.84 (6.25-7.45)
TRgradient	0.28 (0.16-0.39)	6.95 (6.51-7.46)
Sinus of Valsalva Size	0.62 (0.55-0.67)	0.24 (0.22-0.26)
Ascending Aorta Size	0.16 (0.02-0.45)	0.41 (0.29-0.63)
IVC Diameter	0.70 (0.65-0.75)	2.85 (2.65-3.06)

1. Clinical Utility and Workflow Integration: EchoPrime's precision improves with additional views and videos (Figure 2A), indicating potential value as a diagnostic support tool. However, this approach relies on high-quality, multi-view echocardiographic data, which requires skilled sonographers or physicians. This dependency may limit its feasibility in resource-constrained settings.

Reply: Thank you for your comments regarding EchoPrime as a diagnostic support tool. Echocardiography is the most common form of cardiac imaging, and by including the wide range of images from over a decade of a large academic echo lab, has both high and low quality images and a wide range of physiology not seen in small datasets. To assess EchoPrime's performance in low-quality resource-constrained setting, we report the performance on studies deemed technically challenging and to assess EchoPrime's performance with limited datasets, we report the performance using only one A4C video per study. The following results show EchoPrime works good even in a low-quality and resource-constrained settings.

Furthermore, in this revision, we add two more external validation sites, including an international site where cardiologists perform echocardiography but record more limited/abbreviated studies (averaging 20 videos + images per study rather than conventionally 80 – 100 videos + images per study in American sites) and we show the generalizable strong performance in different practice patterns.

In **Results** "**Leveraging multi-view echocardiographic data improves performance**" section

EchoPrime can additionally be leveraged even in low-quality images or settings where limited views are acquired. We evaluated EchoPrime in the set of studies where the study was described as "technically difficult" in the clinician report and show similar performance (Supplementary Table 8). To test the performance in a single view setting, we test EchoPrime on 2172 studies with only one A4C video per study. The performance across all tasks were similar to EchoPrime performance in the full setting (Supplementary Table 8).

Supplementary Table 8 EchoPrime performance in low-quality resource-constrained setting

	FULL TEST	LOW-QUALITY	SINGLE-VIEW
LV ejection fraction (R^2)	0.83 (0.81-0.85)	*0.80 (0.75-0.84)*	0.78 (0.75-0.80)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	*5.30 (4.86-5.76)*	5.22 (5.02-5.43)
Pacemaker	0.85 (0.82-0.88)	*0.89 (0.84-0.93)*	0.83 (0.80-0.85)
RV systolic function	0.93 (0.91-0.95)	*0.91 (0.85-0.96)*	0.92 (0.89-0.94)
RV dilation	0.89 (0.83-0.95)	*0.85 (0.72-0.96)*	0.92 (0.87-0.95)
LA dilation	0.91 (0.86-0.96)	*0.95 (0.89-0.99)*	0.92 (0.87-0.96)
RA dilation	0.90 (0.82-0.97)	*0.89 (0.70-1.00)*	0.90 (0.82-0.97)
Mitraclip	0.99 (0.99-1.00)	*1.00 (0.99-1.00)*	0.97 (0.94-0.99)
Mitral annular calcification	0.96 (0.95-0.97)	*0.97 (0.94-0.98)*	0.94 (0.91-0.96)
Mitral stenosis	0.96 (0.91-0.99)	*0.97 (0.90-1.00)*	0.94 (0.87-0.98)
Mitral regurgitation	0.92 (0.91-0.94)	*0.91 (0.84-0.96)*	0.89 (0.87-0.91)
TAVR	1.00 (0.99-1.00)	*1.00 (0.99-1.00)*	0.90 (0.87-0.93)
Bicuspid AV	0.83 (0.67-0.96)	*0.72 (0.44-1.00)*	0.78 (0.62-0.92)
Aortic stenosis	0.98 (0.96-0.99)	*0.93 (0.83-1.00)*	0.80 (0.75-0.85)
Aortic regurgitation	0.88 (0.83-0.93)	*0.78 (0.38-0.99)*	0.75 (0.68-0.82)
Tricuspid regurgitation	0.95 (0.93-0.97)	*0.98 (0.97-0.99)*	0.92 (0.88-0.94)
Pericardial effusion	0.98 (0.95-0.99)	*0.94 (0.91-0.97)*	0.93 (0.89-0.97)
Aortic root dilation	0.91 (0.83-0.98)	*0.96 (0.91-1.00)*	0.81 (0.70-0.91)
Dilated IVC	0.84 (0.82-0.86)	*0.82 (0.77-0.87)*	0.80 (0.78-0.83)

PA pressure (R^2)	0.43 (0.39-0.47)	*0.32 (0.17-0.44)*	0.36 (0.31-0.40)
PA pressure (MAE)	7.30 (6.95-7.67)	*7.76 (6.96-8.63)*	7.67 (7.29-8.06)

The model's focus on classification tasks (rather than regression tasks beyond EF estimation) could also restrict its usefulness for comprehensive echocardiography, which depends on detailed measurements and morphological assessments essential for patient care.

Reply: Thank you for this important question. To clarify, in contrast to prior work, which focus on a limited set of binary and continuous tasks. EchoPrime generates the entire report, which comprises of 15 sections of over 1000 different statements. These statements comprise the full range of common cardiovascular diseases and include both classification (often multiclass) and regression/continuous tasks. When there are so many potential outputs, classification tasks (yes/no whether the model predicted an attribute in the same way as a cardiologist) is the easiest way to produce metrics, however there are multiple other regression tasks and other metrics that can be evaluated. In addition to LVEF and RVSP as regression tasks, our model also produces a wall motion score, TR gradient and other quantitative metrics (IVC diameter, Sinus of Valsalva) in an echo report. In this revision, we provide additional metrics to assess EchoPrime's performance on these regression tasks in Supplementary Table 7.

In **Results** "**Automated echocardiogram interpretation without supervised learning**" section

> *On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging (Supplementary Tables 2-5), and without additional fine-tuning or training, EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6). EchoPrime had a mean AUC of 0.92 on CSMC, 0.89 on Stanford Healthcare (SHC), 0.86 on Beth Israel Deaconess Medical Center (MIMIC) and 0.86 on Chang Gung Memorial Hospital (CGMH) in Taiwan across 17 classification tasks and 11 regression tasks (Table 2, Supplementary Table 7).*

Supplementary Table 7 EchoPrime performance on regression tasks on the internal held-out test set

Task	R^2	MAE
Ejection Fraction	0.83 (0.81-0.85)	4.79 (4.60-4.97)
Pulmonary Artery Pressure	0.43 (0.39-0.47)	7.30 (7.07-7.80)
Wall Motion Score	0.77 (0.74-0.80)	0.10 (0.10-0.11)
TAPSE	0.24 (0.05-0.49)	0.35 (0.28-0.45)
MV Area	0.34 (0.20-0.44)	0.81 (0.71-0.93)
Transaortic velocity	0.65 (0.58-0.72)	23.84 (20.38-27.82)

Transaortic gradient	0.36 (0.29-0.42)	6.84 (6.25-7.45)
TRgradient	0.28 (0.16-0.39)	6.95 (6.51-7.46)
Sinus of Valsalva Size	0.62 (0.55-0.67)	0.24 (0.22-0.26)
Ascending Aorta Size	0.16 (0.02-0.45)	0.41 (0.29-0.63)
IVC Diameter	0.70 (0.65-0.75)	2.85 (2.65-3.06)

If EchoPrime is intended for high-pre-test probability populations, its added value is unclear since comprehensive studies and specialist reviews are already standard. Alternatively, if the goal is to "clear" low-risk patients, its negative predictive value (NPV) should be compared to simpler and more accessible tools. For instance, Elias et al.'s ECG-based model detects aortic valve regurgitation with an AUC of 0.84 (JACC, 2022), close to EchoPrime's AUC of 0.88. Given that ECG is less resource-intensive and more widely available, the small AUC improvement raises questions about EchoPrime's cost-effectiveness. The authors should clarify EchoPrime's NPV compared to existing tools in low- and intermediate-pre-test populations, its advantages in high-pre-test settings where comprehensive studies are already performed, and how it accommodates the detailed assessments needed in echocardiography. These points are essential to evaluate its clinical value and potential integration into workflows.

Reply: Thank you for your thoughtful comments. The comparison with an AI ECG model is potentially misleading and comparing apples to oranges. ECGs tend to be more commonly obtained in healthy individuals, and a relatively healthy cohort with few abnormal cases will inflate performance metrics. This bias will be reflected in a high reported NPV while the model would not perform as well in such an echo-lab based cohort. In contrast, in our paper, we compare with many other echocardiography AI models and show similar to superior performance.

Echocardiography is the most common form of cardiac imaging and is used in both screening (low pretest probability patients) and patients with known cardiovascular disease (high pretest probability patients). Notably, echocardiography is the definitive test for many of these cardiovascular diseases, while an ECG based model is to highlight suspicion and would require a confirmatory test (like echocardiography). In this setting, a direct comparison of AUC might be misleading, because small differences in AUC might seem like small difference but is the difference between finding suspicion for a disease *(in an ECG setting)* and diagnosing a disease in the definitive testing modality of choice (*in the ehco setting)*.

We would like to clarify that EchoPrime is not intended for clearing low-risk patients or as a screening tool in low-pre-test probability populations. Instead, its application is integrated within the standard echocardiography workflow, where echocardiograms are obtained, an initial report is drafted by the sonographer (but often is not due to time constraints), and the final report is reviewed and modified by a cardiologist. EchoPrime's primary goal is to assist sonographers and cardiologists in interpreting echocardiographic studies by automating preliminary report generation. This approach offers key advantages including:

1. **Enhanced efficiency:** Automating the report generation process streamlines workflow, allowing for faster interpretation and turnaround.
2. **Consistency:** By reducing inter-physician variability, EchoPrime helps standardize interpretations, ensuring more consistent quality of care.

2. Observed Discrepancy in Outputs: Testing the model on my MacBook Pro (macOS Apple M3 Max, 36 GB) with a cardiac ultrasound not provided in the GitHub repository, I found a discrepancy between the EF reported in the generated text (55%) and the EF derived from output logits (61%) using the provided Jupyter Notebook. This inconsistency raises questions about which value should be considered accurate and whether outputs are standardized across formats. It would be helpful for the authors to explain the rationale behind this discrepancy, such as whether it reflects differences in post-processing methods or thresholds used for generating outputs.

Reply: Thank you for your rigorous evaluation of our code. In our codebase, we included intermediate outputs for demonstration purposes. These debugging outputs are not directly reflected in the results and meant to show successful individual report retrieval. EchoPrime embeds the study in the latent space and then finds the closest report embeddings, but the output is the weighted average of the 50 closest reports. In your example, the 55% LVEF output appears from mapping to the closest retrieved report while 61% is the result of retrieving 50 closest reports and then averaged LVEF across them (the intended final EchoPrime output). For metric purposes, the intended final EchoPrime output was used for assessment.

3. While the manuscript does not explicitly discuss pressure estimations such as right atrial pressure (RAP), testing the software using the provided Jupyter Notebook generated a text report that included an estimated RAP value (12 mmHg, see above). This raises questions about whether such estimations are validated or reliable. Are these pressure values directly derived from model predictions, and if so, have they been validated?

Reply: EchoPrime can predict any cardiac features present in a standard echocardiography report, and goes beyond 17 classification tasks and 2 regression tasks. We selected 19 representative features for analysis that encompass most cardiac structures to give insight into the overall performance, however there are over 1000 possible output phrases.

In the reviewers output, it looks like 12mmHg is for the TR max velocity (RV/RA pressure gradient), not the right atrial pressure (RAP). Per reviewer's request, in this revision we add right TR (Tricuspid Valve) gradient as a clinical evaluation metric in Supplementary Table 7.

Supplementary Table 7 EchoPrime performance on regression tasks on the internal held-out test set

Task	R^2	MAE
Ejection Fraction	0.83 (0.81-0.85)	4.79 (4.60-4.97)

Pulmonary Artery Pressure	0.43 (0.39-0.47)	7.30 (7.07-7.80)
Wall Motion Score	0.77 (0.74-0.80)	0.10 (0.10-0.11)
TAPSE	0.24 (0.05-0.49)	0.35 (0.28-0.45)
MV Area	0.34 (0.20-0.44)	0.81 (0.71-0.93)
Transaortic velocity	0.65 (0.58-0.72)	23.84 (20.38-27.82)
Transaortic gradient	0.36 (0.29-0.42)	6.84 (6.25-7.45)
TR gradient	**0.28 (0.16-0.39)**	**6.95 (6.51-7.46)**
Sinus of Valsalva Size	0.62 (0.55-0.67)	0.24 (0.22-0.26)
Ascending Aorta Size	0.16 (0.02-0.45)	0.41 (0.29-0.63)
IVC Diameter	0.70 (0.65-0.75)	2.85 (2.65-3.06)

4. Benchmarking Against Comprehensive Tools: The manuscript provides a comparison with EchoCLIP and BioMedCLIP. However, EchoPrime's focus on binary classifications raises questions about its clinical utility relative to commercially available tools like US2AI (https://us2.ai/product/), which deliver fully automated, multi-view quantitative echocardiography reports. US2AI includes continuous outputs including LV/RV morphology, and diastolic function, which are critical for standard echocardiographic reporting. While the lack of access to proprietary systems like US2AI is understandable, at least a discussion of how EchoPrime compares to or complements such tools would help articulate its standalone value and potential niche in clinical workflows.

Reply: Thank you for this important question. Because US2.AI is a commercial product is blackbox and code not publicly available, we are not able to benchmark performance metrics, however we note that US2.AI does only segmentation / quantitative assessments, and is not able to do any of the qualitative report findings EchoPrime reports. As the reviewer notes, the echocardiography report is primarily qualitative rather than quantitative assessments (more classification tasks than regression tasks) and US2AI and similar products cannot perform any of the classification tasks.

In addition to benchmarking with all available publicly available echocardiography AI models, EchoPrime is the first foundation model designed for comprehensive echocardiography evaluation and captures the full range of likely interpretation statements. Furthermore, unlike US2AI, which was trained on only a few thousand studies, EchoPrime leverages a training dataset of 12 million echocardiography videos, the largest yet training dataset for echocardiography AI and provides a more robust and generalized understanding of cardiac physiology.

5. Statistical and Clinical Significance: The manuscript highlights small percentage improvements in performance over prior models, such as a 2% increase in AUC for tricuspid regurgitation. However, it is unclear whether these improvements are statistically significant or

clinically meaningful. Without p-values or confidence intervals (at least I did not find them), the reported gains could reflect expected variability rather than true advancements.

Reply: EchoPrime outperforms previous foundation models and matches or outperforms task specific models. Often there's a small improvement to task-specific models (like 2% in TR mentioned), however EchoPrime is comprehensive in nature and is a single model rather than requiring many models for an entire echocardiography to report. EchoPrime can predict all these features (in an unsupervised manner), whereas a task-specific model means that a model is specifically fine-tuned for that feature. The task specific improvements in AUC for such a comprehensive general model, highlight EchoPrime's exceptional generalizability, demonstrating its ability to match or surpass task-specific models without feature-specific fine-tuning.

As per reviewer's request we add confidence intervals to all AUC, R2 and MAE metrics in the paper and explain how we calculated them in methods. We also report confidence intervals for task specific models when that is available in the original respective papers.

In **Methods "Evaluation and Benchmarking" section**:

> *To calculate the confidence intervals for performance metrics, we randomly sample the dataset with replacement, maintaining the same sample size as the original dataset, and compute the performance metrics. This process is repeated 10,000 times to generate a distribution of the metrics. Finally, determine the 2.5th and 97.5th percentiles to obtain the bootstrapped 95% confidence intervals.*

In Results "**Automated echocardiogram interpretation without supervised learning**" section:

> *EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6).*

Supplementary Table 6 *EchoPrime vs Task Specific Models*

	ECHOPRIME INTERNAL	TASK SPECIFIC INTERNAL	ECHOPRIME EXTERNAL	TASK SPECIFIC EXTERNAL
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.81	**0.79 (0.76-0.82)**	0.77
LV ejection fraction (MAE)	4.79 (4.60-4.97)	**4.10**	**4.14 (3.94-4.34)**	6.00
Pacemaker	0.85 (0.82-0.88)	-	0.84 (0.81-0.87)	-
RV systolic function	0.93 (0.91-0.95)	-	0.94 (0.88-1.00)	-
RV dilation	0.89 (0.83-0.95)	-	0.85 (0.80-0.90)	-
LA dilation	**0.91 (0.86-0.96)**	0.85	0.73 (0.70-0.75)	-
RA dilation	0.90 (0.82-0.97)	-	0.77 (0.74-0.81)	-
Mitraclip	0.99 (0.99-1.00)	-	0.98 (0.94-1.00)	-
Mitral annular calcification	0.96 (0.95-0.97)	-	0.96 (0.94-0.98)	-

Mitral stenosis	*0.96 (0.91-0.99)*	-	*0.92 (0.86-0.98)*	-
Mitral regurgitation	***0.92 (0.91-0.94)***	***0.92 (0.90-0.93)***	*0.91 (0.89-0.93)*	***0.95 (0.92–0.97)***
TAVR	*1.00 (0.99-1.00)*	-	*0.97 (0.90-1.00)*	-
Bicuspid AV	*0.83 (0.67-0.96)*	-	*0.82 (0.73-0.90)*	-
Aortic stenosis	***0.98 (0.96-0.99)***	***0.98 (0.97- 0.99)***	***0.96 (0.92-0.99)***	*0.95 (0.94- 0.96)*
Aortic regurgitation	*0.88 (0.83-0.93)*	-	*0.89 (0.80-0.97)*	-
Tricuspid regurgitation	***0.95 (0.93-0.97)***	*0.93 (0.91- 0.94)*	*0.88 (0.86-0.91)*	***0.95 (0.94 - 0.96)***
Pericardial effusion	***0.98 (0.95-0.99)***	*0.94 (0.75- 0.99)*	*0.89 (0.77-0.98)*	-
Aortic root dilation	*0.91 (0.83-0.98)*	-	*0.97 (0.94-0.99)*	-
Dilated IVC	*0.84 (0.82-0.86)*	-	*0.86 (0.81-0.91)*	-
PA pressure (R^2)	*0.43 (0.39-0.47)*	-	*0.36 (0.30-0.42)*	-
PA pressure (MAE)	*7.30 (6.95-7.67)*	-	*7.97 (7.35-8.66)*	-

6. Performance in Challenging Imaging Conditions: the manuscript does not address EchoPrime's performance under suboptimal imaging conditions, such as poor acoustic windows, which are common in real-world practice. Understanding its robustness in these scenarios is crucial for assessing its clinical applicability. I believe including an analysis of performance stratified by image quality would enhance the model's credibility for diverse clinical settings.

Reply: Thank you for your suggestion. Echocardiography is the most common form of cardiac imaging, and by including the wide range of images from over a decade of a large academic echo lab, has both high and low quality images. To assess EchoPrime's performance in low-quality resource-constrained setting, we report the performance on studies deemed technically challenging. The following results show EchoPrime works good even in a low-quality images.

In **Results** "**Leveraging multi-view echocardiographic data improves performance**" section

> *EchoPrime can additionally be leveraged even in low-quality images or settings where limited views are acquired. We evaluated EchoPrime in the set of studies where the study was described as "technically difficult" in the clinician report and show similar performance (Supplementary Table 8). To test the performance in a single view setting, we test EchoPrime on 2172 studies with only one A4C video per study. The performance across all tasks were similar to EchoPrime performance in the full setting (Supplementary Table 8).*

Supplementary Table 8 *EchoPrime performance in low-quality resource-constrained setting*

	FULL TEST	***LOW-QUALITY***	***SINGLE-VIEW***

LV ejection fraction (R^2)	*0.83 (0.81-0.85)*	*0.80 (0.75-0.84)*	*0.78 (0.75-0.80)*
LV ejection fraction (MAE)	*4.79 (4.60-4.97)*	*5.30 (4.86-5.76)*	*5.22 (5.02-5.43)*
Pacemaker	*0.85 (0.82-0.88)*	*0.89 (0.84-0.93)*	*0.83 (0.80-0.85)*
RV systolic function	*0.93 (0.91-0.95)*	*0.91 (0.85-0.96)*	*0.92 (0.89-0.94)*
RV dilation	*0.89 (0.83-0.95)*	*0.85 (0.72-0.96)*	*0.92 (0.87-0.95)*
LA dilation	*0.91 (0.86-0.96)*	*0.95 (0.89-0.99)*	*0.92 (0.87-0.96)*
RA dilation	*0.90 (0.82-0.97)*	*0.89 (0.70-1.00)*	*0.90 (0.82-0.97)*
Mitraclip	*0.99 (0.99-1.00)*	*1.00 (0.99-1.00)*	*0.97 (0.94-0.99)*
Mitral annular calcification	*0.96 (0.95-0.97)*	*0.97 (0.94-0.98)*	*0.94 (0.91-0.96)*
Mitral stenosis	*0.96 (0.91-0.99)*	*0.97 (0.90-1.00)*	*0.94 (0.87-0.98)*
Mitral regurgitation	*0.92 (0.91-0.94)*	*0.91 (0.84-0.96)*	*0.89 (0.87-0.91)*
TAVR	*1.00 (0.99-1.00)*	*1.00 (0.99-1.00)*	*0.90 (0.87-0.93)*
Bicuspid AV	*0.83 (0.67-0.96)*	*0.72 (0.44-1.00)*	*0.78 (0.62-0.92)*
Aortic stenosis	*0.98 (0.96-0.99)*	*0.93 (0.83-1.00)*	*0.80 (0.75-0.85)*
Aortic regurgitation	*0.88 (0.83-0.93)*	*0.78 (0.38-0.99)*	*0.75 (0.68-0.82)*
Tricuspid regurgitation	*0.95 (0.93-0.97)*	*0.98 (0.97-0.99)*	*0.92 (0.88-0.94)*
Pericardial effusion	*0.98 (0.95-0.99)*	*0.94 (0.91-0.97)*	*0.93 (0.89-0.97)*
Aortic root dilation	*0.91 (0.83-0.98)*	*0.96 (0.91-1.00)*	*0.81 (0.70-0.91)*
Dilated IVC	*0.84 (0.82-0.86)*	*0.82 (0.77-0.87)*	*0.80 (0.78-0.83)*
PA pressure (R^2)	*0.43 (0.39-0.47)*	*0.32 (0.17-0.44)*	*0.36 (0.31-0.40)*
PA pressure (MAE)	*7.30 (6.95-7.67)*	*7.76 (6.96-8.63)*	*7.67 (7.29-8.06)*

7. Implementation and User-Friendliness: running the software on a MacBook using a CPU backend required adjustments. To improve user-friendliness and ensure reproducibility, the README should specify exact requirements for setting up the virtual environment. Providing a Dockerized version would ensure cross-platform compatibility and consistency. Currently, it is unclear whether the exact Python modules and dependencies used align with those intended by the authors, potentially impacting reproducibility and performance.

Reply: Thank you for your feedback. We add requirements.txt and a Dockerfile to improve the accessibility of the shared code. https://github.com/echonet/EchoPrime/blob/main/Dockerfile https://github.com/echonet/EchoPrime/blob/main/requirements.txt

8. Non-Echocardiographic Disease Predictions (Minor): the manuscript refers to STEMI and cardiac amyloidosis as "non-echocardiographic primary diseases," which may not fully capture their diagnostic pathways. Echocardiography is commonly used to diagnose cardiac amyloidosis, offering structural and functional insights, and may play a significant role in assessing cardiac function in STEMI patients. Clarifying the term "non-echocardiographic traits" would prevent potential misunderstandings and better align the claims with established clinical practices.

Reply: Thank you for this comment. We sought to clarify that echocardiography is not the definitive test of choice for diagnosing STEMI or cardiac amyloidosis. ECGs or nuclear imaging is often used to confirm the diagnosis, while echocardiography can find corroborating evidence that is insufficient to confirm diagnosis.

Similar to what the reviewer described with AI-ECG models, even when performing well, STEMI and Amyloidosis diagnoses are not something that commonly appears in the report written by a sonographer or echocardiography. We clarify in this revision, and we modify the term "non-echocardiographic traits" with "Multimodal cardiovascular diagnoses" throughout the text.

In **Methods** "**Disease Diagnosis with Probing**" section:

> ***Multimodal Disease Diagnosis***
> *We apply k-nearest neighbors (KNN) and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyhloidosis as two diseases for which echocardiography is not the definitive test echocardiography can find corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study, and used along with clinician-assigned diagnoses to fit the models. For KNN probing, we use KNeighborsClassifier, and for linear probing, we apply LogisticRegression, both from scikit-learn library.*

Referee #2:

Congratulations to the authors as you continue your progressive refinement of the use of AI in echocardiography. Expanding from your recent foundational interpretive model EchoCLIP, which was trained on one million echo video clips, you have now used over 12 million echo clips (essentially all echocardiograms done at Cedars Sinai over a decade, along with their de-identified echo reports) to train a successor foundation model (EchoPrime), which has many advantages over its predecessor. Among the advances in modeling is contrastive learning to form a joint image-text representation, a more sophisticated view classifier that allows an anatomic attention model to understand which of the 58 characterized views are most likely to show the

finding of current interest, and use of the full video clip for assessment rather than individual frames. The model was tested on hold-out sets of 4044 total echo studies from both Cedars Sinai as well as Stanford. EchoPrime claims the ability to predict the presence of hundreds of pathologies encountered over the years but was tested on 19 fairly common features encapsuling chamber size and function and valve stenosis and regurgitation as well as implanted devices (e.g., prosthetic valves, MitraClip, and pacers). Two parameters (LV ejection fraction and RV systolic pressure) were compared to the clinical report by linear regression (r^2 and mean average error), while the other 17 were tested with AUROC for binary variables, significantly outperforming EchoCLIP, the more generalized foundation model BioMedCLIP, and several task specific models.

Reply: Thank you for your comments regarding our work on AI in echocardiography. As you noted in your review, EchoPrime is able to predict the presence of hundreds of pathologies across anatomical 15 sections of over 1000 different statements that cardiologists frequently make in interpreting echocardiograms. The comprehensiveness of EchoPrime's interpretation is also shown in additional metrics of EchoPrime performance in more tasks requested by the reviewers (Supplementary Table 7) These results were not previously described in the first version of manuscript but are evaluation metrics from the same model without more training.

Supplementary Table 7 *EchoPrime performance on regression tasks on the internal held-out test set*

Task	*R^2*	*MAE*
Ejection Fraction	*0.83 (0.81-0.85)*	*4.79 (4.60-4.97)*
Pulmonary Artery Pressure	*0.43 (0.39-0.47)*	*7.30 (7.07-7.80)*
Wall Motion Score	*0.77 (0.74-0.80)*	*0.10 (0.10-0.11)*
TAPSE	*0.24 (0.05-0.49)*	*0.35 (0.28-0.45)*
MV Area	*0.34 (0.20-0.44)*	*0.81 (0.71-0.93)*
Transaortic velocity	*0.65 (0.58-0.72)*	*23.84 (20.38-27.82)*
Transaortic gradient	*0.36 (0.29-0.42)*	*6.84 (6.25-7.45)*
TRgradient	*0.28 (0.16-0.39)*	*6.95 (6.51-7.46)*
Sinus of Valsalva Size	*0.62 (0.55-0.67)*	*0.24 (0.22-0.26)*
Ascending Aorta Size	*0.16 (0.02-0.45)*	*0.41 (0.29-0.63)*
IVC Diameter	*0.70 (0.65-0.75)*	*2.85 (2.65-3.06)*

One important echo capability not well demonstrated was detection and quantitation of regional wall motion abnormalities. It is unclear if EchoPrime is not capable of doing this natively or if it just wasn't part of the test suite but should be mentioned in either case.

Reply: Thank you for this important comment, EchoPrime is able to describe regional wall motion abnormalities. We add wall motion score as an additional metric to support previous argument.

Supplementary Table 7 *EchoPrime performance on regression tasks on the internal held-out test set*

Task	*R^2*	*MAE*
Ejection Fraction	*0.83 (0.81-0.85)*	*4.79 (4.60-4.97)*
Pulmonary Artery Pressure	*0.43 (0.39-0.47)*	*7.30 (7.07-7.80)*
Wall Motion Score	***0.77 (0.74-0.80)***	***0.10 (0.10-0.11)***
TAPSE	*0.24 (0.05-0.49)*	*0.35 (0.28-0.45)*
MV Area	*0.34 (0.20-0.44)*	*0.81 (0.71-0.93)*
Transaortic velocity	*0.65 (0.58-0.72)*	*23.84 (20.38-27.82)*
Transaortic gradient	*0.36 (0.29-0.42)*	*6.84 (6.25-7.45)*
TRgradient	*0.28 (0.16-0.39)*	*6.95 (6.51-7.46)*
Sinus of Valsalva Size	*0.62 (0.55-0.67)*	*0.24 (0.22-0.26)*
Ascending Aorta Size	*0.16 (0.02-0.45)*	*0.41 (0.29-0.63)*
IVC Diameter	*0.70 (0.65-0.75)*	*2.85 (2.65-3.06)*

This becomes quite relevant in their auxiliary metric of detecting STEMI (which probably should be spelled out at first mention for the medically uninitiated). It is unclear just what you mean by STEMI; is it an acute infarct or does a chronic regional wall motion abnormality count? Does the model understand the connection between wall motion and coronary territories?

Reply: This comparison is in the limited set of patients that had both ECGs and echocardiograms. Patients diagnosed with STEMI through conventional clinical definition (ST changes on ECG) were used to see if EchoPrime can identify these patients compared to patients with chronic WMA and other chronic diseases. We modify the main text to clarify that the labels were obtained from other modalities:
In **Methods** "**Disease Diagnosis with Probing**" section:

> ***Multimodal Disease Diagnosis***
> *We apply k-nearest neighbors (KNN) and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyhloidosis as two diseases for which echocardiography is not the definitive test*

echocardiography can find corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study, and used along with clinician-assigned diagnoses to fit the models. For KNN probing, we use KNeighborsClassifier, and for linear probing, we apply LogisticRegression, both from scikit-learn library.

I'm also confused by predicate for the STEMI (and amyloidosis), where you state that these are diagnoses "not typically diagnosed with echocardiography," as I would consider both of these to be "bread and butter" echo diagnoses. It also appears that you only considered the Ap4 view for STEMI, which would miss many RCA and LCx infarcts. Similarly, your focus on amyloid uses only the PLAx view, which would miss many of the key findings in amyloid (mid-septal stripe and apical strain sparing, for example).

Reply: Thank you for this comment. We sought to clarify that echocardiography is not the definitive test of choice for diagnosing STEMI or cardiac amyloidosis. ECGs, nuclear medicine, or MRI tests will confirm the diagnosis, while it is correct that echocardiography can find corroborating evidence. We clarify in this revision by replacing the phrase "non-echocardiographic diagnoses" with "multimodal cardiovascular diagnoses" and how this is an approach to use EchoPrime's pretrained weights for non-'echocardiographic' tasks.

In **Methods** "**Disease Diagnosis with Probing**" section:

Multimodal Disease Diagnosis
We apply k-nearest neighbors (KNN) and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyhloidosis as two diseases for which echocardiography is not the definitive test echocardiography can find corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study, and used along with clinician-assigned diagnoses to fit the models. For KNN probing, we use KNeighborsClassifier, and for linear probing, we apply LogisticRegression, both from scikit-learn library.

This model clearly is an important advance in AI-guided echo interpretation and, after appropriate revision, quite worthy of publication. One possible limitation is that the model does not actually write the interpretation language but, as I understand it, rather finds the best fitting verbiage from the 275K reports generated at Cedars Sinai over the last ten years (admittedly, these will cover most every nuance!).

Reply: Thank you for this important feedback. EchoPrime does not generate but retrieves the

report section-by-section. We see that as beneficial to minimize hallucinations and downstream implementation (as most echo labs use structured reporting and will require output in specific phrase formats to be placed into the clinical report).There are 15 sections and 275k candidate reports, meaning that EchoPrime picks one from 15^(275_000) candidate possibilities, which we believe covers a comprehensive and wide range of echocardiographic findings and interpretations.

Another critique is the choice of batch size = 32. The key point of contrastive learning is to bring like embeddings (text reports and their corresponding videos, i.e., the positive example) closer together while pushing negative pairs far apart. This necessitates a sufficiently large batch size to create enough differentiation within the latent embedding space. You trained on only 2 A6000's, raising the possibility of a limitation of computing power, which would warrant some further discussion and ideally ablation studies showing the impact of different batch sizes.

Reply: Thank you for making this point. In this revision, we show new experiments were we train EchoPrime with batch sizes of 16, 32 and 50, and show all batch sizes have similar performance with likely asymptoming gains around a batch size of 32 (original hyperparameter choice). While these ablations offer useful insights into batch size effects, the differences observed with batch size 50 are minimal and do not justify a switch in the primary model configuration.

Supplementary Table 10 *EchoPrime fine-tuned on different batch sizes*

BATCH SIZE	***16***	***32***	***50***
LV ejection fraction (R^2)	*0.81 (0.79-0.83)*	***0.83 (0.81-0.85)***	*0.80 (0.78-0.82)*
LV ejection fraction (MAE)	*4.87 (4.69-5.06)*	***4.79 (4.60-4.97)***	*4.94 (4.75-5.13)*
Pacemaker	*0.85 (0.82-0.87)*	*0.85 (0.82-0.88)*	***0.87 (0.84-0.89)***
RV systolic function	*0.93 (0.90-0.95)*	*0.93 (0.91-0.95)*	*0.93 (0.90-0.95)*
RV dilation	*0.88 (0.82-0.94)*	*0.89 (0.83-0.95)*	***0.92 (0.87-0.96)***
LA dilation	*0.92 (0.87-0.96)*	*0.91 (0.86-0.96)*	***0.94 (0.90-0.97)***
RA dilation	*0.94 (0.88-0.98)*	*0.90 (0.82-0.97)*	*0.94 (0.88-0.99)*
Mitraclip	*0.99 (0.99-1.00)*	*0.99 (0.99-1.00)*	*0.99 (0.99-1.00)*
Mitral annular calcification	*0.96 (0.94-0.97)*	*0.96 (0.95-0.97)*	*0.95 (0.92-0.97)*
Mitral stenosis	*0.91 (0.82-0.97)*	***0.96 (0.91-0.99)***	*0.92 (0.85-0.99)*
Mitral regurgitation	*0.93 (0.92-0.95)*	*0.92 (0.91-0.94)*	*0.93 (0.92-0.95)*
TAVR	*0.99 (0.98-1.00)*	*1.00 (0.99-1.00)*	*0.99 (0.98-1.00)*
Bicuspid AV	*0.83 (0.67-0.96)*	*0.83 (0.67-0.96)*	*0.83 (0.67-0.96)*
Aortic stenosis	*0.98 (0.96-0.99)*	*0.98 (0.96-0.99)*	*0.98 (0.96-0.99)*

Aortic regurgitation	*0.87 (0.82-0.92)*	*0.88 (0.83-0.93)*	***0.90 (0.85-0.94)***
Tricuspid regurgitation	*0.95 (0.94-0.97)*	*0.95 (0.93-0.97)*	***0.96 (0.94-0.97)***
Pericardial effusion	***0.99 (0.98-0.99)***	*0.98 (0.95-0.99)*	*0.97 (0.93-0.99)*
Aortic root dilation	***0.92 (0.83-0.99)***	*0.91 (0.83-0.98)*	*0.91 (0.82-0.98)*
Dilated IVC	*0.83 (0.81-0.85)*	***0.84 (0.82-0.86)***	*0.83 (0.81-0.85)*
PA pressure (R^2)	*0.42 (0.37-0.46)*	***0.43 (0.39-0.47)***	*0.42 (0.38-0.47)*
PA pressure (MAE)	*7.31 (6.94-7.68)*	*7.30 (6.95-7.67)*	***7.26 (6.89-7.62)***

A potential critique of multiple instance learning is the risk that your inference pipeline is over-engineered, carrying strong inductive bias with a classifier and the weighting objective. Have you experimented at all with learning the anatomical priors outside of a two-step system, i.e. a learned pooling of embeddings to create the study and anatomical embeddings?

Reply: Thank you for this important point. In Figure 2A and newly added Supplementary Table 9 we directly compare the performance with and without anatomical attention and show the approach with anatomical attention is superior. The trained recognition of important features from the most relevant videos (anatomic attention) is important and helpful, however there are many approaches that might be proposed for pooling and ensembling approaches. In our discussion, we describe this as a direction for future investigation.

In **Results "Leveraging multi-view echocardiographic data improves performance**" section:

To demonstrated the impact of integrating multiple clips, multiple videos, and multiple views, we compared the performance of the EchoPrime encoder when input is a single frame vs. single video vs. multiple videos vs. multiple videos with multiple instance attention weighting (**Error! Reference source not found.**A). The gradual improvement across multiple tasks reflects the relevance of temporal and view-specific information to each task and the importance of the anatomic attention module (Supplementary Table 9).

In **Discussion**:

> Additional work is required to continue advancing medical AI models, including the development of multimodal models that integrate complementary diagnostic modalities and recognize task-relevant features as well as clinical trials of AI in a clinical setting.

Supplementary Table 9 EchoPrime performance with and without anatomical weighting

	WITHOUT	**WITH**	**STANFORD W/O**	**STANFORD WITH**
LV ejection fraction (R^2)	0.80 (0.78-0.82)	**0.83 (0.81-0.85)**	0.78 (0.75-0.81)	0.79 (0.76-0.82)

LV ejection fraction (MAE)	4.97 (4.78-5.17)	**4.79 (4.60-4.97)**	4.22 (4.01-4.42)	4.14 (3.94-4.34)
Pacemaker	0.82 (0.80-0.85)	**0.85 (0.82-0.88)**	0.80 (0.77-0.83)	0.84 (0.81-0.87)
RV systolic function	0.92 (0.90-0.94)	**0.93 (0.91-0.95)**	0.91 (0.80-1.00)	0.94 (0.88-1.00)
RV dilation	**0.90 (0.85-0.95)**	0.89 (0.83-0.95)	0.83 (0.77-0.88)	0.85 (0.80-0.90)
LA dilation	0.91 (0.86-0.96)	0.91 (0.86-0.96)	0.71 (0.68-0.73)	0.73 (0.70-0.75)
RA dilation	**0.92 (0.85-0.98)**	0.90 (0.82-0.97)	0.76 (0.73-0.80)	0.77 (0.74-0.81)
Mitraclip	0.98 (0.96-0.99)	**0.99 (0.99-1.00)**	0.97 (0.93-1.00)	0.98 (0.94-1.00)
Mitral annular calcification	0.95 (0.93-0.96)	**0.96 (0.95-0.97)**	0.96 (0.93-0.98)	0.96 (0.94-0.98)
Mitral stenosis	0.96 (0.90-0.99)	0.96 (0.91-0.99)	0.93 (0.86-0.98)	0.92 (0.86-0.98)
Mitral regurgitation	0.91 (0.89-0.93)	**0.92 (0.91-0.94)**	0.89 (0.85-0.93)	0.91 (0.89-0.93)
TAVR	0.99 (0.99-1.00)	**1.00 (0.99-1.00)**	0.99 (0.98-1.00)	0.97 (0.90-1.00)
Bicuspid AV	0.82 (0.67-0.96)	**0.83 (0.67-0.96)**	0.78 (0.69-0.87)	0.82 (0.73-0.90)
Aortic stenosis	0.96 (0.93-0.98)	**0.98 (0.96-0.99)**	0.94 (0.89-0.99)	0.96 (0.92-0.99)
Aortic regurgitation	0.83 (0.77-0.88)	**0.88 (0.83-0.93)**	0.84 (0.72-0.94)	0.89 (0.80-0.97)
Tricuspid regurgitation	0.94 (0.91-0.96)	**0.95 (0.93-0.97)**	0.84 (0.79-0.88)	0.88 (0.86-0.91)
Pericardial effusion	0.95 (0.92-0.98)	**0.98 (0.95-0.99)**	0.88 (0.75-0.98)	0.89 (0.77-0.98)
Aortic root dilation	0.88 (0.79-0.96)	**0.91 (0.83-0.98)**	0.94 (0.90-0.98)	0.97 (0.94-0.99)
Dilated IVC	0.82 (0.80-0.84)	**0.84 (0.82-0.86)**	0.82 (0.74-0.88)	0.86 (0.81-0.91)
PA pressure (R^2)	0.41 (0.37-0.45)	**0.43 (0.39-0.47)**	0.33 (0.27-0.39)	0.36 (0.30-0.42)
PA pressure (MAE)	7.44 (7.10-7.82)	**7.30 (6.95-7.67)**	7.89 (7.21-8.60)	7.97 (7.35-8.66)

Retrieval Augmented Interpretation (page 14, par 2): I'm not sure why this was described uniquely in comparison to Retrieval Augmented Generation. If Retrieval Augmented Interpretation has been described elsewhere this should be cited in this section. Otherwise,

Retrieval Augmented Interpretation seems to describe how CLIP models are used as zero shot __ (where __ is a regression task, classifier, or 'report generator'). Probe text is encoded using the text encoder and a similarity metric is used to retrieve the most relevant text based on the similarity between the video and text encodings. So, a description of the 'historical' reports that are retrieved is necessary here. Are these reports all from the training set or a different set of reports?

Reply:
We note the difference between RAI and RAG because RAG is traditionally connected to large language models. RAI is also different than traditional CLIP retrieval because it retrieves the report section-by-section. The historical reports are all reports from the training set. We modify the text to clarify the source of candidate reports:

*In **Methods** "**Retrieval Augmented Interpretation**" section*

> *Using pre-trained parametric memory (EchoPrime weights) and non-parametric memory (clinical reports from the train set), our approach leverages anatomic attention to weight the cross-modal retrieval based on each section in an echocardiogram report.*

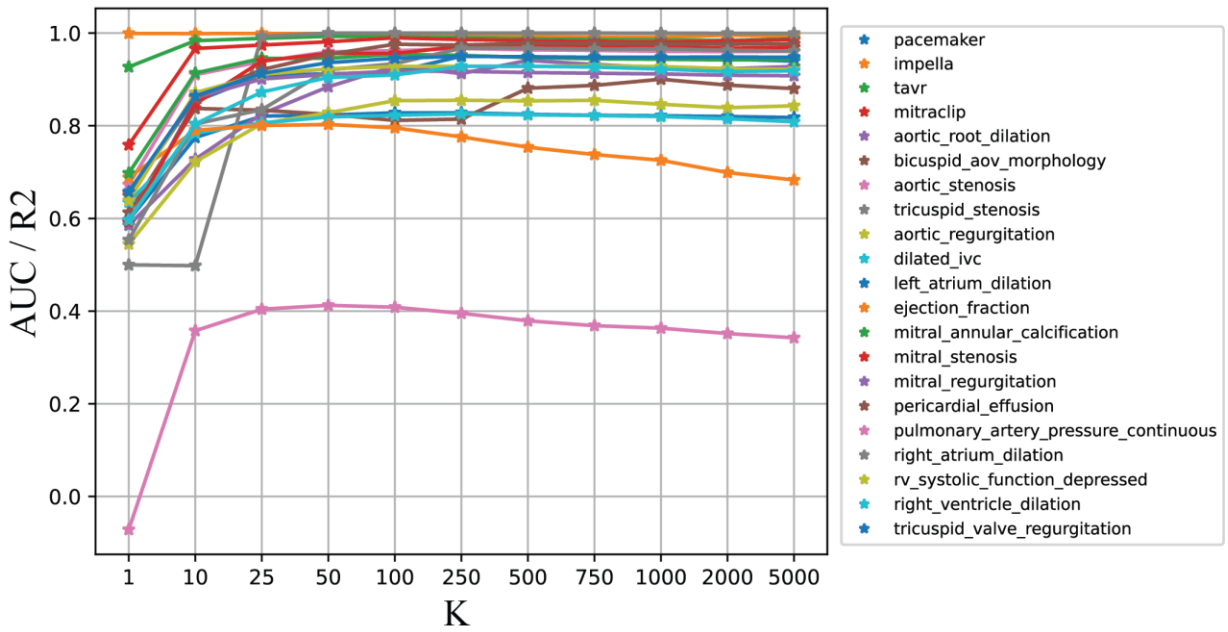
In addition, you average 50 retrieved reports to generate the interpretation. Evidence for k=50 reports being optimal? Also, what is meant by averaging since this is at the text level?

Reply:

We add the plot (Supplementary Figure 4) showing that k=50 is optimal hyperparameter including considerations of performance and runtime. Specifically, for values k>50 the performance doesn't improve but the runtime increases, that is why k=50 was chosen.
The averaging is happening at the feature level. Features are extracted with regex and custom defined phrases (https://github.com/echonet/EchoPrime/blob/main/per_section.json).

*In **Methods "Retrieval Augmented Interpretation"** section*

> *Hyperparameter k=50 is chosen because it achieves the optimal trade-off between performance and runtime (Supplementary Figure 4).*



Supplementary Figure 4 *Performance of Retrieval Based Interpretation as we increase the number of candidates reports to average across. For binary tasks the metric is AUROC and for regression tasks R2 score.*

Along the same lines there is no discussion as to why mViT was chosen as the video encoder and no ablations for other video encoding architectures. Additionally, the same can be said about using PubMedBERT for the text encoder. In fact, in the previous iteration, EchoCLIP dedicated significant discussion space to describing a unique tokenization technique that took advantage of the templated structure of Echo reports. This was shown to be better than standard Byte Pair Encoding with the CLIP text encoder. I think the switch to PubMedBERT and Wordpiece encoding warrants deeper discussion and some comparison studies with the EchoCLIP text encoder and the text encoder used in EchoPrime.

Reply: Thank you for your insightful comments. Due to the significant computational demands, we were unable to conduct extensive ablation studies on alternative video and text encoder architectures. For context, training our main checkpoint required approximately five weeks on our compute cluster. Our primary focus was on demonstrating EchoPrime's practical accuracy and its superior performance compared to prior biomedical foundation models. For instance, the comparison of EchoPrime, a model that uses PubMedBERT, and EchoCLIP-R, a model that uses custom BPE, is presented in Table 3. While additional architectural ablations could provide further insights, we believe that EchoPrime's end-to-end improvements already establish its effectiveness over previous approaches.

As per reviewer request me motivate our architectural choices.

*In **Methods "Contrastive Visual-Language Training"** section:*

> *mVIT was chosen as the video encoder because, in the original CLIP paper[4], a transformer-based encoder outperformed a ResNet-based encoder. Additionally, mVIT effectively captures temporal dynamics, as evidenced by performance changes when video frames are shuffled—a phenomenon not observed in other vision transformers[30]. PubMedBERT was selected as the text encoder due to its pretraining on 21GB of medical text, giving our model a head start in understanding medical literature.*

Table 1: Cross-modal retrieval metrics on a test set of 2,254 videos-text report pairs. Lower rank is better, and higher recall is better. Best performance is bolded

	Videos-To-Text					Text-To-Videos				
	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10
EchoPrime	**2.86**	**1**	**70%**	**94%**	**98%**	**3.05**	**1**	**69%**	**94%**	**97%**
EchoCLIP	58.71	9	20%	42%	53%	37.07	5	25%	50%	62%

I would note that there is break in the text continuity between page 5 and 6. It is unclear if there is any lost text here (I suspect not), but it should be corrected.

Reply: Thank you for noticing this. It is indeed a spacing issue after the parentheses, we fix it in the revision.

The authors have provided model code with all initialization parameters at their github. While I have not implemented this for my own testing, it does appear that it is in standard form and should be readily implemented with the appropriate subroutine libraries, which they specified. In addition, you provide a video example of EchoPrime in operation, interpreting a relatively normal echo. This overall is effective in showing model operation, but I am puzzled by the lack of any spectral Doppler frames in the example data. It is simply impossible to estimate RV pressure, diastolic function, or valve stenosis without Doppler, so this should be redone with the appropriate needed images. It should be noted that you have not provided the echo training data, but this seems reasonable. Not only would this likely exceed a petabyte in size, but despite all best efforts to strip identifying information from the images, such PHI might still be present. Of note, you do provide a sample casein DICOM format, though again I only see video clips, no M-mode nor spectral Doppler. Unclear if this is the same case shown in the model operation example. You certainly need to include the single frame images and specify if this is the case shown in the MOV file.

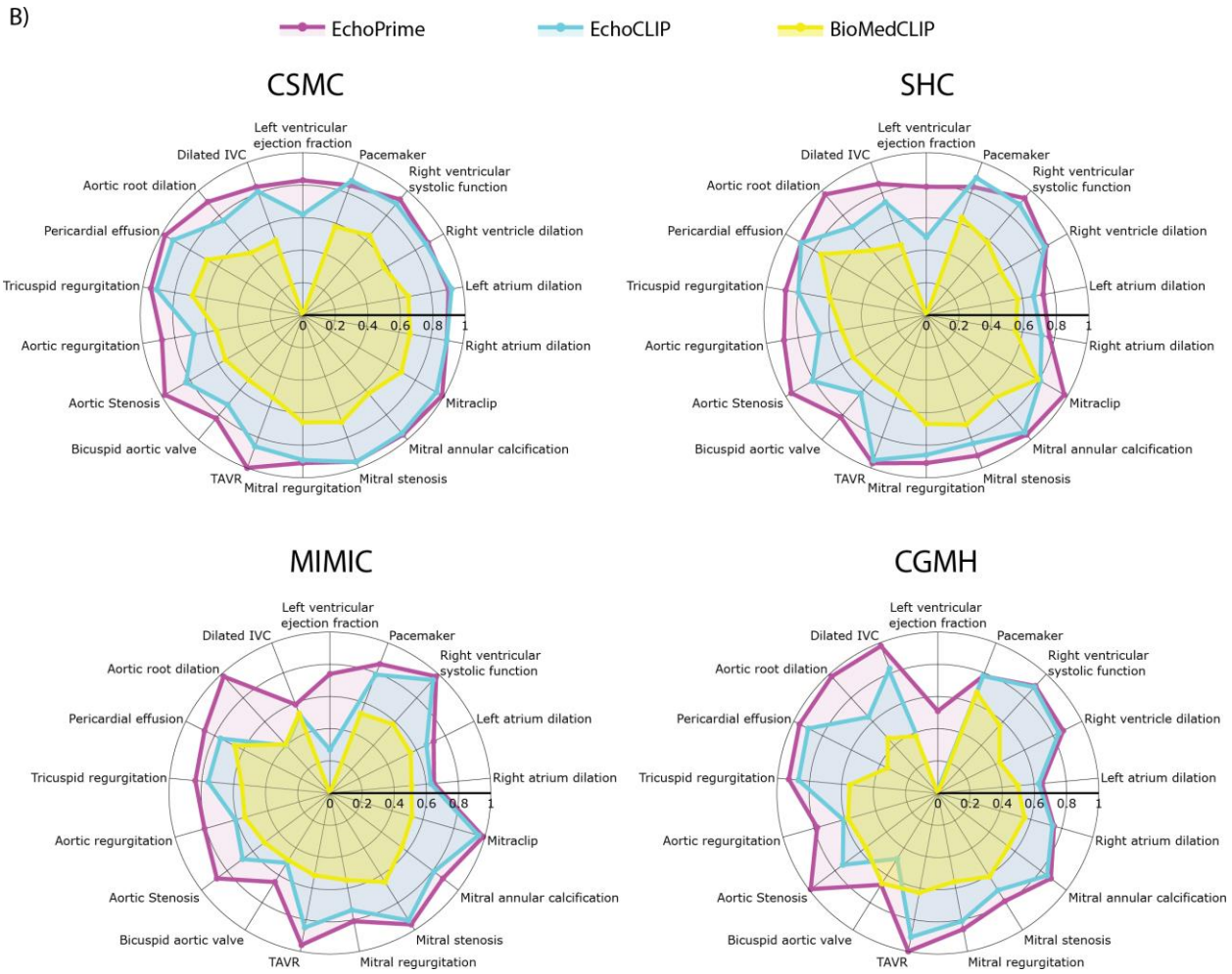
Reply: EchoPrime was trained on B-mode and Doppler videos and used a video-based encoder. Much/all of the Spectral doppler and M mode images information images are obtainable from

standard echocardiographic videos from the same study. For example, the M mode has time in the x axis and pixel intensity in the y, however all that necessary information is inherently already present in the B-mode video. Similarly, the Spectral doppler information is already present in the color Doppler videos and we hypothesize can be extracted with a video encoder from other videos in the study.

To confirm this finding, in this revision, we introduce two new external validation sets (echocardiogram studies from MIMIC and CGMH which have B-mode, Doppler, Spectral doppler, and M mode images) and show EchoPrime performs similarly in these datasets as our other internal and external test datasets even with ignoring Spectral doppler and M mode images.

*In **Results "Automated echocardiogram interpretation without supervised learning" section***

> *On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging (Figure2B, Supplementary Tables 2-5)*



Supplementary Table 4 *Performance of foundation models for biomedical imaging on MIMIC dataset from Boston*

MIMIC DATASET			
	BioMed CLIP	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-2.48 (-3.26--1.85)*	*0.27 (0.13-0.40)*	***0.74 (0.66-0.80)***
LV ejection fraction (MAE)	*26.96 (25.31-28.58)*	*10.37 (9.30-11.51)*	***6.21 (5.54-6.92)***
Pacemaker	*0.53 (0.39-0.68)*	*0.79 (0.66-0.90)*	***0.86 (0.80-0.92)***
RV systolic function	*0.58 (0.02-0.90)*	*0.95 (0.89-0.99)*	***0.98 (0.97-0.99)***
RV dilation	*-*	*-*	*-*
LA dilation	*0.56 (0.52-0.59)*	*0.67 (0.64-0.71)*	***0.72 (0.68-0.75)***
RA dilation	*0.51 (0.47-0.54)*	*0.63 (0.59-0.67)*	***0.65 (0.61-0.69)***
Mitraclip	*0.53 (0.29-0.99)*	*0.96 (0.86-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.56 (0.51-0.60)*	*0.81 (0.77-0.85)*	***0.88 (0.85-0.91)***
Mitral stenosis	*0.65 (0.48-0.98)*	*0.93 (0.80-1.00)*	***0.96 (0.95-0.98)***
Mitral regurgitation	*0.55 (0.52-0.58)*	*0.74 (0.71-0.76)*	***0.81 (0.79-0.83)***
TAVR	*0.52 (0.44-0.60)*	*0.85 (0.79-0.90)*	***0.96 (0.94-0.98)***
Bicuspid AV	*0.49 (0.46-0.56)*	*0.51 (0.42-0.61)*	***0.65 (0.52-0.78)***
Aortic stenosis	*0.51 (0.48-0.54)*	*0.68 (0.64-0.73)*	***0.88 (0.84-0.91)***
Aortic regurgitation	*0.55 (0.49-0.61)*	*0.61 (0.54-0.67)*	***0.81 (0.75-0.86)***
Tricuspid regurgitation	*0.55 (0.52-0.58)*	*0.76 (0.73-0.78)*	***0.84 (0.82-0.86)***
Pericardial effusion	*0.66 (0.61-0.72)*	*0.76 (0.70-0.81)*	***0.87 (0.83-0.92)***
Aortic root dilation	*0.41 (0.40-0.42)*	*0.40 (0.39-0.40)*	***0.98 (0.96-0.99)***
Dilated IVC	*0.53 (0.44-0.63)*	*0.52 (0.40-0.64)*	***0.59 (0.49-0.69)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Supplementary Table 5 *Performance of foundation models for biomedical imaging on CGMH dataset from Taiwan*

CGMH DATASET

	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	*-8.31 (-9.74--7.06)*	*0.02 (-0.13-0.15)*	***0.51 (0.44-0.57)***
LV ejection fraction (MAE)	*32.64 (31.89-33.38)*	*9.15 (8.75-9.54)*	***6.46 (6.19-6.74)***
Pacemaker	*0.67 (0.56-0.80)*	***0.78 (0.59-0.92)***	***0.78 (0.63-0.92)***
RV systolic function	*0.57 (0.41-0.72)*	*0.89 (0.84-0.94)*	***0.90 (0.79-0.98)***
RV dilation	*0.43 (0.30-0.61)*	*0.84 (0.65-0.97)*	***0.87 (0.69-0.97)***
LA dilation	*0.50 (0.47-0.53)*	*0.63 (0.60-0.65)*	***0.65 (0.62-0.67)***
RA dilation	*0.56 (0.47-0.64)*	*0.74 (0.65-0.83)*	***0.75 (0.67-0.84)***
Mitraclip	*-*	*-*	*-*
Mitral annular calcification	*0.55 (0.40-0.70)*	*0.85 (0.69-0.97)*	***0.88 (0.75-0.96)***
Mitral stenosis	*0.61 (0.48-0.79)*	*0.71 (0.51-0.92)*	***0.79 (0.59-0.98)***
Mitral regurgitation	*0.56 (0.50-0.63)*	*0.81 (0.76-0.86)*	***0.86 (0.82-0.89)***
TAVR	*0.63 (0.50-0.77)*	*0.91 (0.83-0.97)*	***1.00 (1.00-1.00)***
Bicuspid AV	*0.66 (0.43-1.00)*	*0.48 (0.35-0.77)*	***0.67 (0.45-1.00)***
Aortic stenosis	*0.56 (0.43-0.71)*	*0.74 (0.59-0.88)*	***0.99 (0.98-1.00)***
Aortic regurgitation	*0.58 (0.50-0.66)*	*0.61 (0.53-0.69)*	***0.78 (0.70-0.84)***
Tricuspid regurgitation	*0.55 (0.49-0.61)*	*0.87 (0.83-0.91)*	***0.93 (0.90-0.95)***
Pericardial effusion	*0.35 (0.16-0.65)*	*0.90 (0.69-1.00)*	***0.96 (0.91-1.00)***
Aortic root dilation	*0.46 (0.39-0.62)*	*0.64 (0.41-0.86)*	***0.98 (0.97-1.00)***
Dilated IVC	*0.38 (0.37-0.39)*	*0.83 (0.50-1.00)*	***0.98 (0.94-1.00)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Referee #3 (Remarks to the Author):

A. Summary of Key Results
The authors propose a novel vision-language model trained on 12 million samples to interpret echocardiography studies by performing view classification, mapping videos into a video-text latent space, and retrieving study interpretations. Experiments on both internal and external datasets demonstrate that EchoPrime outperforms both task-specific approaches and prior foundation models. The authors introduce EchoPrime, which incorporates multi-view information to enhance the accuracy of echocardiography interpretations—an innovative approach. Additionally, they collected an extensive dataset of 12 million samples, highlighting the scale of their work.

Reply: Thank you for your feedback. EchoPrime is the largest AI model in echocardiography in terms of the dataset size used for training. It was trained on an unparalleled dataset of 12 million echocardiogram videos, synthesizing all videos within each study and their corresponding 100–500-word reports, and this scale of training data helps explain its unparalleled performance on a wide range of clinical interpretation tasks.

B. Originality and Significance
However, although the authors claim EchoPrime can assist physicians in the automated preliminary assessment of comprehensive echocardiography after rigorous clinical evaluation, no human evaluation has been conducted; the assessment relies solely on metrics such as AUC, limiting the study's significance. While the dataset used is comprehensive, the analysis is purely retrospective. Prospective experiments would align more directly with clinical applications, providing a more robust validation of EchoPrime's effectiveness in real-world clinical settings. The lack of prospective analysis significantly limits the study's overall validity and impact.

Reply: Thank you for this important consideration. In this model development work, we have not undergone clinical trial assessment of EchoPrime. That is currently ongoing, and we are excited to present these results likely in late 2025 / early 2026.

Despite this limitation, we sought to comprehensively assess EchoPrime's performance in a wide range of tasks and settings. In this revision, we introduce two new external validation sites (one from Boston and one from Taiwan) to show international validation on a wide range of potential image types/characteristics and patient populations. We think this is a rigorous assessment of generalizability of EchoPrime across many clinical settings and practice patterns, however we recognize that this is still a retrospective evaluation.

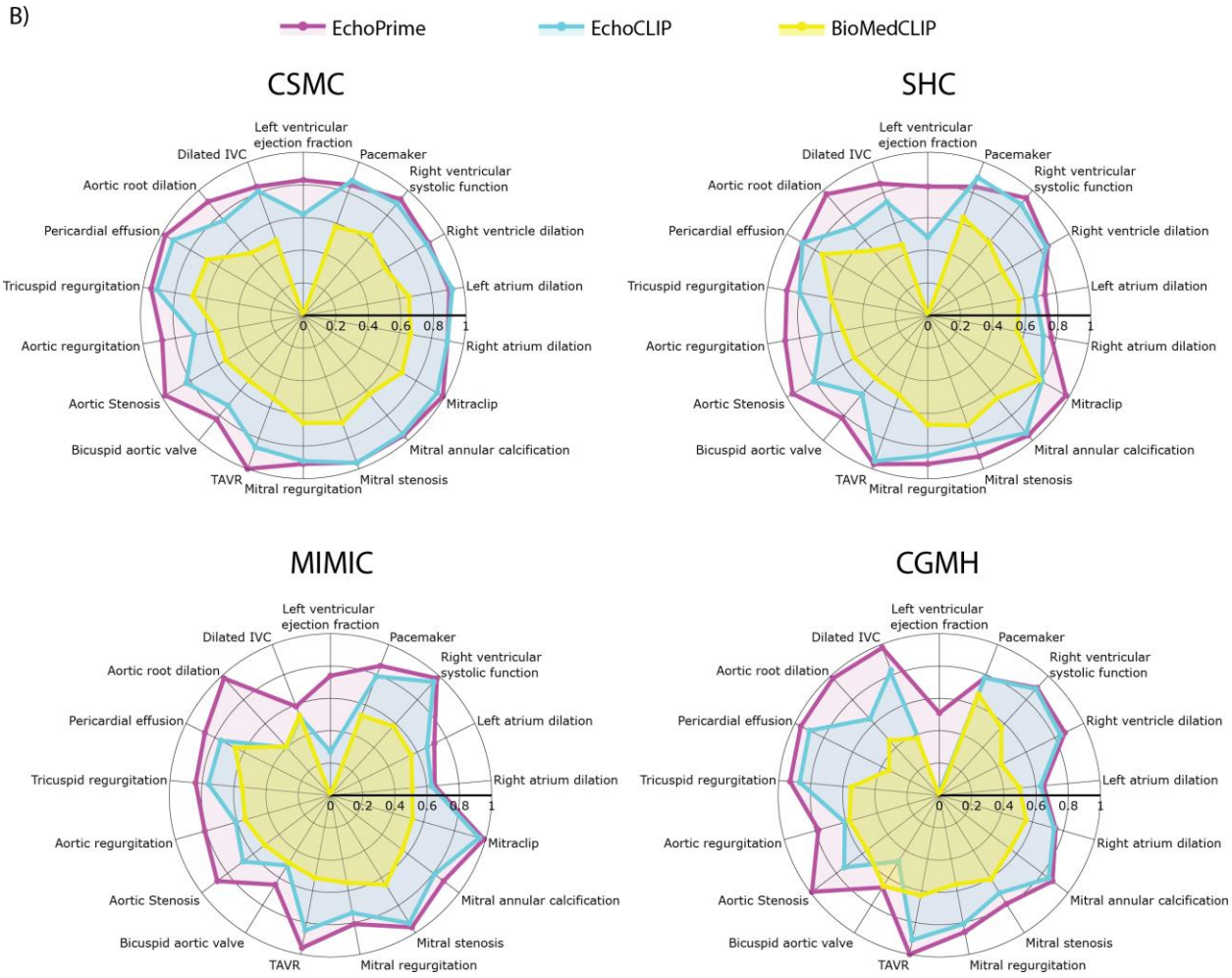


Figure 2B

In this revision, we provide more discussion of this critical point in the discussion portion of the manuscript, highlighting the need for prospective evaluation.

*In the **Discussion**:*

> *A few limitations are worth considering. Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations under diverse and variable image acquisition protocols to ensure robust and generalizable performance and to measure its practical value to physicians. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information.*

C. Data & Methodology

Video Encoder: The authors state that their video encoder is trained on a mViT checkpoint;

however, mViT dates back to 2021. Recent advancements such as BEVT [1] and VideoMAE [2] could enhance relevance and demonstrate a thorough exploration of state-of-the-art methods.

Reply: Thank you for your comment. BEVT and VideoMAE propose strategies for training foundation models using unsupervised masking, which is particularly beneficial when only videos are available. In our project, however, we have access to both videos and their corresponding reports. To leverage the additional information from the reports, we adopt a contrastive learning approach to build a joint video-text representation space.

Foundation models in the video-language domain generally fall into three categories: contrastive, masking, and generative models [1]. EchoPrime belongs to the contrastive family, and while comparisons across these three families are undoubtedly valuable, such an analysis is beyond the scope of this paper. Our focus is on demonstrating the effectiveness of the contrastive approach for this specific application. With the design choices we have implemented, EchoPrime has superior performance across a wide range of tasks and exceeds or parallels all publicly available echocardiographic AI models. Further work will continue to improve upon this model as the AI space is rapidly evolving, but we believe this is a important milestone in AI for echocardiography.

In our discussion, we now note the importance of new models that might be interesting avenues for future investigation.

[1.] Bordes, F. *et al.* An Introduction to Vision-Language Modeling. Preprint at http://arxiv.org/abs/2405.17247 (2024).

*In the **Discussion**:*

> *A few limitations are worth considering. Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations under diverse and variable image acquisition protocols to ensure robust and generalizable performance and to measure its practical value to physicians. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information.*

Text Encoder: EchoPrime supports over 500 tokens and utilizes WordPiece encoding within a BERT architecture. The authors state, “This marks a significant improvement over previous medical foundation models like BioMedCLIP and EchoCLIP, which process only single-view, individual images and handle text up to 77 tokens.” However, no ablation studies on token length are provided, which are necessary to validate the benefit of increasing the token limit. Comparisons with other models alone do not substantiate that this design choice is significantly advantageous.

Reply: Thank you for your comment. The main motivation behind extending the context length of the text encoder is to be able to fit full echocardiography reports during training. For example,

the following report requires 404 tokens by Wordpiece tokenizer, but would be truncated by CLIP's encoder (which only allows for 77 tokens):

What EchoPrime sees:

'Aorta: The aortic root is normal in size. Sinus of Valsalva: 3.3 cm. The ascending aorta is normal in size. Ascending Aorta 3.5 cm. Aortic Valve: No significant aortic stenosis or insufficiency. Trileaflet aortic valve. The aortic cusps appear mildly thickened. The peak transaortic gradient is 5 mmHg The mean transaortic gradient is 3 mmHg Atrial Septum: The interatrial septum is normal in appearance. IVC: The inferior vena cava is of normal size. The IVC diameter is 18 mm The inferior vena cava shows a normal respiratory collapse consistent with normal right atrial pressure (3 mmHg). Left Atrium: The left atrium is normal in size. Left Ventricle: Normal left ventricular size by linear cavity dimension. Normal left ventricular size by volume Mild left ventricular hypertrophy. Normal left ventricular systolic function. LV Ejection Fraction is 60 %. Mild diastolic dysfunction. There is reversal of the E to A ratio and/or prolonged deceleration time consistent with impaired left ventricular relaxation. Doppler parameters and/or lateral mitral annular (E`) velocities are consistent with normal left ventricular filling pressures. Mitral Valve: The mitral valve demonstrates normal leaflet morphology. The mitral valve demonstrates normal function with trace physiologic regurgitation. Pericardium: Normal pericardium with no pericardial effusion. Pulmonary Artery: Estimated PA Pressure is 16 mmHg. PA systolic pressure is normal. Pulmonary Veins: Pulmonary veins are normal in appearance and pulse Doppler interrogation shows normal systolic predominant flow. Pulmonic Valve: Normal pulmonic valve appearance. Normal pulmonic valve function with trace physiologic regurgitation. Resting Segmental Wall Motion Analysis: Total wall motion score is 1.00. There are no regional wall motion abnormalities Right Atrium: The right atrium is normal in size. Right Ventricle: Normal right ventricular size. Normal right ventricular systolic function. Tricuspid Valve: Normal appearance of the tricuspid valve. Est RV/RA pressure gradient is 13 mmHg. Estimated peak RVSP is 16 mmHg. There is trivial tricuspid regurgitation.'

What EchoCLIP sees:

<|startoftext|>aorta : the aortic root is normal in size . sinus of valsalva : 3 . 3 cm . the ascending aorta is normal in size . ascending aorta 3 . 5 cm . aortic valve : no significant aortic stenosis or insufficiency . trileaflet aortic valve . the aortic cusps appear mildly thickened . the

The ability to process the full echocardiography reports is a fundamental aspect of EchoPrime's design, allowing it to leverage the complete context provided by the detailed reports during training, while a much longer context is not useful if reports are too short to utilize a much longer context. While we appreciate the suggestion for ablation studies on token length, this has significant computational cost and the choice of context length was rationaly designed based on the ability to accommodating full reports. Truncating reports, as required by models with limited token capacity like CLIP, likely results in loss of performance.

Method: The evaluation of EchoPrime lacks comprehensiveness. While the authors selected two foundation models, BioMedCLIP and EchoCLIP, additional comparisons, such as with PubMedCLIP [3], PMC-CLIP [4], and ROCO [5], would provide a broader context.

Reply: Thank you for your feedback. In the original manuscript, we chose benchmarks that have highest likelihood of the strongest performance (BioMedCLIP has the strongest performance among foundation models trained on publicly available data and EchoCLIP is the most comparable echocardiographic foundation model). In this revision, we had the recommended comparison models (PubMedCLIP and PMC-CLIP) and show that they perform worse than EchoPrime (**Supplementary Tables 2)**. Of note, ROCO. [5] doesn't introduce a new machine learning model but rather is a dataset that can be used for training contrastive models (such as

PubMedCLIP and PMC-CLIP).

The performance of general medical data foundation models is poor on echocardiography tasks (likely due to the small size of echocardiography videos present in their training set). Specifically, PubMedCLIP achieves an average of 0.50 across AUROC metrics, which is essentially random performance, and PMC-CLIP performs slightly better than random with 0.532 average across AUROC metrics. BioMedCLIP achieves an average of 0.608 across AUROC metrics but still far worse than EchoCLIP (average AUC =0.866) and EchoPrime (average AUC = 0.922), foundation models trained specifically for echocardiography.

In the ***Results*** *"****Automated echocardiogram interpretation without supervised learning****"*

> *We evaluated EchoPrime on a diverse set of benchmarks for cardiac structure and function across datasets from four different institutions (****Error! Reference source not found.****). On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging [17,19–21] (Figure 2B,* ***Supplementary Tables 2****-5),*

Supplementary Table 2 *Performance of foundation models for biomedical imaging on Cedars-Sinai test set*

CEDARS-SINAI TEST SET					
	PubMedCLIP	*PMC-CLIP*	*BioMed CLIP*	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-0.32 (-0.40--0.25)*	*-0.38 (-0.46--0.31)*	*-2.60 (-3.00 - -2.20)*	*0.62 (0.59-0.65)*	***0.83 (0.81-0.85)***
LV ejection fraction (MAE)	*14.68 (14.30-15.06)*	*14.03 (13.58-14.49)*	*26.93 (26.42-27.44)*	*7.00 (6.74-7.26)*	***4.79 (4.60-4.97)***
Pacemaker	*0.50 (0.50-0.50)*	*0.51 (0.47-0.54)*	*0.58 (0.55-0.62)*	***0.88 (0.85-0.91)***	*0.85 (0.82-0.88)*
RV systolic function	*0.50 (0.49-0.51)*	*0.53 (0.48-0.58)*	*0.64 (0.59-0.69)*	*0.89 (0.86-0.92)*	***0.93 (0.91-0.95)***
RV dilation	*0.50 (0.45-0.56)*	*0.59 (0.50-0.67)*	*0.58 (0.52-0.64)*	*0.88 (0.83-0.92)*	***0.89 (0.83-0.95)***
LA dilation	*0.50 (0.50-0.50)*	*0.57 (0.48-0.65)*	*0.66 (0.58-0.74)*	***0.93 (0.89-0.96)***	*0.91 (0.86-0.96)*
RA dilation	*0.50 (0.50-0.50)*	*0.58 (0.47-0.68)*	*0.67 (0.59-0.75)*	*0.90 (0.82-0.97)*	***0.90 (0.82-0.97)***
Mitraclip	*0.50 (0.50-0.50)*	*0.49 (0.43-0.55)*	*0.70 (0.63-0.76)*	*0.95 (0.92-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.50 (0.46-0.55)*	*0.59 (0.54-0.64)*	*0.63 (0.58-0.68)*	*0.95 (0.93-0.96)*	***0.96 (0.95-0.97)***
Mitral stenosis	*0.50 (0.50-0.50)*	*0.55 (0.44-0.66)*	*0.70 (0.61-0.79)*	***0.96 (0.91-0.99)***	***0.96 (0.91-0.99)***
Mitral regurgitation	*0.50 (0.50-0.50)*	*0.59 (0.55-0.63)*	*0.66 (0.63-0.69)*	*0.89 (0.87-0.91)*	***0.92 (0.91-0.94)***
TAVR	*0.50 (0.50-0.51)*	*0.51 (0.46-0.55)*	*0.54 (0.50-0.57)*	*0.86 (0.83-0.89)*	***1.00 (0.99-1.00)***
Bicuspid AV	*0.50 (0.50-0.50)*	*0.51 (0.47-0.60)*	*0.52 (0.51-0.53)*	*0.72 (0.58-0.85)*	***0.83 (0.67-0.96)***

Aortic stenosis	*0.50 (0.50-0.52)*	*0.48 (0.42-0.55)*	*0.55 (0.53-0.57)*	*0.83 (0.79-0.87)*	***0.98 (0.96-0.99)***
Aortic regurgitation	*0.50 (0.50-0.50)*	*0.50 (0.41-0.58)*	*0.54 (0.53-0.55)*	*0.68 (0.60-0.76)*	***0.88 (0.83-0.93)***
Tricuspid regurgitation	*0.49 (0.45-0.53)*	*0.57 (0.51-0.63)*	*0.69 (0.64-0.74)*	*0.91(0.89-0.93)*	***0.95 (0.93-0.97)***
Pericardial effusion	*0.54 (0.45-0.62)*	*0.50 (0.42-0.57)*	*0.68 (0.59-0.77)*	*0.92 (0.89-0.95)*	***0.98 (0.95-0.99)***
Aortic root dilation	*0.50 (0.50-0.50)*	*0.44 (0.35-0.55)*	*0.50 (0.49-0.51)*	*0.76 (0.63-0.88)*	***0.91 (0.83-0.98)***
Dilated IVC	*0.50 (0.50-0.51)*	*0.53 (0.50-0.56)*	*0.49 (0.47-0.50)*	*0.81 (0.79-0.83)*	***0.84 (0.82-0.86)***
PA pressure (R^2)	*-0.59 (-0.74--0.46)*	*-0.83 (-1.07--0.62)*	*-0.19 (-0.24 - -0.14)*	*0.35 (0.29-0.41)*	***0.43 (0.39-0.47)***
PA pressure (MAE)	*11.27 (9.98-12.62)*	*13.65 (13.04-14.28*	*10.62 (10.12-11.12)*	*7.84 (7.45- 8.23)*	***7.30 (6.95-7.67)***

Data: For the view classification task, the origin of the training dataset is unclear. Is this dataset part of the 12,124,168 echocardiography videos from CSMC as listed in Table 1? If not, please specify the data source and provide statistical details for this specific dataset.

Reply: Thank you for pointing this out. The view classification dataset consists of 77,426 echocardiograms, which is indeed a subset of the 12,124,168 training videos. This subset manually labeled by sonographers into 58 different view categories

We modify the text to be more specific about the source of the view classification dataset.

In the methods:

> We first trained a view classifier to provide a fine-grained characterization of echocardiogram views as defined by the American Society of Echocardiography guidelines[36]. We selected 77,426 echocardiogram videos, a subset of videos from the training set, for cardiac sonographers to label into 58 different view categories and trained an image-based ConvNextBase[37] view classification model on 224 x 224 images.

Presentation: The metric represented in Figure 2b is not specified. Including a description in the figure caption would clarify this. Furthermore, providing figures for data statistics would enhance understanding of the dataset's distribution and characteristics.

Reply: We clarify the metrics used in Figure 2B.

> ***(B)*** *Comparison of EchoPrime, EchoCLIP, and BioMedCLIP on a range of echocardiography interpretation tasks on cohorts from four different clinical centers. AUROC is shown for classification tasks and R2 score for regression tasks*

Figure 1A visualizes characteristics of the training dataset. Table 1 shows detailed statistics of each dataset used.

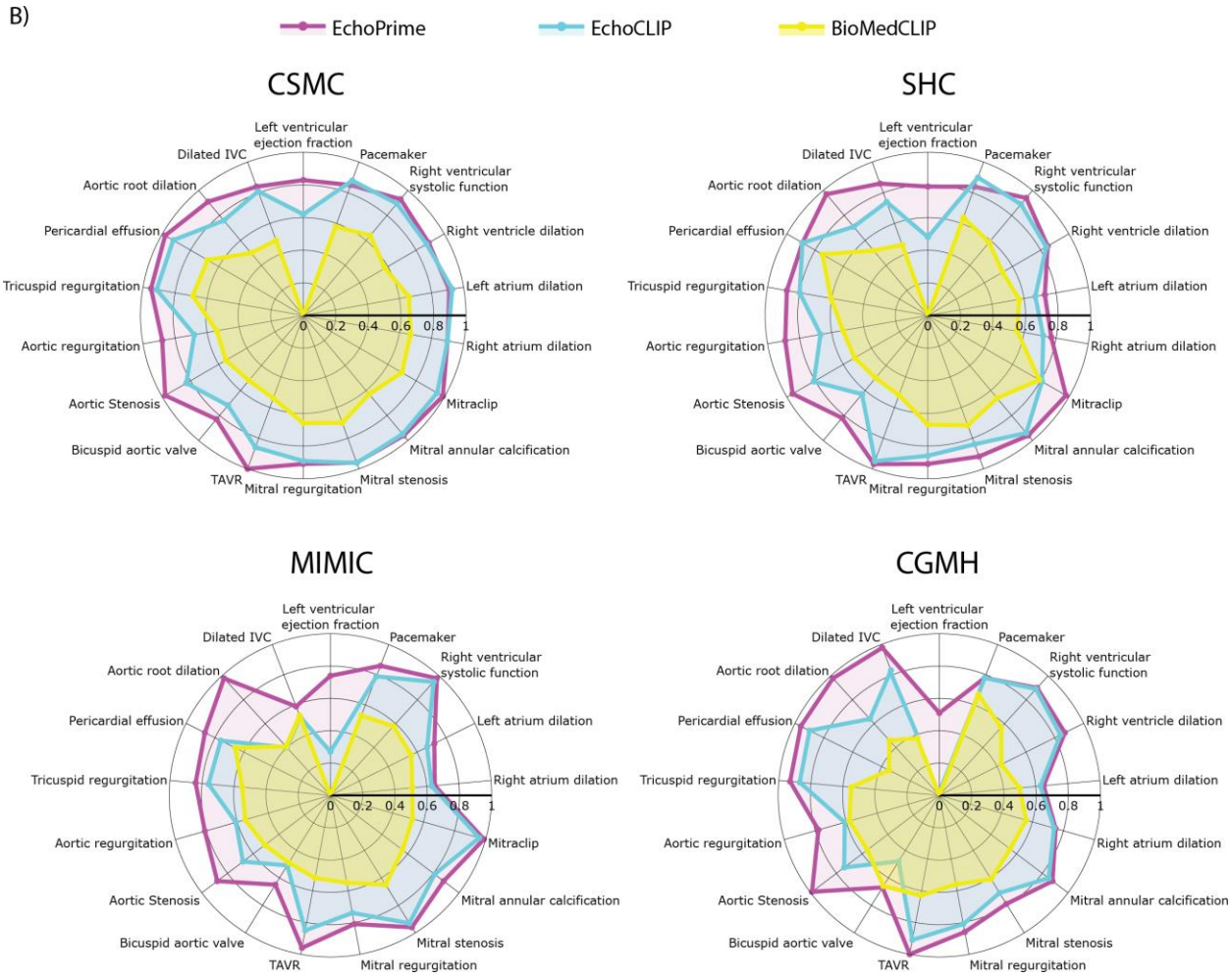


Figure 2**B**

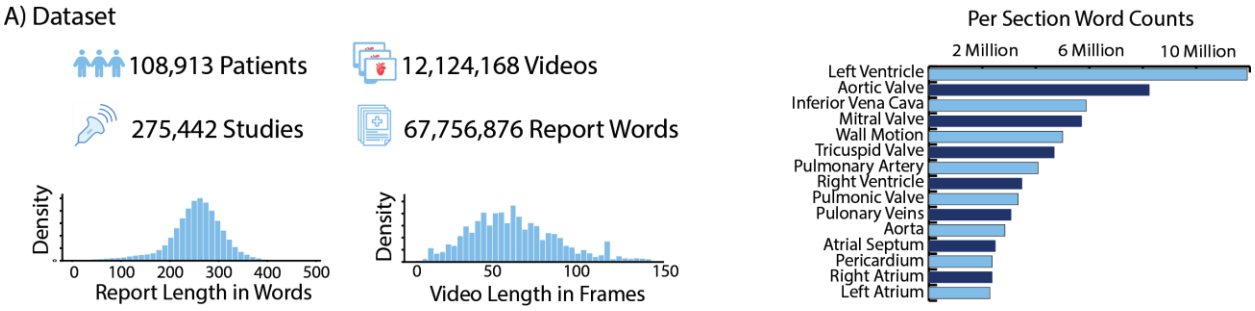


Figure 1A

Table 2: *Clinical characteristics of the study cohorts.*

	Cedars Sinai	***Stanford***	***MIMIC***	***CGMH***
Videos	*12,124,168*	*91,746*	*75,768*	*24,724*

Studies (N)	*275,442*	*1,792*	*2,431*	*1,188*
Patients	*108,913*	*1,779*	*1,767*	*1,188*
Age (mean (s.d.))	*66.31 (16.7)*	*60.42 (17.3)*	*67.67 (13.5)*	*65.87 (16.4)*
Female (%)	*117,324 (42.6)*	*843 (47.0)*	*909 (51.4)*	*583 (49.1)*
Race (%)				
Non-Hispanic White	*166,091 (60.3)*	*900 (50.2)*	*1,234 (69.8)*	*0 (0.0)*
Black	*35,624 (12.9)*	*76 (4.2)*	*343 (19.4)*	*0 (0.0)*
Hispanic	*27,194 (9.9)*	*241 (13.5)*	*68 (3.8)*	*0 (0.0)*
Asian	*20,299 (7.4)*	*301 (16.8)*	*45 (2.5)*	*1188 (100.00)*
Other	*19,832 (7.2)*	*161 (9.0)*	*42 (2.4)*	*0 (0.0)*
Unknown	*5,321 (1.9)*	*85 (4.7)*	*35 (2.0)*	*0 (0.0)*
Pacific Islander	*984 (0.4)*	*28 (1.6)*	*0 (0.0)*	*0 (0.0)*
Conditions (%)				
Hypertension	*112,919 (41.0)*	*1,055 (59.9)*	*1,510 (85.5)*	-
Heart failure	*93,875 (34.1)*	*746 (41.6)*	*996 (56.4)*	-
Atrial fibrillation	*60,448 (21.9)*	*555 (30.0)*	*510 (28.9)*	-
Chronic kidney disease	*50,302 (18.3)*	*580 (32.4)*	*755 (42.7)*	-
Diabetes mellitus	*46,656 (16.9)*	*168 (9.3)*	*486 (27.5)*	-
Pulmonary artery disease	*28,903 (10.5)*	*41 (2.3)*	*141 (8.0)*	-
Cerebrovascular accident	*18,442 (6.7)*	*104 (5.8)*	*160 (9.1)*	-
Myocardial infarction	*18,283 (6.6)*	*296 (16.5)*	*507 (28.7)*	-

D. Use of Statistics and Treatment of Uncertainties

The authors claim that EchoPrime matches or exceeds the performance of task-specific models, but they have not employed statistical methods to verify whether EchoPrime's results differ significantly from those of task-specific models. The lack of statistical measures in the paper is a

major limitation. Including confidence intervals for the results is strongly recommended to demonstrate the robustness of the performance comparisons.

Reply: Thank you for this important point. There was a substantial difference in the point estimate of performance for EchoPrime vs. benchmarks (BioMedCLIP,EchoCLIP), but we recognize the value of confidence intervals to evaluate true statistical difference. In this revision, we add confidence intervals to Supplementary Tables 2-6. Task Specific performance is reported from the cited literature and confidence intervals were included when they were reported in the original papers.

In **Methods "Evaluation and Benchmarking" section**:

> *To calculate the confidence intervals for performance metrics, we randomly sample the dataset with replacement, maintaining the same sample size as the original dataset, and compute the performance metrics. This process is repeated 10,000 times to generate a distribution of the metrics. Finally, determine the 2.5th and 97.5th percentiles to obtain the bootstrapped 95% confidence intervals.*

In Results "**Automated echocardiogram interpretation without supervised learning**" section:

> *EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6).*

Supplementary Table 2 *Performance of foundation models for biomedical imaging on Cedars-Sinai test set*

CEDARS-SINAI TEST SET					
	PubMedCLIP	*PMC-CLIP*	*BioMed CLIP*	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-0.32 (-0.40--0.25)*	*-0.38 (-0.46--0.31)*	*-2.60 (-3.00 - -2.20)*	*0.62 (0.59-0.65)*	***0.83 (0.81-0.85)***
LV ejection fraction (MAE)	*14.68 (14.30-15.06)*	*14.03 (13.58-14.49)*	*26.93 (26.42-27.44)*	*7.00 (6.74-7.26)*	***4.79 (4.60-4.97)***
Pacemaker	*0.50 (0.50-0.50)*	*0.51 (0.47-0.54)*	*0.58 (0.55-0.62)*	***0.88 (0.85-0.91)***	*0.85 (0.82-0.88)*
RV systolic function	*0.50 (0.49-0.51)*	*0.53 (0.48-0.58)*	*0.64 (0.59-0.69)*	*0.89 (0.86-0.92)*	***0.93 (0.91-0.95)***
RV dilation	*0.50 (0.45-0.56)*	*0.59 (0.50-0.67)*	*0.58 (0.52-0.64)*	*0.88 (0.83-0.92)*	***0.89 (0.83-0.95)***
LA dilation	*0.50 (0.50-0.50)*	*0.57 (0.48-0.65)*	*0.66 (0.58-0.74)*	***0.93 (0.89-0.96)***	*0.91 (0.86-0.96)*
RA dilation	*0.50 (0.50-0.50)*	*0.58 (0.47-0.68)*	*0.67 (0.59-0.75)*	*0.90 (0.82-0.97)*	***0.90 (0.82-0.97)***
Mitraclip	*0.50 (0.50-0.50)*	*0.49 (0.43-0.55)*	*0.70 (0.63-0.76)*	*0.95 (0.92-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.50 (0.46-0.55)*	*0.59 (0.54-0.64)*	*0.63 (0.58-0.68)*	*0.95 (0.93-0.96)*	***0.96 (0.95-0.97)***
Mitral stenosis	*0.50 (0.50-0.50)*	*0.55 (0.44-0.66)*	*0.70 (0.61-0.79)*	***0.96 (0.91-0.99)***	***0.96 (0.91-0.99)***

Mitral regurgitation	*0.50 (0.50-0.50)*	*0.59 (0.55-0.63)*	*0.66 (0.63-0.69)*	*0.89 (0.87-0.91)*	***0.92 (0.91-0.94)***
TAVR	*0.50 (0.50-0.51)*	*0.51 (0.46-0.55)*	*0.54 (0.50-0.57)*	*0.86 (0.83-0.89)*	***1.00 (0.99-1.00)***
Bicuspid AV	*0.50 (0.50-0.50)*	*0.51 (0.47-0.60)*	*0.52 (0.51-0.53)*	*0.72 (0.58-0.85)*	***0.83 (0.67-0.96)***
Aortic stenosis	*0.50 (0.50-0.52)*	*0.48 (0.42-0.55)*	*0.55 (0.53-0.57)*	*0.83 (0.79-0.87)*	***0.98 (0.96-0.99)***
Aortic regurgitation	*0.50 (0.50-0.50)*	*0.50 (0.41-0.58)*	*0.54 (0.53-0.55)*	*0.68 (0.60-0.76)*	***0.88 (0.83-0.93)***
Tricuspid regurgitation	*0.49 (0.45-0.53)*	*0.57 (0.51-0.63)*	*0.69 (0.64-0.74)*	*0.91(0.89-0.93)*	***0.95 (0.93-0.97)***
Pericardial effusion	*0.54 (0.45-0.62)*	*0.50 (0.42-0.57)*	*0.68 (0.59-0.77)*	*0.92 (0.89-0.95)*	***0.98 (0.95-0.99)***
Aortic root dilation	*0.50 (0.50-0.50)*	*0.44 (0.35-0.55)*	*0.50 (0.49-0.51)*	*0.76 (0.63-0.88)*	***0.91 (0.83-0.98)***
Dilated IVC	*0.50 (0.50-0.51)*	*0.53 (0.50-0.56)*	*0.49 (0.47-0.50)*	*0.81 (0.79-0.83)*	***0.84 (0.82-0.86)***
PA pressure (R^2)	*-0.59 (-0.74--0.46)*	*-0.83 (-1.07--0.62)*	*-0.19 (-0.24 - -0.14)*	*0.35 (0.29-0.41)*	***0.43 (0.39-0.47)***
PA pressure (MAE)	*11.27 (9.98-12.62)*	*13.65 (13.04-14.28*	*10.62 (10.12-11.12)*	*7.84 (7.45- 8.23)*	***7.30 (6.95-7.67)***

Supplementary Table 3 *Performance of foundation models for biomedical imaging on Stanford dataset*

STANFORD DATASET			
	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	-3.50 (-4.15- -3.03)	0.48 (0.40-0.55)	**0.79 (0.76-0.82)**
LV ejection fraction (MAE)	23.00 (22.68-23.79)	6.23 (5.89-6.44)	**4.14 (3.94-4.34)**
Pacemaker	0.64 (0.60-0.68)	**0.90 (0.88-0.93)**	0.84 (0.81-0.87)
RV systolic function	0.59 (0.02-0.92)	0.89 (0.73-1.00)	**0.94 (0.88-1.00)**
RV dilation	0.53 (0.49-0.58)	0.84 (0.79-0.89)	**0.85 (0.80-0.90)**
LA dilation	0.57 (0.55-0.60)	0.67 (0.65-0.70)	**0.73 (0.70-0.75)**
RA dilation	0.56 (0.52-0.59)	0.72 (0.69-0.75)	**0.77 (0.74-0.81)**
Mitraclip	0.79 (0.78-0.81)	0.81 (0.44-1.00)	**0.98 (0.94-1.00)**
Mitral annular calcification	0.66 (0.56-0.76)	0.94 (0.90-0.97)	**0.96 (0.94-0.98)**
Mitral stenosis	0.72 (0.63-0.81)	0.84 (0.75-0.92)	**0.92 (0.86-0.98)**
Mitral regurgitation	0.67 (0.57-0.71)	0.86 (0.82-0.90)	**0.91 (0.89-0.93)**
TAVR	0.52 (0.47-0.69)	0.95 (0.93-0.97)	**0.97 (0.90-1.00)**
Bicuspid AV	0.50 (0.47-0.54)	0.63 (0.53-0.73)	**0.82 (0.73-0.90)**
Aortic stenosis	0.52 (0.48-0.55)	0.81 (0.73-0.89)	**0.96 (0.92-0.99)**
Aortic regurgitation	0.53 (0.48-0.61)	0.67 (0.54-0.80)	**0.89 (0.80-0.97)**
Tricuspid regurgitation	0.61 (0.55-0.66)	0.80 (0.75-0.85)	**0.88 (0.86-0.91)**
Pericardial effusion	0.75 (0.60-0.66)	0.89 (0.78-0.98)	**0.89 (0.77-0.98)**
Aortic root dilation	0.52 (0.50-0.56)	0.71 (0.62-0.79)	**0.97 (0.94-0.99)**
Dilated IVC	0.46 (0.42-0.51)	0.75 (0.66-0.82)	**0.86 (0.81-0.91)**
PA pressure (R^2)	-0.25(-0.34- -0.17)	0.22 (0.13-0.30)	**0.36 (0.30-0.42)**
PA pressure (MAE)	10.92 (10.00-11.86)	8.80 (8.08-9.52)	**7.97 (7.35-8.66)**

Supplementary Table 4 *Performance of foundation models for biomedical imaging on MIMIC dataset from Boston*

MIMIC DATASET			
	BioMed CLIP	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-2.48 (-3.26--1.85)*	*0.27 (0.13-0.40)*	***0.74 (0.66-0.80)***
LV ejection fraction (MAE)	*26.96 (25.31-28.58)*	*10.37 (9.30-11.51)*	***6.21 (5.54-6.92)***
Pacemaker	*0.53 (0.39-0.68)*	*0.79 (0.66-0.90)*	***0.86 (0.80-0.92)***
RV systolic function	*0.58 (0.02-0.90)*	*0.95 (0.89-0.99)*	***0.98 (0.97-0.99)***
RV dilation	*-*	*-*	*-*
LA dilation	*0.56 (0.52-0.59)*	*0.67 (0.64-0.71)*	***0.72 (0.68-0.75)***
RA dilation	*0.51 (0.47-0.54)*	*0.63 (0.59-0.67)*	***0.65 (0.61-0.69)***
Mitraclip	*0.53 (0.29-0.99)*	*0.96 (0.86-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.56 (0.51-0.60)*	*0.81 (0.77-0.85)*	***0.88 (0.85-0.91)***
Mitral stenosis	*0.65 (0.48-0.98)*	*0.93 (0.80-1.00)*	***0.96 (0.95-0.98)***
Mitral regurgitation	*0.55 (0.52-0.58)*	*0.74 (0.71-0.76)*	***0.81 (0.79-0.83)***
TAVR	*0.52 (0.44-0.60)*	*0.85 (0.79-0.90)*	***0.96 (0.94-0.98)***
Bicuspid AV	*0.49 (0.46-0.56)*	*0.51 (0.42-0.61)*	***0.65 (0.52-0.78)***
Aortic stenosis	*0.51 (0.48-0.54)*	*0.68 (0.64-0.73)*	***0.88 (0.84-0.91)***
Aortic regurgitation	*0.55 (0.49-0.61)*	*0.61 (0.54-0.67)*	***0.81 (0.75-0.86)***
Tricuspid regurgitation	*0.55 (0.52-0.58)*	*0.76 (0.73-0.78)*	***0.84 (0.82-0.86)***
Pericardial effusion	*0.66 (0.61-0.72)*	*0.76 (0.70-0.81)*	***0.87 (0.83-0.92)***
Aortic root dilation	*0.41 (0.40-0.42)*	*0.40 (0.39-0.40)*	***0.98 (0.96-0.99)***
Dilated IVC	*0.53 (0.44-0.63)*	*0.52 (0.40-0.64)*	***0.59 (0.49-0.69)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Supplementary Table 5 *Performance of foundation models for biomedical imaging on CGMH dataset from Taiwan*

CGMH DATASET			
	BioMed CLIP	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-8.31 (-9.74--7.06)*	*0.02 (-0.13-0.15)*	***0.51 (0.44-0.57)***
LV ejection fraction (MAE)	*32.64 (31.89-33.38)*	*9.15 (8.75-9.54)*	***6.46 (6.19-6.74)***
Pacemaker	*0.67 (0.56-0.80)*	***0.78 (0.59-0.92)***	***0.78 (0.63-0.92)***
RV systolic function	*0.57 (0.41-0.72)*	*0.89 (0.84-0.94)*	***0.90 (0.79-0.98)***
RV dilation	*0.43 (0.30-0.61)*	*0.84 (0.65-0.97)*	***0.87 (0.69-0.97)***
LA dilation	*0.50 (0.47-0.53)*	*0.63 (0.60-0.65)*	***0.65 (0.62-0.67)***
RA dilation	*0.56 (0.47-0.64)*	*0.74 (0.65-0.83)*	***0.75 (0.67-0.84)***
Mitraclip	*-*	*-*	*-*
Mitral annular calcification	*0.55 (0.40-0.70)*	*0.85 (0.69-0.97)*	***0.88 (0.75-0.96)***
Mitral stenosis	*0.61 (0.48-0.79)*	*0.71 (0.51-0.92)*	***0.79 (0.59-0.98)***
Mitral regurgitation	*0.56 (0.50-0.63)*	*0.81 (0.76-0.86)*	***0.86 (0.82-0.89)***
TAVR	*0.63 (0.50-0.77)*	*0.91 (0.83-0.97)*	***1.00 (1.00-1.00)***
Bicuspid AV	*0.66 (0.43-1.00)*	*0.48 (0.35-0.77)*	***0.67 (0.45-1.00)***
Aortic stenosis	*0.56 (0.43-0.71)*	*0.74 (0.59-0.88)*	***0.99 (0.98-1.00)***
Aortic regurgitation	*0.58 (0.50-0.66)*	*0.61 (0.53-0.69)*	***0.78 (0.70-0.84)***
Tricuspid regurgitation	*0.55 (0.49-0.61)*	*0.87 (0.83-0.91)*	***0.93 (0.90-0.95)***
Pericardial effusion	*0.35 (0.16-0.65)*	*0.90 (0.69-1.00)*	***0.96 (0.91-1.00)***
Aortic root dilation	*0.46 (0.39-0.62)*	*0.64 (0.41-0.86)*	***0.98 (0.97-1.00)***
Dilated IVC	*0.38 (0.37-0.39)*	*0.83 (0.50-1.00)*	***0.98 (0.94-1.00)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Supplementary Table 6 *EchoPrime vs Task Specific Models*

	ECHOPRIME INTERNAL	***TASK SPECIFIC INTERNAL***	***ECHOPRIME EXTERNAL***	***TASK SPECIFIC EXTERNAL***
LV ejection fraction (R^2)	***0.83 (0.81-0.85)***	*0.81*	***0.79 (0.76-0.82)***	*0.77*
LV ejection fraction (MAE)	*4.79 (4.60-4.97)*	***4.10***	***4.14 (3.94-4.34)***	*6.00*
Pacemaker	*0.85 (0.82-0.88)*	-	*0.84 (0.81-0.87)*	-
RV systolic function	*0.93 (0.91-0.95)*	-	*0.94 (0.88-1.00)*	-
RV dilation	*0.89 (0.83-0.95)*	-	*0.85 (0.80-0.90)*	-
LA dilation	***0.91 (0.86-0.96)***	*0.85*	*0.73 (0.70-0.75)*	-
RA dilation	*0.90 (0.82-0.97)*	-	*0.77 (0.74-0.81)*	-
Mitraclip	*0.99 (0.99-1.00)*	-	*0.98 (0.94-1.00)*	-
Mitral annular calcification	*0.96 (0.95-0.97)*	-	*0.96 (0.94-0.98)*	-
Mitral stenosis	*0.96 (0.91-0.99)*	-	*0.92 (0.86-0.98)*	-
Mitral regurgitation	***0.92 (0.91-0.94)***	***0.92 (0.90-0.93)***	*0.91 (0.89-0.93)*	***0.95 (0.92–0.97)***
TAVR	*1.00 (0.99-1.00)*	-	*0.97 (0.90-1.00)*	-
Bicuspid AV	*0.83 (0.67-0.96)*	-	*0.82 (0.73-0.90)*	-
Aortic stenosis	***0.98 (0.96-0.99)***	***0.98 (0.97- 0.99)***	***0.96 (0.92-0.99)***	*0.95 (0.94- 0.96)*
Aortic regurgitation	*0.88 (0.83-0.93)*	-	*0.89 (0.80-0.97)*	-
Tricuspid regurgitation	***0.95 (0.93-0.97)***	*0.93 (0.91- 0.94)*	*0.88 (0.86-0.91)*	***0.95 (0.94 - 0.96)***
Pericardial effusion	***0.98 (0.95-0.99)***	*0.94 (0.75- 0.99)*	*0.89 (0.77-0.98)*	-
Aortic root dilation	*0.91 (0.83-0.98)*	-	*0.97 (0.94-0.99)*	-
Dilated IVC	*0.84 (0.82-0.86)*	-	*0.86 (0.81-0.91)*	-
PA pressure (R^2)	*0.43 (0.39-0.47)*	-	*0.36 (0.30-0.42)*	-
PA pressure (MAE)	*7.30 (6.95-7.67)*	-	*7.97 (7.35-8.66)*	-

E. Conclusions: Robustness, Validity, Reliability

While the model demonstrates promising results, the robustness of these findings is limited by the absence of human evaluation, lack of prospective analysis, and exclusive reliance on quantitative metrics such as AUC. Additionally, the absence of statistical measures (e.g., confidence intervals) weakens the reliability of the conclusions. Inclusion of additional comparison methods such as PubMedCLIP [3], PMC-CLIP [4], and ROCO [5], or at least reference to them, would improve the comprehensiveness and validity of the evaluation.

Reply: *Thank you for your feedback. In the original manuscript, we chose benchmarks that have highest likelihood of the strongest performance In this revision, we add additional comparison models (PubMedCLIP and PMC-CLIP) and show that they perform worse than EchoPrime (**Supplementary Tables 2)**.*

Supplementary Table 2 Performance of foundation models for biomedical imaging on Cedars-Sinai test set

CEDARS-SINAI TEST SET					
	PubMedCLIP	PMC-CLIP	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	-0.32 (-0.40--0.25)	-0.38 (-0.46--0.31)	-2.60 (-3.00 - -2.20)	0.62 (0.59-0.65)	**0.83 (0.81-0.85)**
LV ejection fraction (MAE)	14.68 (14.30-15.06)	14.03 (13.58-14.49)	26.93 (26.42-27.44)	7.00 (6.74-7.26)	**4.79 (4.60-4.97)**
Pacemaker	0.50 (0.50-0.50)	0.51 (0.47-0.54)	0.58 (0.55-0.62)	**0.88 (0.85-0.91)**	0.85 (0.82-0.88)
RV systolic function	0.50 (0.49-0.51)	0.53 (0.48-0.58)	0.64 (0.59-0.69)	0.89 (0.86-0.92)	**0.93 (0.91-0.95)**
RV dilation	0.50 (0.45-0.56)	0.59 (0.50-0.67)	0.58 (0.52-0.64)	0.88 (0.83-0.92)	**0.89 (0.83-0.95)**
LA dilation	0.50 (0.50-0.50)	0.57 (0.48-0.65)	0.66 (0.58-0.74)	**0.93 (0.89-0.96)**	0.91 (0.86-0.96)
RA dilation	0.50 (0.50-0.50)	0.58 (0.47-0.68)	0.67 (0.59-0.75)	0.90 (0.82-0.97)	**0.90 (0.82-0.97)**
Mitraclip	0.50 (0.50-0.50)	0.49 (0.43-0.55)	0.70 (0.63-0.76)	0.95 (0.92-1.00)	**0.99 (0.99-1.00)**
Mitral annular calcification	0.50 (0.46-0.55)	0.59 (0.54-0.64)	0.63 (0.58-0.68)	0.95 (0.93-0.96)	**0.96 (0.95-0.97)**
Mitral stenosis	0.50 (0.50-0.50)	0.55 (0.44-0.66)	0.70 (0.61-0.79)	**0.96 (0.91-0.99)**	**0.96 (0.91-0.99)**
Mitral regurgitation	0.50 (0.50-0.50)	0.59 (0.55-0.63)	0.66 (0.63-0.69)	0.89 (0.87-0.91)	**0.92 (0.91-0.94)**
TAVR	0.50 (0.50-0.51)	0.51 (0.46-0.55)	0.54 (0.50-0.57)	0.86 (0.83-0.89)	**1.00 (0.99-1.00)**
Bicuspid AV	0.50 (0.50-0.50)	0.51 (0.47-0.60)	0.52 (0.51-0.53)	0.72 (0.58-0.85)	**0.83 (0.67-0.96)**
Aortic stenosis	0.50 (0.50-0.52)	0.48 (0.42-0.55)	0.55 (0.53-0.57)	0.83 (0.79-0.87)	**0.98 (0.96-0.99)**
Aortic regurgitation	0.50 (0.50-0.50)	0.50 (0.41-0.58)	0.54 (0.53-0.55)	0.68 (0.60-0.76)	**0.88 (0.83-0.93)**
Tricuspid regurgitation	0.49 (0.45-0.53)	0.57 (0.51-0.63)	0.69 (0.64-0.74)	0.91(0.89-0.93)	**0.95 (0.93-0.97)**
Pericardial effusion	0.54 (0.45-0.62)	0.50 (0.42-0.57)	0.68 (0.59-0.77)	0.92 (0.89-0.95)	**0.98 (0.95-0.99)**
Aortic root dilation	0.50 (0.50-0.50)	0.44 (0.35-0.55)	0.50 (0.49-0.51)	0.76 (0.63-0.88)	**0.91 (0.83-0.98)**
Dilated IVC	0.50 (0.50-0.51)	0.53 (0.50-0.56)	0.49 (0.47-0.50)	0.81 (0.79-0.83)	**0.84 (0.82-0.86)**

PA pressure (R^2)	-0.59 (-0.74--0.46)	-0.83 (-1.07--0.62)	-0.19 (-0.24 - -0.14)	0.35 (0.29-0.41)	**0.43 (0.39-0.47)**
PA pressure (MAE)	11.27 (9.98-12.62)	13.65 (13.04-14.28	10.62 (10.12-11.12)	7.84 (7.45- 8.23)	**7.30 (6.95-7.67)**

Referee #3 (Remarks on code availability):

It is not very clear about the checkpoint of the model in the Github repo.

Reply: Checkpoint of the model can be found under the releases in the GitHub repository under model_data/weights/echo_prime_encoder.pt. During this revision, we have clarified our README, added requirements.txt and Dockerfile to make it easier for a broader audience to access and work with our model.

Reviewer Comments

Referee #1 (Remarks to the Author):

The manuscript introduces EchoPrime, a multi-task model for echocardiography that claims state-of-the-art performance across 17 binary and 2 continuous tasks, utilizing a contrastive encoder and anatomical attention module. I have some comments following the points below:

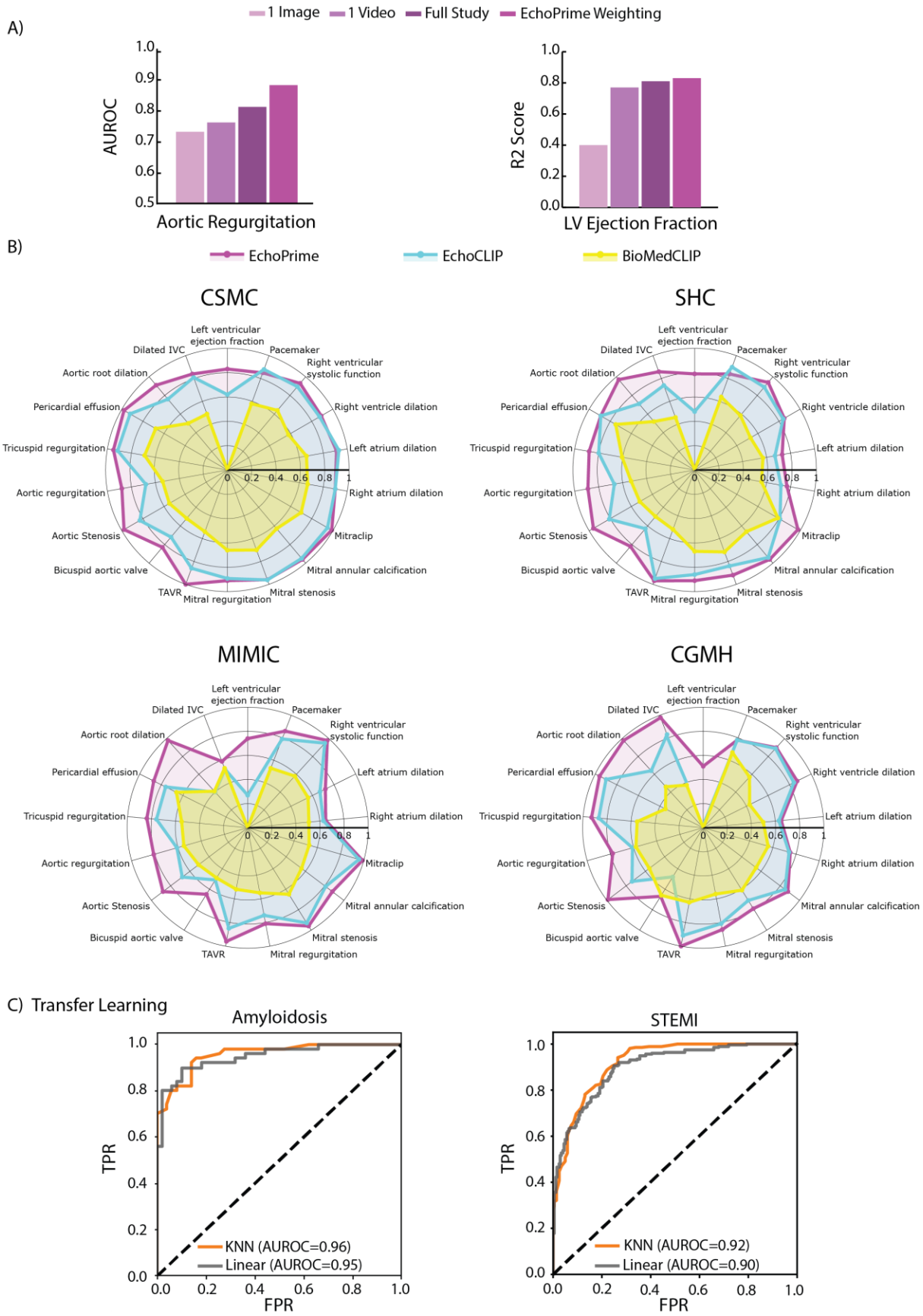
Reply: We thank the reviewer for their feedback and time in reviewing this manuscript. In contrast to prior work, which focus on either individual tasks or a limited set of binary and continuous tasks, EchoPrime generates the entire report, which comprises of 15 sections of over 1000 different statements that cardiologists frequently make in interpreting echocardiograms. The original manuscript benchmark metrics sought to highlight 19 relevant tasks for comparison with previous models, but EchoPrime is able to do many more tasks than just the 19 tasks mentioned.

These EchoPrime statements comprise the full range of common cardiovascular diseases and include both classification (often multiclass) and regression/continuous tasks. To clarify this point, in this revision, we demonstrate EchoPrime's performance on an increased range of classification tasks and comparisons with 11 regression tasks.

Furthermore, to emphasize the generalizability of EchoPrime's performance, we introduce two new external validation sites (including an international validation site) which replicate the strong performance across multiple settings, image acquisition protocols, and workflows.

In **Results** "**Automated echocardiogram interpretation without supervised learning**" section:

> *This approach allows a single foundation model to be applied directly to all echocardiography interpretation tasks, enabling the interpretation of hundreds of pathologies across 15 anatomical sections resulting in over 1000 different report statements that cardiologists frequently make in interpreting echocardiograms.*
>
> *We evaluated EchoPrime on a diverse set of benchmarks for cardiac structure and function across datasets from four different institutions (**Error! Reference source not found.**). On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging (Supplementary Tables 2-5), and without additional fine-tuning or training, EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6). EchoPrime had a mean AUC of 0.92 on CSMC, 0.89 on Stanford Healthcare (SHC), 0.86 on Beth Israel Deaconess Medical Center (MIMIC) and 0.86 on Chang Gung Memorial Hospital (CGMH) in Taiwan across 17 classification tasks and 11 regression tasks (Table 2, Supplementary Table 7).*


A)
1 Image
1 Video
Full Study
EchoPrime Weighting
AUROC
Aortic Regurgitation
R2 Score
LV Ejection Fraction
B)
EchoPrime
EchoCLIP
BioMedCLIP
CSMC
SHC
MIMIC
CGMH
Left ventricular ejection fraction
Pacemaker
Right ventricular systolic function
Right ventricle dilation
Left atrium dilation
Right atrium dilation
Mitraclip
Mitral annular calcification
Mitral stenosis
Mitral regurgitation
TAVR
Bicuspid aortic valve
Aortic Stenosis
Aortic regurgitation
Tricuspid regurgitation
Pericardial effusion
Aortic root dilation
Dilated IVC
C) Transfer Learning
Amyloidosis
STEMI
TPR
FPR
KNN (AUROC=0.96)
Linear (AUROC=0.95)
KNN (AUROC=0.92)
Linear (AUROC=0.90)

Supplementary Table 7 EchoPrime performance on regression tasks

Task	R^2	MAE
Ejection Fraction	0.83 (0.81-0.85)	4.79 (4.60-4.97)
Pulmonary Artery Pressure	0.43 (0.39-0.47)	7.30 (7.07-7.80)
Wall Motion Score	0.77 (0.74-0.80)	0.10 (0.10-0.11)
TAPSE	0.24 (0.05-0.49)	0.35 (0.28-0.45)
MV Area	0.34 (0.20-0.44)	0.81 (0.71-0.93)
Transaortic velocity	0.65 (0.58-0.72)	23.84 (20.38-27.82)
Transaortic gradient	0.36 (0.29-0.42)	6.84 (6.25-7.45)
TRgradient	0.28 (0.16-0.39)	6.95 (6.51-7.46)
Sinus of Valsalva Size	0.62 (0.55-0.67)	0.24 (0.22-0.26)
Ascending Aorta Size	0.16 (0.02-0.45)	0.41 (0.29-0.63)
IVC Diameter	0.70 (0.65-0.75)	2.85 (2.65-3.06)

1. Clinical Utility and Workflow Integration: EchoPrime's precision improves with additional views and videos (Figure 2A), indicating potential value as a diagnostic support tool. However, this approach relies on high-quality, multi-view echocardiographic data, which requires skilled sonographers or physicians. This dependency may limit its feasibility in resource-constrained settings.

Reply: Thank you for your comments regarding EchoPrime as a diagnostic support tool. Echocardiography is the most common form of cardiac imaging, and by including the wide range of images from over a decade of a large academic echo lab, has both high and low quality images and a wide range of physiology not seen in small datasets. To assess EchoPrime's performance in low-quality resource-constrained setting, we report the performance on studies deemed technically challenging and to assess EchoPrime's performance with limited datasets, we report the performance using only one A4C video per study. The following results show EchoPrime works good even in a low-quality and resource-constrained settings.

Furthermore, in this revision, we add two more external validation sites, including an international site where cardiologists perform echocardiography but record more limited/abbreviated studies (averaging 20 videos + images per study rather than conventionally 80 – 100 videos + images per study in American sites) and we show the generalizable strong performance in different practice patterns.

In **Results** "**Leveraging multi-view echocardiographic data improves performance**" section

EchoPrime can additionally be leveraged even in low-quality images or settings where limited views are acquired. We evaluated EchoPrime in the set of studies where the study was described as "technically difficult" in the clinician report and show similar performance (Supplementary Table 8). To test the performance in a single view setting, we test EchoPrime on 2172 studies with only one A4C video per study. The performance across all tasks were similar to EchoPrime performance in the full setting (Supplementary Table 8).

Supplementary Table 8 EchoPrime performance in low-quality resource-constrained setting

	FULL TEST	LOW-QUALITY	SINGLE-VIEW
LV ejection fraction (R^2)	0.83 (0.81-0.85)	*0.80 (0.75-0.84)*	0.78 (0.75-0.80)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	*5.30 (4.86-5.76)*	5.22 (5.02-5.43)
Pacemaker	0.85 (0.82-0.88)	*0.89 (0.84-0.93)*	0.83 (0.80-0.85)
RV systolic function	0.93 (0.91-0.95)	*0.91 (0.85-0.96)*	0.92 (0.89-0.94)
RV dilation	0.89 (0.83-0.95)	*0.85 (0.72-0.96)*	0.92 (0.87-0.95)
LA dilation	0.91 (0.86-0.96)	*0.95 (0.89-0.99)*	0.92 (0.87-0.96)
RA dilation	0.90 (0.82-0.97)	*0.89 (0.70-1.00)*	0.90 (0.82-0.97)
Mitraclip	0.99 (0.99-1.00)	*1.00 (0.99-1.00)*	0.97 (0.94-0.99)
Mitral annular calcification	0.96 (0.95-0.97)	*0.97 (0.94-0.98)*	0.94 (0.91-0.96)
Mitral stenosis	0.96 (0.91-0.99)	*0.97 (0.90-1.00)*	0.94 (0.87-0.98)
Mitral regurgitation	0.92 (0.91-0.94)	*0.91 (0.84-0.96)*	0.89 (0.87-0.91)
TAVR	1.00 (0.99-1.00)	*1.00 (0.99-1.00)*	0.90 (0.87-0.93)
Bicuspid AV	0.83 (0.67-0.96)	*0.72 (0.44-1.00)*	0.78 (0.62-0.92)
Aortic stenosis	0.98 (0.96-0.99)	*0.93 (0.83-1.00)*	0.80 (0.75-0.85)
Aortic regurgitation	0.88 (0.83-0.93)	*0.78 (0.38-0.99)*	0.75 (0.68-0.82)
Tricuspid regurgitation	0.95 (0.93-0.97)	*0.98 (0.97-0.99)*	0.92 (0.88-0.94)
Pericardial effusion	0.98 (0.95-0.99)	*0.94 (0.91-0.97)*	0.93 (0.89-0.97)
Aortic root dilation	0.91 (0.83-0.98)	*0.96 (0.91-1.00)*	0.81 (0.70-0.91)
Dilated IVC	0.84 (0.82-0.86)	*0.82 (0.77-0.87)*	0.80 (0.78-0.83)

PA pressure (R^2)	0.43 (0.39-0.47)	*0.32 (0.17-0.44)*	0.36 (0.31-0.40)
PA pressure (MAE)	7.30 (6.95-7.67)	*7.76 (6.96-8.63)*	7.67 (7.29-8.06)

The model's focus on classification tasks (rather than regression tasks beyond EF estimation) could also restrict its usefulness for comprehensive echocardiography, which depends on detailed measurements and morphological assessments essential for patient care.

Reply: Thank you for this important question. To clarify, in contrast to prior work, which focus on a limited set of binary and continuous tasks. EchoPrime generates the entire report, which comprises of 15 sections of over 1000 different statements. These statements comprise the full range of common cardiovascular diseases and include both classification (often multiclass) and regression/continuous tasks. When there are so many potential outputs, classification tasks (yes/no whether the model predicted an attribute in the same way as a cardiologist) is the easiest way to produce metrics, however there are multiple other regression tasks and other metrics that can be evaluated. In addition to LVEF and RVSP as regression tasks, our model also produces a wall motion score, TR gradient and other quantitative metrics (IVC diameter, Sinus of Valsalva) in an echo report. In this revision, we provide additional metrics to assess EchoPrime's performance on these regression tasks in Supplementary Table 7.

In **Results** "**Automated echocardiogram interpretation without supervised learning**" section

> *On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging (Supplementary Tables 2-5), and without additional fine-tuning or training, EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6). EchoPrime had a mean AUC of 0.92 on CSMC, 0.89 on Stanford Healthcare (SHC), 0.86 on Beth Israel Deaconess Medical Center (MIMIC) and 0.86 on Chang Gung Memorial Hospital (CGMH) in Taiwan across 17 classification tasks and 11 regression tasks (Table 2, Supplementary Table 7).*

Supplementary Table 7 EchoPrime performance on regression tasks on the internal held-out test set

Task	R^2	MAE
Ejection Fraction	0.83 (0.81-0.85)	4.79 (4.60-4.97)
Pulmonary Artery Pressure	0.43 (0.39-0.47)	7.30 (7.07-7.80)
Wall Motion Score	0.77 (0.74-0.80)	0.10 (0.10-0.11)
TAPSE	0.24 (0.05-0.49)	0.35 (0.28-0.45)
MV Area	0.34 (0.20-0.44)	0.81 (0.71-0.93)
Transaortic velocity	0.65 (0.58-0.72)	23.84 (20.38-27.82)

Transaortic gradient	0.36 (0.29-0.42)	6.84 (6.25-7.45)
TRgradient	0.28 (0.16-0.39)	6.95 (6.51-7.46)
Sinus of Valsalva Size	0.62 (0.55-0.67)	0.24 (0.22-0.26)
Ascending Aorta Size	0.16 (0.02-0.45)	0.41 (0.29-0.63)
IVC Diameter	0.70 (0.65-0.75)	2.85 (2.65-3.06)

If EchoPrime is intended for high-pre-test probability populations, its added value is unclear since comprehensive studies and specialist reviews are already standard. Alternatively, if the goal is to "clear" low-risk patients, its negative predictive value (NPV) should be compared to simpler and more accessible tools. For instance, Elias et al.'s ECG-based model detects aortic valve regurgitation with an AUC of 0.84 (JACC, 2022), close to EchoPrime's AUC of 0.88. Given that ECG is less resource-intensive and more widely available, the small AUC improvement raises questions about EchoPrime's cost-effectiveness. The authors should clarify EchoPrime's NPV compared to existing tools in low- and intermediate-pre-test populations, its advantages in high-pre-test settings where comprehensive studies are already performed, and how it accommodates the detailed assessments needed in echocardiography. These points are essential to evaluate its clinical value and potential integration into workflows.

Reply: Thank you for your thoughtful comments. The comparison with an AI ECG model is potentially misleading and comparing apples to oranges. ECGs tend to be more commonly obtained in healthy individuals, and a relatively healthy cohort with few abnormal cases will inflate performance metrics. This bias will be reflected in a high reported NPV while the model would not perform as well in such an echo-lab based cohort. In contrast, in our paper, we compare with many other echocardiography AI models and show similar to superior performance.

Echocardiography is the most common form of cardiac imaging and is used in both screening (low pretest probability patients) and patients with known cardiovascular disease (high pretest probability patients). Notably, echocardiography is the definitive test for many of these cardiovascular diseases, while an ECG based model is to highlight suspicion and would require a confirmatory test (like echocardiography). In this setting, a direct comparison of AUC might be misleading, because small differences in AUC might seem like small difference but is the difference between finding suspicion for a disease *(in an ECG setting)* and diagnosing a disease in the definitive testing modality of choice (*in the ehco setting)*.

We would like to clarify that EchoPrime is not intended for clearing low-risk patients or as a screening tool in low-pre-test probability populations. Instead, its application is integrated within the standard echocardiography workflow, where echocardiograms are obtained, an initial report is drafted by the sonographer (but often is not due to time constraints), and the final report is reviewed and modified by a cardiologist. EchoPrime's primary goal is to assist sonographers and cardiologists in interpreting echocardiographic studies by automating preliminary report generation. This approach offers key advantages including:

1. **Enhanced efficiency:** Automating the report generation process streamlines workflow, allowing for faster interpretation and turnaround.
2. **Consistency:** By reducing inter-physician variability, EchoPrime helps standardize interpretations, ensuring more consistent quality of care.

2. Observed Discrepancy in Outputs: Testing the model on my MacBook Pro (macOS Apple M3 Max, 36 GB) with a cardiac ultrasound not provided in the GitHub repository, I found a discrepancy between the EF reported in the generated text (55%) and the EF derived from output logits (61%) using the provided Jupyter Notebook. This inconsistency raises questions about which value should be considered accurate and whether outputs are standardized across formats. It would be helpful for the authors to explain the rationale behind this discrepancy, such as whether it reflects differences in post-processing methods or thresholds used for generating outputs.

Reply: Thank you for your rigorous evaluation of our code. In our codebase, we included intermediate outputs for demonstration purposes. These debugging outputs are not directly reflected in the results and meant to show successful individual report retrieval. EchoPrime embeds the study in the latent space and then finds the closest report embeddings, but the output is the weighted average of the 50 closest reports. In your example, the 55% LVEF output appears from mapping to the closest retrieved report while 61% is the result of retrieving 50 closest reports and then averaged LVEF across them (the intended final EchoPrime output). For metric purposes, the intended final EchoPrime output was used for assessment.

3. While the manuscript does not explicitly discuss pressure estimations such as right atrial pressure (RAP), testing the software using the provided Jupyter Notebook generated a text report that included an estimated RAP value (12 mmHg, see above). This raises questions about whether such estimations are validated or reliable. Are these pressure values directly derived from model predictions, and if so, have they been validated?

Reply: EchoPrime can predict any cardiac features present in a standard echocardiography report, and goes beyond 17 classification tasks and 2 regression tasks. We selected 19 representative features for analysis that encompass most cardiac structures to give insight into the overall performance, however there are over 1000 possible output phrases.

In the reviewers output, it looks like 12mmHg is for the TR max velocity (RV/RA pressure gradient), not the right atrial pressure (RAP). Per reviewer's request, in this revision we add right TR (Tricuspid Valve) gradient as a clinical evaluation metric in Supplementary Table 7.

Supplementary Table 7 EchoPrime performance on regression tasks on the internal held-out test set

Task	R^2	MAE
Ejection Fraction	0.83 (0.81-0.85)	4.79 (4.60-4.97)

Pulmonary Artery Pressure	0.43 (0.39-0.47)	7.30 (7.07-7.80)
Wall Motion Score	0.77 (0.74-0.80)	0.10 (0.10-0.11)
TAPSE	0.24 (0.05-0.49)	0.35 (0.28-0.45)
MV Area	0.34 (0.20-0.44)	0.81 (0.71-0.93)
Transaortic velocity	0.65 (0.58-0.72)	23.84 (20.38-27.82)
Transaortic gradient	0.36 (0.29-0.42)	6.84 (6.25-7.45)
TR gradient	**0.28 (0.16-0.39)**	**6.95 (6.51-7.46)**
Sinus of Valsalva Size	0.62 (0.55-0.67)	0.24 (0.22-0.26)
Ascending Aorta Size	0.16 (0.02-0.45)	0.41 (0.29-0.63)
IVC Diameter	0.70 (0.65-0.75)	2.85 (2.65-3.06)

4. Benchmarking Against Comprehensive Tools: The manuscript provides a comparison with EchoCLIP and BioMedCLIP. However, EchoPrime's focus on binary classifications raises questions about its clinical utility relative to commercially available tools like US2AI (https://us2.ai/product/), which deliver fully automated, multi-view quantitative echocardiography reports. US2AI includes continuous outputs including LV/RV morphology, and diastolic function, which are critical for standard echocardiographic reporting. While the lack of access to proprietary systems like US2AI is understandable, at least a discussion of how EchoPrime compares to or complements such tools would help articulate its standalone value and potential niche in clinical workflows.

Reply: Thank you for this important question. Because US2.AI is a commercial product is blackbox and code not publicly available, we are not able to benchmark performance metrics, however we note that US2.AI does only segmentation / quantitative assessments, and is not able to do any of the qualitative report findings EchoPrime reports. As the reviewer notes, the echocardiography report is primarily qualitative rather than quantitative assessments (more classification tasks than regression tasks) and US2AI and similar products cannot perform any of the classification tasks.

In addition to benchmarking with all available publicly available echocardiography AI models, EchoPrime is the first foundation model designed for comprehensive echocardiography evaluation and captures the full range of likely interpretation statements. Furthermore, unlike US2AI, which was trained on only a few thousand studies, EchoPrime leverages a training dataset of 12 million echocardiography videos, the largest yet training dataset for echocardiography AI and provides a more robust and generalized understanding of cardiac physiology.

5. Statistical and Clinical Significance: The manuscript highlights small percentage improvements in performance over prior models, such as a 2% increase in AUC for tricuspid regurgitation. However, it is unclear whether these improvements are statistically significant or

clinically meaningful. Without p-values or confidence intervals (at least I did not find them), the reported gains could reflect expected variability rather than true advancements.

Reply: EchoPrime outperforms previous foundation models and matches or outperforms task specific models. Often there's a small improvement to task-specific models (like 2% in TR mentioned), however EchoPrime is comprehensive in nature and is a single model rather than requiring many models for an entire echocardiography to report. EchoPrime can predict all these features (in an unsupervised manner), whereas a task-specific model means that a model is specifically fine-tuned for that feature. The task specific improvements in AUC for such a comprehensive general model, highlight EchoPrime's exceptional generalizability, demonstrating its ability to match or surpass task-specific models without feature-specific fine-tuning.

As per reviewer's request we add confidence intervals to all AUC, R2 and MAE metrics in the paper and explain how we calculated them in methods. We also report confidence intervals for task specific models when that is available in the original respective papers.

In **Methods "Evaluation and Benchmarking" section**:

> *To calculate the confidence intervals for performance metrics, we randomly sample the dataset with replacement, maintaining the same sample size as the original dataset, and compute the performance metrics. This process is repeated 10,000 times to generate a distribution of the metrics. Finally, determine the 2.5th and 97.5th percentiles to obtain the bootstrapped 95% confidence intervals.*

In Results "**Automated echocardiogram interpretation without supervised learning**" section:

> *EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6).*

Supplementary Table 6 *EchoPrime vs Task Specific Models*

	ECHOPRIME INTERNAL	TASK SPECIFIC INTERNAL	ECHOPRIME EXTERNAL	TASK SPECIFIC EXTERNAL
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.81	**0.79 (0.76-0.82)**	0.77
LV ejection fraction (MAE)	4.79 (4.60-4.97)	**4.10**	**4.14 (3.94-4.34)**	6.00
Pacemaker	0.85 (0.82-0.88)	-	0.84 (0.81-0.87)	-
RV systolic function	0.93 (0.91-0.95)	-	0.94 (0.88-1.00)	-
RV dilation	0.89 (0.83-0.95)	-	0.85 (0.80-0.90)	-
LA dilation	**0.91 (0.86-0.96)**	0.85	0.73 (0.70-0.75)	-
RA dilation	0.90 (0.82-0.97)	-	0.77 (0.74-0.81)	-
Mitraclip	0.99 (0.99-1.00)	-	0.98 (0.94-1.00)	-
Mitral annular calcification	0.96 (0.95-0.97)	-	0.96 (0.94-0.98)	-

Mitral stenosis	*0.96 (0.91-0.99)*	-	*0.92 (0.86-0.98)*	-
Mitral regurgitation	***0.92 (0.91-0.94)***	***0.92 (0.90-0.93)***	*0.91 (0.89-0.93)*	***0.95 (0.92–0.97)***
TAVR	*1.00 (0.99-1.00)*	-	*0.97 (0.90-1.00)*	-
Bicuspid AV	*0.83 (0.67-0.96)*	-	*0.82 (0.73-0.90)*	-
Aortic stenosis	***0.98 (0.96-0.99)***	***0.98 (0.97- 0.99)***	***0.96 (0.92-0.99)***	*0.95 (0.94- 0.96)*
Aortic regurgitation	*0.88 (0.83-0.93)*	-	*0.89 (0.80-0.97)*	-
Tricuspid regurgitation	***0.95 (0.93-0.97)***	*0.93 (0.91- 0.94)*	*0.88 (0.86-0.91)*	***0.95 (0.94 - 0.96)***
Pericardial effusion	***0.98 (0.95-0.99)***	*0.94 (0.75- 0.99)*	*0.89 (0.77-0.98)*	-
Aortic root dilation	*0.91 (0.83-0.98)*	-	*0.97 (0.94-0.99)*	-
Dilated IVC	*0.84 (0.82-0.86)*	-	*0.86 (0.81-0.91)*	-
PA pressure (R^2)	*0.43 (0.39-0.47)*	-	*0.36 (0.30-0.42)*	-
PA pressure (MAE)	*7.30 (6.95-7.67)*	-	*7.97 (7.35-8.66)*	-

6. Performance in Challenging Imaging Conditions: the manuscript does not address EchoPrime's performance under suboptimal imaging conditions, such as poor acoustic windows, which are common in real-world practice. Understanding its robustness in these scenarios is crucial for assessing its clinical applicability. I believe including an analysis of performance stratified by image quality would enhance the model's credibility for diverse clinical settings.

Reply: Thank you for your suggestion. Echocardiography is the most common form of cardiac imaging, and by including the wide range of images from over a decade of a large academic echo lab, has both high and low quality images. To assess EchoPrime's performance in low-quality resource-constrained setting, we report the performance on studies deemed technically challenging. The following results show EchoPrime works good even in a low-quality images.

In **Results** "**Leveraging multi-view echocardiographic data improves performance**" section

> *EchoPrime can additionally be leveraged even in low-quality images or settings where limited views are acquired. We evaluated EchoPrime in the set of studies where the study was described as "technically difficult" in the clinician report and show similar performance (Supplementary Table 8). To test the performance in a single view setting, we test EchoPrime on 2172 studies with only one A4C video per study. The performance across all tasks were similar to EchoPrime performance in the full setting (Supplementary Table 8).*

Supplementary Table 8 *EchoPrime performance in low-quality resource-constrained setting*

	FULL TEST	***LOW-QUALITY***	***SINGLE-VIEW***

LV ejection fraction (R^2)	*0.83 (0.81-0.85)*	*0.80 (0.75-0.84)*	*0.78 (0.75-0.80)*
LV ejection fraction (MAE)	*4.79 (4.60-4.97)*	*5.30 (4.86-5.76)*	*5.22 (5.02-5.43)*
Pacemaker	*0.85 (0.82-0.88)*	*0.89 (0.84-0.93)*	*0.83 (0.80-0.85)*
RV systolic function	*0.93 (0.91-0.95)*	*0.91 (0.85-0.96)*	*0.92 (0.89-0.94)*
RV dilation	*0.89 (0.83-0.95)*	*0.85 (0.72-0.96)*	*0.92 (0.87-0.95)*
LA dilation	*0.91 (0.86-0.96)*	*0.95 (0.89-0.99)*	*0.92 (0.87-0.96)*
RA dilation	*0.90 (0.82-0.97)*	*0.89 (0.70-1.00)*	*0.90 (0.82-0.97)*
Mitraclip	*0.99 (0.99-1.00)*	*1.00 (0.99-1.00)*	*0.97 (0.94-0.99)*
Mitral annular calcification	*0.96 (0.95-0.97)*	*0.97 (0.94-0.98)*	*0.94 (0.91-0.96)*
Mitral stenosis	*0.96 (0.91-0.99)*	*0.97 (0.90-1.00)*	*0.94 (0.87-0.98)*
Mitral regurgitation	*0.92 (0.91-0.94)*	*0.91 (0.84-0.96)*	*0.89 (0.87-0.91)*
TAVR	*1.00 (0.99-1.00)*	*1.00 (0.99-1.00)*	*0.90 (0.87-0.93)*
Bicuspid AV	*0.83 (0.67-0.96)*	*0.72 (0.44-1.00)*	*0.78 (0.62-0.92)*
Aortic stenosis	*0.98 (0.96-0.99)*	*0.93 (0.83-1.00)*	*0.80 (0.75-0.85)*
Aortic regurgitation	*0.88 (0.83-0.93)*	*0.78 (0.38-0.99)*	*0.75 (0.68-0.82)*
Tricuspid regurgitation	*0.95 (0.93-0.97)*	*0.98 (0.97-0.99)*	*0.92 (0.88-0.94)*
Pericardial effusion	*0.98 (0.95-0.99)*	*0.94 (0.91-0.97)*	*0.93 (0.89-0.97)*
Aortic root dilation	*0.91 (0.83-0.98)*	*0.96 (0.91-1.00)*	*0.81 (0.70-0.91)*
Dilated IVC	*0.84 (0.82-0.86)*	*0.82 (0.77-0.87)*	*0.80 (0.78-0.83)*
PA pressure (R^2)	*0.43 (0.39-0.47)*	*0.32 (0.17-0.44)*	*0.36 (0.31-0.40)*
PA pressure (MAE)	*7.30 (6.95-7.67)*	*7.76 (6.96-8.63)*	*7.67 (7.29-8.06)*

7. Implementation and User-Friendliness: running the software on a MacBook using a CPU backend required adjustments. To improve user-friendliness and ensure reproducibility, the README should specify exact requirements for setting up the virtual environment. Providing a Dockerized version would ensure cross-platform compatibility and consistency. Currently, it is unclear whether the exact Python modules and dependencies used align with those intended by the authors, potentially impacting reproducibility and performance.

Reply: Thank you for your feedback. We add requirements.txt and a Dockerfile to improve the accessibility of the shared code. https://github.com/echonet/EchoPrime/blob/main/Dockerfile https://github.com/echonet/EchoPrime/blob/main/requirements.txt

8. Non-Echocardiographic Disease Predictions (Minor): the manuscript refers to STEMI and cardiac amyloidosis as "non-echocardiographic primary diseases," which may not fully capture their diagnostic pathways. Echocardiography is commonly used to diagnose cardiac amyloidosis, offering structural and functional insights, and may play a significant role in assessing cardiac function in STEMI patients. Clarifying the term "non-echocardiographic traits" would prevent potential misunderstandings and better align the claims with established clinical practices.

Reply: Thank you for this comment. We sought to clarify that echocardiography is not the definitive test of choice for diagnosing STEMI or cardiac amyloidosis. ECGs or nuclear imaging is often used to confirm the diagnosis, while echocardiography can find corroborating evidence that is insufficient to confirm diagnosis.

Similar to what the reviewer described with AI-ECG models, even when performing well, STEMI and Amyloidosis diagnoses are not something that commonly appears in the report written by a sonographer or echocardiography. We clarify in this revision, and we modify the term "non-echocardiographic traits" with "Multimodal cardiovascular diagnoses" throughout the text.

In **Methods** "**Disease Diagnosis with Probing**" section:

> ***Multimodal Disease Diagnosis***
> *We apply k-nearest neighbors (KNN) and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyhloidosis as two diseases for which echocardiography is not the definitive test echocardiography can find corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study, and used along with clinician-assigned diagnoses to fit the models. For KNN probing, we use KNeighborsClassifier, and for linear probing, we apply LogisticRegression, both from scikit-learn library.*

Referee #2:

Congratulations to the authors as you continue your progressive refinement of the use of AI in echocardiography. Expanding from your recent foundational interpretive model EchoCLIP, which was trained on one million echo video clips, you have now used over 12 million echo clips (essentially all echocardiograms done at Cedars Sinai over a decade, along with their de-identified echo reports) to train a successor foundation model (EchoPrime), which has many advantages over its predecessor. Among the advances in modeling is contrastive learning to form a joint image-text representation, a more sophisticated view classifier that allows an anatomic attention model to understand which of the 58 characterized views are most likely to show the

finding of current interest, and use of the full video clip for assessment rather than individual frames. The model was tested on hold-out sets of 4044 total echo studies from both Cedars Sinai as well as Stanford. EchoPrime claims the ability to predict the presence of hundreds of pathologies encountered over the years but was tested on 19 fairly common features encapsuling chamber size and function and valve stenosis and regurgitation as well as implanted devices (e.g., prosthetic valves, MitraClip, and pacers). Two parameters (LV ejection fraction and RV systolic pressure) were compared to the clinical report by linear regression (r^2 and mean average error), while the other 17 were tested with AUROC for binary variables, significantly outperforming EchoCLIP, the more generalized foundation model BioMedCLIP, and several task specific models.

Reply: Thank you for your comments regarding our work on AI in echocardiography. As you noted in your review, EchoPrime is able to predict the presence of hundreds of pathologies across anatomical 15 sections of over 1000 different statements that cardiologists frequently make in interpreting echocardiograms. The comprehensiveness of EchoPrime's interpretation is also shown in additional metrics of EchoPrime performance in more tasks requested by the reviewers (Supplementary Table 7) These results were not previously described in the first version of manuscript but are evaluation metrics from the same model without more training.

Supplementary Table 7 *EchoPrime performance on regression tasks on the internal held-out test set*

Task	*R^2*	*MAE*
Ejection Fraction	*0.83 (0.81-0.85)*	*4.79 (4.60-4.97)*
Pulmonary Artery Pressure	*0.43 (0.39-0.47)*	*7.30 (7.07-7.80)*
Wall Motion Score	*0.77 (0.74-0.80)*	*0.10 (0.10-0.11)*
TAPSE	*0.24 (0.05-0.49)*	*0.35 (0.28-0.45)*
MV Area	*0.34 (0.20-0.44)*	*0.81 (0.71-0.93)*
Transaortic velocity	*0.65 (0.58-0.72)*	*23.84 (20.38-27.82)*
Transaortic gradient	*0.36 (0.29-0.42)*	*6.84 (6.25-7.45)*
TRgradient	*0.28 (0.16-0.39)*	*6.95 (6.51-7.46)*
Sinus of Valsalva Size	*0.62 (0.55-0.67)*	*0.24 (0.22-0.26)*
Ascending Aorta Size	*0.16 (0.02-0.45)*	*0.41 (0.29-0.63)*
IVC Diameter	*0.70 (0.65-0.75)*	*2.85 (2.65-3.06)*

One important echo capability not well demonstrated was detection and quantitation of regional wall motion abnormalities. It is unclear if EchoPrime is not capable of doing this natively or if it just wasn't part of the test suite but should be mentioned in either case.

Reply: Thank you for this important comment, EchoPrime is able to describe regional wall motion abnormalities. We add wall motion score as an additional metric to support previous argument.

Supplementary Table 7 *EchoPrime performance on regression tasks on the internal held-out test set*

Task	*R^2*	*MAE*
Ejection Fraction	*0.83 (0.81-0.85)*	*4.79 (4.60-4.97)*
Pulmonary Artery Pressure	*0.43 (0.39-0.47)*	*7.30 (7.07-7.80)*
Wall Motion Score	***0.77 (0.74-0.80)***	***0.10 (0.10-0.11)***
TAPSE	*0.24 (0.05-0.49)*	*0.35 (0.28-0.45)*
MV Area	*0.34 (0.20-0.44)*	*0.81 (0.71-0.93)*
Transaortic velocity	*0.65 (0.58-0.72)*	*23.84 (20.38-27.82)*
Transaortic gradient	*0.36 (0.29-0.42)*	*6.84 (6.25-7.45)*
TRgradient	*0.28 (0.16-0.39)*	*6.95 (6.51-7.46)*
Sinus of Valsalva Size	*0.62 (0.55-0.67)*	*0.24 (0.22-0.26)*
Ascending Aorta Size	*0.16 (0.02-0.45)*	*0.41 (0.29-0.63)*
IVC Diameter	*0.70 (0.65-0.75)*	*2.85 (2.65-3.06)*

This becomes quite relevant in their auxiliary metric of detecting STEMI (which probably should be spelled out at first mention for the medically uninitiated). It is unclear just what you mean by STEMI; is it an acute infarct or does a chronic regional wall motion abnormality count? Does the model understand the connection between wall motion and coronary territories?

Reply: This comparison is in the limited set of patients that had both ECGs and echocardiograms. Patients diagnosed with STEMI through conventional clinical definition (ST changes on ECG) were used to see if EchoPrime can identify these patients compared to patients with chronic WMA and other chronic diseases. We modify the main text to clarify that the labels were obtained from other modalities:
In **Methods** "**Disease Diagnosis with Probing**" section:

> ***Multimodal Disease Diagnosis***
> *We apply k-nearest neighbors (KNN) and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyhloidosis as two diseases for which echocardiography is not the definitive test*

echocardiography can find corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study, and used along with clinician-assigned diagnoses to fit the models. For KNN probing, we use KNeighborsClassifier, and for linear probing, we apply LogisticRegression, both from scikit-learn library.

I'm also confused by predicate for the STEMI (and amyloidosis), where you state that these are diagnoses "not typically diagnosed with echocardiography," as I would consider both of these to be "bread and butter" echo diagnoses. It also appears that you only considered the Ap4 view for STEMI, which would miss many RCA and LCx infarcts. Similarly, your focus on amyloid uses only the PLAx view, which would miss many of the key findings in amyloid (mid-septal stripe and apical strain sparing, for example).

Reply: Thank you for this comment. We sought to clarify that echocardiography is not the definitive test of choice for diagnosing STEMI or cardiac amyloidosis. ECGs, nuclear medicine, or MRI tests will confirm the diagnosis, while it is correct that echocardiography can find corroborating evidence. We clarify in this revision by replacing the phrase "non-echocardiographic diagnoses" with "multimodal cardiovascular diagnoses" and how this is an approach to use EchoPrime's pretrained weights for non-'echocardiographic' tasks.

In **Methods** "**Disease Diagnosis with Probing**" section:

Multimodal Disease Diagnosis
We apply k-nearest neighbors (KNN) and linear probing on top of EchoPrime for multimodal cardiovascular diagnoses prediction. We focused on STEMI and cardiac amyhloidosis as two diseases for which echocardiography is not the definitive test echocardiography can find corroborating evidence. In this setting, patients with clinically diagnosed STEMI or cardiac amyloidosis were identified and the EchoPrime embeddings were used to evaluate whether linear probing can identify patients with this diagnosis despite not being in the echo report text. Study embeddings are obtained by averaging the embeddings of all videos within each study, and used along with clinician-assigned diagnoses to fit the models. For KNN probing, we use KNeighborsClassifier, and for linear probing, we apply LogisticRegression, both from scikit-learn library.

This model clearly is an important advance in AI-guided echo interpretation and, after appropriate revision, quite worthy of publication. One possible limitation is that the model does not actually write the interpretation language but, as I understand it, rather finds the best fitting verbiage from the 275K reports generated at Cedars Sinai over the last ten years (admittedly, these will cover most every nuance!).

Reply: Thank you for this important feedback. EchoPrime does not generate but retrieves the

report section-by-section. We see that as beneficial to minimize hallucinations and downstream implementation (as most echo labs use structured reporting and will require output in specific phrase formats to be placed into the clinical report).There are 15 sections and 275k candidate reports, meaning that EchoPrime picks one from 15^(275_000) candidate possibilities, which we believe covers a comprehensive and wide range of echocardiographic findings and interpretations.

Another critique is the choice of batch size = 32. The key point of contrastive learning is to bring like embeddings (text reports and their corresponding videos, i.e., the positive example) closer together while pushing negative pairs far apart. This necessitates a sufficiently large batch size to create enough differentiation within the latent embedding space. You trained on only 2 A6000's, raising the possibility of a limitation of computing power, which would warrant some further discussion and ideally ablation studies showing the impact of different batch sizes.

Reply: Thank you for making this point. In this revision, we show new experiments were we train EchoPrime with batch sizes of 16, 32 and 50, and show all batch sizes have similar performance with likely asymptoming gains around a batch size of 32 (original hyperparameter choice). While these ablations offer useful insights into batch size effects, the differences observed with batch size 50 are minimal and do not justify a switch in the primary model configuration.

Supplementary Table 10 *EchoPrime fine-tuned on different batch sizes*

BATCH SIZE	***16***	***32***	***50***
LV ejection fraction (R^2)	*0.81 (0.79-0.83)*	***0.83 (0.81-0.85)***	*0.80 (0.78-0.82)*
LV ejection fraction (MAE)	*4.87 (4.69-5.06)*	***4.79 (4.60-4.97)***	*4.94 (4.75-5.13)*
Pacemaker	*0.85 (0.82-0.87)*	*0.85 (0.82-0.88)*	***0.87 (0.84-0.89)***
RV systolic function	*0.93 (0.90-0.95)*	*0.93 (0.91-0.95)*	*0.93 (0.90-0.95)*
RV dilation	*0.88 (0.82-0.94)*	*0.89 (0.83-0.95)*	***0.92 (0.87-0.96)***
LA dilation	*0.92 (0.87-0.96)*	*0.91 (0.86-0.96)*	***0.94 (0.90-0.97)***
RA dilation	*0.94 (0.88-0.98)*	*0.90 (0.82-0.97)*	*0.94 (0.88-0.99)*
Mitraclip	*0.99 (0.99-1.00)*	*0.99 (0.99-1.00)*	*0.99 (0.99-1.00)*
Mitral annular calcification	*0.96 (0.94-0.97)*	*0.96 (0.95-0.97)*	*0.95 (0.92-0.97)*
Mitral stenosis	*0.91 (0.82-0.97)*	***0.96 (0.91-0.99)***	*0.92 (0.85-0.99)*
Mitral regurgitation	*0.93 (0.92-0.95)*	*0.92 (0.91-0.94)*	*0.93 (0.92-0.95)*
TAVR	*0.99 (0.98-1.00)*	*1.00 (0.99-1.00)*	*0.99 (0.98-1.00)*
Bicuspid AV	*0.83 (0.67-0.96)*	*0.83 (0.67-0.96)*	*0.83 (0.67-0.96)*
Aortic stenosis	*0.98 (0.96-0.99)*	*0.98 (0.96-0.99)*	*0.98 (0.96-0.99)*

Aortic regurgitation	*0.87 (0.82-0.92)*	*0.88 (0.83-0.93)*	***0.90 (0.85-0.94)***
Tricuspid regurgitation	*0.95 (0.94-0.97)*	*0.95 (0.93-0.97)*	***0.96 (0.94-0.97)***
Pericardial effusion	***0.99 (0.98-0.99)***	*0.98 (0.95-0.99)*	*0.97 (0.93-0.99)*
Aortic root dilation	***0.92 (0.83-0.99)***	*0.91 (0.83-0.98)*	*0.91 (0.82-0.98)*
Dilated IVC	*0.83 (0.81-0.85)*	***0.84 (0.82-0.86)***	*0.83 (0.81-0.85)*
PA pressure (R^2)	*0.42 (0.37-0.46)*	***0.43 (0.39-0.47)***	*0.42 (0.38-0.47)*
PA pressure (MAE)	*7.31 (6.94-7.68)*	*7.30 (6.95-7.67)*	***7.26 (6.89-7.62)***

A potential critique of multiple instance learning is the risk that your inference pipeline is over-engineered, carrying strong inductive bias with a classifier and the weighting objective. Have you experimented at all with learning the anatomical priors outside of a two-step system, i.e. a learned pooling of embeddings to create the study and anatomical embeddings?

Reply: Thank you for this important point. In Figure 2A and newly added Supplementary Table 9 we directly compare the performance with and without anatomical attention and show the approach with anatomical attention is superior. The trained recognition of important features from the most relevant videos (anatomic attention) is important and helpful, however there are many approaches that might be proposed for pooling and ensembling approaches. In our discussion, we describe this as a direction for future investigation.

In **Results "Leveraging multi-view echocardiographic data improves performance**" section:

To demonstrated the impact of integrating multiple clips, multiple videos, and multiple views, we compared the performance of the EchoPrime encoder when input is a single frame vs. single video vs. multiple videos vs. multiple videos with multiple instance attention weighting (**Error! Reference source not found.**A). The gradual improvement across multiple tasks reflects the relevance of temporal and view-specific information to each task and the importance of the anatomic attention module (Supplementary Table 9).

In **Discussion**:

> Additional work is required to continue advancing medical AI models, including the development of multimodal models that integrate complementary diagnostic modalities and recognize task-relevant features as well as clinical trials of AI in a clinical setting.

Supplementary Table 9 EchoPrime performance with and without anatomical weighting

	WITHOUT	**WITH**	**STANFORD W/O**	**STANFORD WITH**
LV ejection fraction (R^2)	0.80 (0.78-0.82)	**0.83 (0.81-0.85)**	0.78 (0.75-0.81)	0.79 (0.76-0.82)

LV ejection fraction (MAE)	4.97 (4.78-5.17)	**4.79 (4.60-4.97)**	4.22 (4.01-4.42)	4.14 (3.94-4.34)
Pacemaker	0.82 (0.80-0.85)	**0.85 (0.82-0.88)**	0.80 (0.77-0.83)	0.84 (0.81-0.87)
RV systolic function	0.92 (0.90-0.94)	**0.93 (0.91-0.95)**	0.91 (0.80-1.00)	0.94 (0.88-1.00)
RV dilation	**0.90 (0.85-0.95)**	0.89 (0.83-0.95)	0.83 (0.77-0.88)	0.85 (0.80-0.90)
LA dilation	0.91 (0.86-0.96)	0.91 (0.86-0.96)	0.71 (0.68-0.73)	0.73 (0.70-0.75)
RA dilation	**0.92 (0.85-0.98)**	0.90 (0.82-0.97)	0.76 (0.73-0.80)	0.77 (0.74-0.81)
Mitraclip	0.98 (0.96-0.99)	**0.99 (0.99-1.00)**	0.97 (0.93-1.00)	0.98 (0.94-1.00)
Mitral annular calcification	0.95 (0.93-0.96)	**0.96 (0.95-0.97)**	0.96 (0.93-0.98)	0.96 (0.94-0.98)
Mitral stenosis	0.96 (0.90-0.99)	0.96 (0.91-0.99)	0.93 (0.86-0.98)	0.92 (0.86-0.98)
Mitral regurgitation	0.91 (0.89-0.93)	**0.92 (0.91-0.94)**	0.89 (0.85-0.93)	0.91 (0.89-0.93)
TAVR	0.99 (0.99-1.00)	**1.00 (0.99-1.00)**	0.99 (0.98-1.00)	0.97 (0.90-1.00)
Bicuspid AV	0.82 (0.67-0.96)	**0.83 (0.67-0.96)**	0.78 (0.69-0.87)	0.82 (0.73-0.90)
Aortic stenosis	0.96 (0.93-0.98)	**0.98 (0.96-0.99)**	0.94 (0.89-0.99)	0.96 (0.92-0.99)
Aortic regurgitation	0.83 (0.77-0.88)	**0.88 (0.83-0.93)**	0.84 (0.72-0.94)	0.89 (0.80-0.97)
Tricuspid regurgitation	0.94 (0.91-0.96)	**0.95 (0.93-0.97)**	0.84 (0.79-0.88)	0.88 (0.86-0.91)
Pericardial effusion	0.95 (0.92-0.98)	**0.98 (0.95-0.99)**	0.88 (0.75-0.98)	0.89 (0.77-0.98)
Aortic root dilation	0.88 (0.79-0.96)	**0.91 (0.83-0.98)**	0.94 (0.90-0.98)	0.97 (0.94-0.99)
Dilated IVC	0.82 (0.80-0.84)	**0.84 (0.82-0.86)**	0.82 (0.74-0.88)	0.86 (0.81-0.91)
PA pressure (R^2)	0.41 (0.37-0.45)	**0.43 (0.39-0.47)**	0.33 (0.27-0.39)	0.36 (0.30-0.42)
PA pressure (MAE)	7.44 (7.10-7.82)	**7.30 (6.95-7.67)**	7.89 (7.21-8.60)	7.97 (7.35-8.66)

Retrieval Augmented Interpretation (page 14, par 2): I'm not sure why this was described uniquely in comparison to Retrieval Augmented Generation. If Retrieval Augmented Interpretation has been described elsewhere this should be cited in this section. Otherwise,

Retrieval Augmented Interpretation seems to describe how CLIP models are used as zero shot __ (where __ is a regression task, classifier, or 'report generator'). Probe text is encoded using the text encoder and a similarity metric is used to retrieve the most relevant text based on the similarity between the video and text encodings. So, a description of the 'historical' reports that are retrieved is necessary here. Are these reports all from the training set or a different set of reports?

Reply:
We note the difference between RAI and RAG because RAG is traditionally connected to large language models. RAI is also different than traditional CLIP retrieval because it retrieves the report section-by-section. The historical reports are all reports from the training set. We modify the text to clarify the source of candidate reports:

*In **Methods** "**Retrieval Augmented Interpretation**" section*

> *Using pre-trained parametric memory (EchoPrime weights) and non-parametric memory (clinical reports from the train set), our approach leverages anatomic attention to weight the cross-modal retrieval based on each section in an echocardiogram report.*

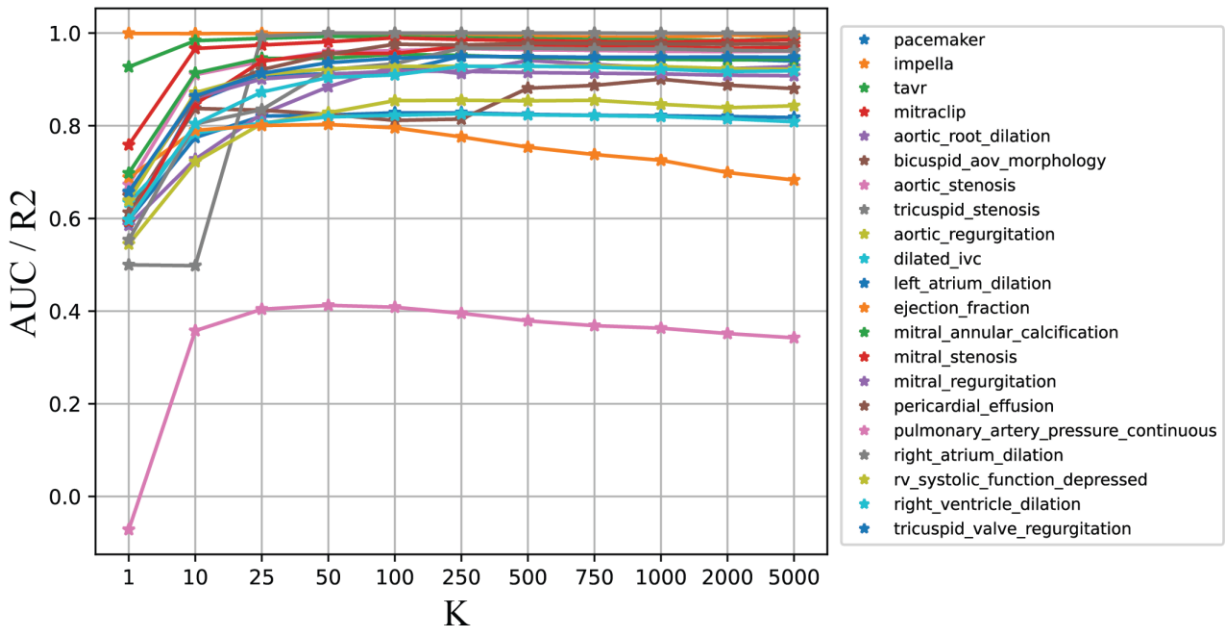
In addition, you average 50 retrieved reports to generate the interpretation. Evidence for k=50 reports being optimal? Also, what is meant by averaging since this is at the text level?

Reply:

We add the plot (Supplementary Figure 4) showing that k=50 is optimal hyperparameter including considerations of performance and runtime. Specifically, for values k>50 the performance doesn't improve but the runtime increases, that is why k=50 was chosen.
The averaging is happening at the feature level. Features are extracted with regex and custom defined phrases (https://github.com/echonet/EchoPrime/blob/main/per_section.json).

*In **Methods "Retrieval Augmented Interpretation"** section*

> *Hyperparameter k=50 is chosen because it achieves the optimal trade-off between performance and runtime (Supplementary Figure 4).*



Supplementary Figure 4 *Performance of Retrieval Based Interpretation as we increase the number of candidates reports to average across. For binary tasks the metric is AUROC and for regression tasks R2 score.*

Along the same lines there is no discussion as to why mViT was chosen as the video encoder and no ablations for other video encoding architectures. Additionally, the same can be said about using PubMedBERT for the text encoder. In fact, in the previous iteration, EchoCLIP dedicated significant discussion space to describing a unique tokenization technique that took advantage of the templated structure of Echo reports. This was shown to be better than standard Byte Pair Encoding with the CLIP text encoder. I think the switch to PubMedBERT and Wordpiece encoding warrants deeper discussion and some comparison studies with the EchoCLIP text encoder and the text encoder used in EchoPrime.

Reply: Thank you for your insightful comments. Due to the significant computational demands, we were unable to conduct extensive ablation studies on alternative video and text encoder architectures. For context, training our main checkpoint required approximately five weeks on our compute cluster. Our primary focus was on demonstrating EchoPrime's practical accuracy and its superior performance compared to prior biomedical foundation models. For instance, the comparison of EchoPrime, a model that uses PubMedBERT, and EchoCLIP-R, a model that uses custom BPE, is presented in Table 3. While additional architectural ablations could provide further insights, we believe that EchoPrime's end-to-end improvements already establish its effectiveness over previous approaches.

As per reviewer request me motivate our architectural choices.

*In **Methods "Contrastive Visual-Language Training"** section:*

> *mVIT was chosen as the video encoder because, in the original CLIP paper[4], a transformer-based encoder outperformed a ResNet-based encoder. Additionally, mVIT effectively captures temporal dynamics, as evidenced by performance changes when video frames are shuffled—a phenomenon not observed in other vision transformers[30]. PubMedBERT was selected as the text encoder due to its pretraining on 21GB of medical text, giving our model a head start in understanding medical literature.*

Table 1: Cross-modal retrieval metrics on a test set of 2,254 videos-text report pairs. Lower rank is better, and higher recall is better. Best performance is bolded

	Videos-To-Text					Text-To-Videos				
	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10
EchoPrime	**2.86**	**1**	**70%**	**94%**	**98%**	**3.05**	**1**	**69%**	**94%**	**97%**
EchoCLIP	58.71	9	20%	42%	53%	37.07	5	25%	50%	62%

I would note that there is break in the text continuity between page 5 and 6. It is unclear if there is any lost text here (I suspect not), but it should be corrected.

Reply: Thank you for noticing this. It is indeed a spacing issue after the parentheses, we fix it in the revision.

The authors have provided model code with all initialization parameters at their github. While I have not implemented this for my own testing, it does appear that it is in standard form and should be readily implemented with the appropriate subroutine libraries, which they specified. In addition, you provide a video example of EchoPrime in operation, interpreting a relatively normal echo. This overall is effective in showing model operation, but I am puzzled by the lack of any spectral Doppler frames in the example data. It is simply impossible to estimate RV pressure, diastolic function, or valve stenosis without Doppler, so this should be redone with the appropriate needed images. It should be noted that you have not provided the echo training data, but this seems reasonable. Not only would this likely exceed a petabyte in size, but despite all best efforts to strip identifying information from the images, such PHI might still be present. Of note, you do provide a sample casein DICOM format, though again I only see video clips, no M-mode nor spectral Doppler. Unclear if this is the same case shown in the model operation example. You certainly need to include the single frame images and specify if this is the case shown in the MOV file.

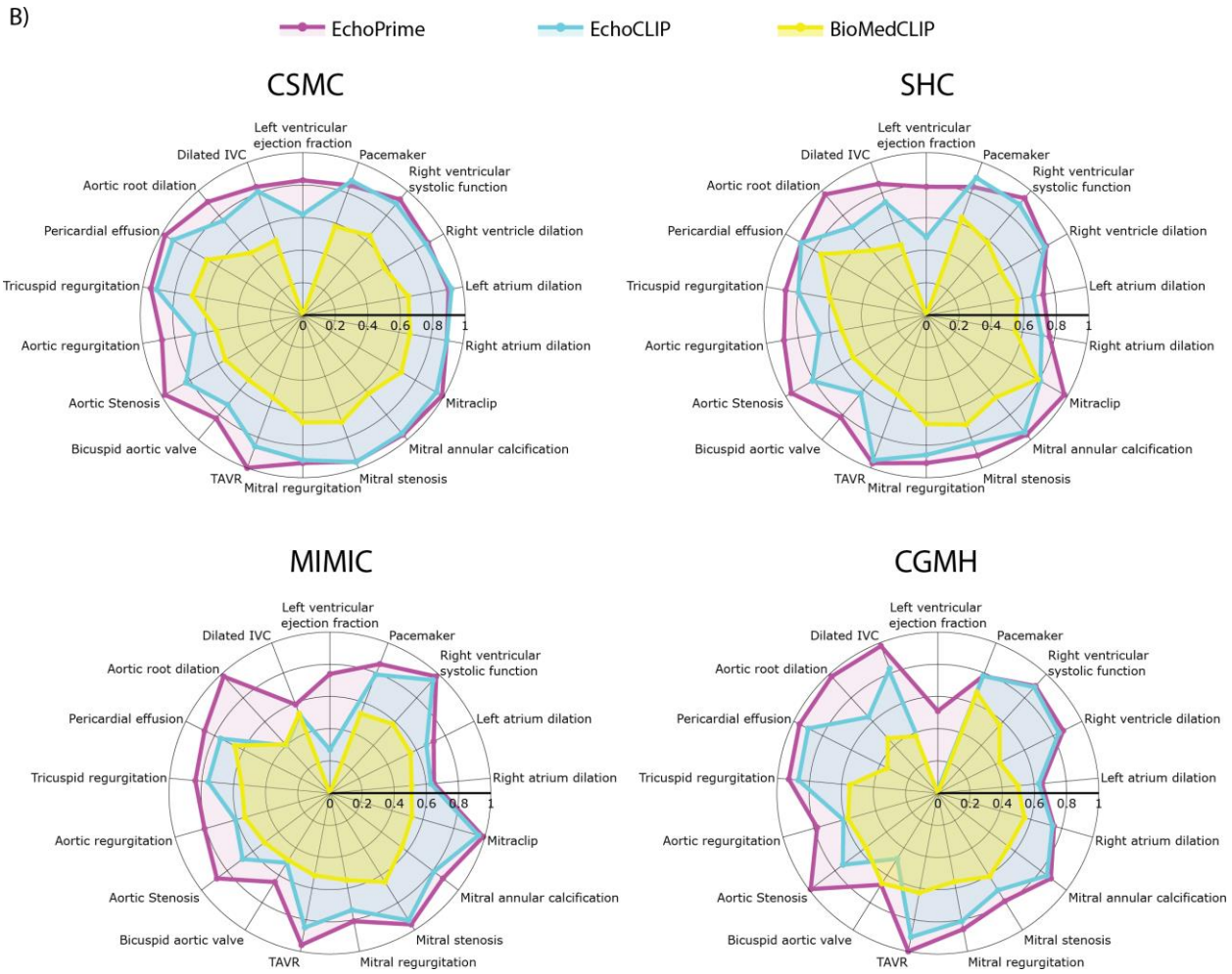
Reply: EchoPrime was trained on B-mode and Doppler videos and used a video-based encoder. Much/all of the Spectral doppler and M mode images information images are obtainable from

standard echocardiographic videos from the same study. For example, the M mode has time in the x axis and pixel intensity in the y, however all that necessary information is inherently already present in the B-mode video. Similarly, the Spectral doppler information is already present in the color Doppler videos and we hypothesize can be extracted with a video encoder from other videos in the study.

To confirm this finding, in this revision, we introduce two new external validation sets (echocardiogram studies from MIMIC and CGMH which have B-mode, Doppler, Spectral doppler, and M mode images) and show EchoPrime performs similarly in these datasets as our other internal and external test datasets even with ignoring Spectral doppler and M mode images.

*In **Results "Automated echocardiogram interpretation without supervised learning" section***

> *On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging (Figure2B, Supplementary Tables 2-5)*



Supplementary Table 4 *Performance of foundation models for biomedical imaging on MIMIC dataset from Boston*

MIMIC DATASET			
	BioMed CLIP	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-2.48 (-3.26--1.85)*	*0.27 (0.13-0.40)*	***0.74 (0.66-0.80)***
LV ejection fraction (MAE)	*26.96 (25.31-28.58)*	*10.37 (9.30-11.51)*	***6.21 (5.54-6.92)***
Pacemaker	*0.53 (0.39-0.68)*	*0.79 (0.66-0.90)*	***0.86 (0.80-0.92)***
RV systolic function	*0.58 (0.02-0.90)*	*0.95 (0.89-0.99)*	***0.98 (0.97-0.99)***
RV dilation	*-*	*-*	*-*
LA dilation	*0.56 (0.52-0.59)*	*0.67 (0.64-0.71)*	***0.72 (0.68-0.75)***
RA dilation	*0.51 (0.47-0.54)*	*0.63 (0.59-0.67)*	***0.65 (0.61-0.69)***
Mitraclip	*0.53 (0.29-0.99)*	*0.96 (0.86-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.56 (0.51-0.60)*	*0.81 (0.77-0.85)*	***0.88 (0.85-0.91)***
Mitral stenosis	*0.65 (0.48-0.98)*	*0.93 (0.80-1.00)*	***0.96 (0.95-0.98)***
Mitral regurgitation	*0.55 (0.52-0.58)*	*0.74 (0.71-0.76)*	***0.81 (0.79-0.83)***
TAVR	*0.52 (0.44-0.60)*	*0.85 (0.79-0.90)*	***0.96 (0.94-0.98)***
Bicuspid AV	*0.49 (0.46-0.56)*	*0.51 (0.42-0.61)*	***0.65 (0.52-0.78)***
Aortic stenosis	*0.51 (0.48-0.54)*	*0.68 (0.64-0.73)*	***0.88 (0.84-0.91)***
Aortic regurgitation	*0.55 (0.49-0.61)*	*0.61 (0.54-0.67)*	***0.81 (0.75-0.86)***
Tricuspid regurgitation	*0.55 (0.52-0.58)*	*0.76 (0.73-0.78)*	***0.84 (0.82-0.86)***
Pericardial effusion	*0.66 (0.61-0.72)*	*0.76 (0.70-0.81)*	***0.87 (0.83-0.92)***
Aortic root dilation	*0.41 (0.40-0.42)*	*0.40 (0.39-0.40)*	***0.98 (0.96-0.99)***
Dilated IVC	*0.53 (0.44-0.63)*	*0.52 (0.40-0.64)*	***0.59 (0.49-0.69)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Supplementary Table 5 *Performance of foundation models for biomedical imaging on CGMH dataset from Taiwan*

CGMH DATASET

	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	*-8.31 (-9.74--7.06)*	*0.02 (-0.13-0.15)*	***0.51 (0.44-0.57)***
LV ejection fraction (MAE)	*32.64 (31.89-33.38)*	*9.15 (8.75-9.54)*	***6.46 (6.19-6.74)***
Pacemaker	*0.67 (0.56-0.80)*	***0.78 (0.59-0.92)***	***0.78 (0.63-0.92)***
RV systolic function	*0.57 (0.41-0.72)*	*0.89 (0.84-0.94)*	***0.90 (0.79-0.98)***
RV dilation	*0.43 (0.30-0.61)*	*0.84 (0.65-0.97)*	***0.87 (0.69-0.97)***
LA dilation	*0.50 (0.47-0.53)*	*0.63 (0.60-0.65)*	***0.65 (0.62-0.67)***
RA dilation	*0.56 (0.47-0.64)*	*0.74 (0.65-0.83)*	***0.75 (0.67-0.84)***
Mitraclip	*-*	*-*	*-*
Mitral annular calcification	*0.55 (0.40-0.70)*	*0.85 (0.69-0.97)*	***0.88 (0.75-0.96)***
Mitral stenosis	*0.61 (0.48-0.79)*	*0.71 (0.51-0.92)*	***0.79 (0.59-0.98)***
Mitral regurgitation	*0.56 (0.50-0.63)*	*0.81 (0.76-0.86)*	***0.86 (0.82-0.89)***
TAVR	*0.63 (0.50-0.77)*	*0.91 (0.83-0.97)*	***1.00 (1.00-1.00)***
Bicuspid AV	*0.66 (0.43-1.00)*	*0.48 (0.35-0.77)*	***0.67 (0.45-1.00)***
Aortic stenosis	*0.56 (0.43-0.71)*	*0.74 (0.59-0.88)*	***0.99 (0.98-1.00)***
Aortic regurgitation	*0.58 (0.50-0.66)*	*0.61 (0.53-0.69)*	***0.78 (0.70-0.84)***
Tricuspid regurgitation	*0.55 (0.49-0.61)*	*0.87 (0.83-0.91)*	***0.93 (0.90-0.95)***
Pericardial effusion	*0.35 (0.16-0.65)*	*0.90 (0.69-1.00)*	***0.96 (0.91-1.00)***
Aortic root dilation	*0.46 (0.39-0.62)*	*0.64 (0.41-0.86)*	***0.98 (0.97-1.00)***
Dilated IVC	*0.38 (0.37-0.39)*	*0.83 (0.50-1.00)*	***0.98 (0.94-1.00)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Referee #3 (Remarks to the Author):

A. Summary of Key Results
The authors propose a novel vision-language model trained on 12 million samples to interpret echocardiography studies by performing view classification, mapping videos into a video-text latent space, and retrieving study interpretations. Experiments on both internal and external datasets demonstrate that EchoPrime outperforms both task-specific approaches and prior foundation models. The authors introduce EchoPrime, which incorporates multi-view information to enhance the accuracy of echocardiography interpretations—an innovative approach. Additionally, they collected an extensive dataset of 12 million samples, highlighting the scale of their work.

Reply: Thank you for your feedback. EchoPrime is the largest AI model in echocardiography in terms of the dataset size used for training. It was trained on an unparalleled dataset of 12 million echocardiogram videos, synthesizing all videos within each study and their corresponding 100–500-word reports, and this scale of training data helps explain its unparalleled performance on a wide range of clinical interpretation tasks.

B. Originality and Significance
However, although the authors claim EchoPrime can assist physicians in the automated preliminary assessment of comprehensive echocardiography after rigorous clinical evaluation, no human evaluation has been conducted; the assessment relies solely on metrics such as AUC, limiting the study's significance. While the dataset used is comprehensive, the analysis is purely retrospective. Prospective experiments would align more directly with clinical applications, providing a more robust validation of EchoPrime's effectiveness in real-world clinical settings. The lack of prospective analysis significantly limits the study's overall validity and impact.

Reply: Thank you for this important consideration. In this model development work, we have not undergone clinical trial assessment of EchoPrime. That is currently ongoing, and we are excited to present these results likely in late 2025 / early 2026.

Despite this limitation, we sought to comprehensively assess EchoPrime's performance in a wide range of tasks and settings. In this revision, we introduce two new external validation sites (one from Boston and one from Taiwan) to show international validation on a wide range of potential image types/characteristics and patient populations. We think this is a rigorous assessment of generalizability of EchoPrime across many clinical settings and practice patterns, however we recognize that this is still a retrospective evaluation.

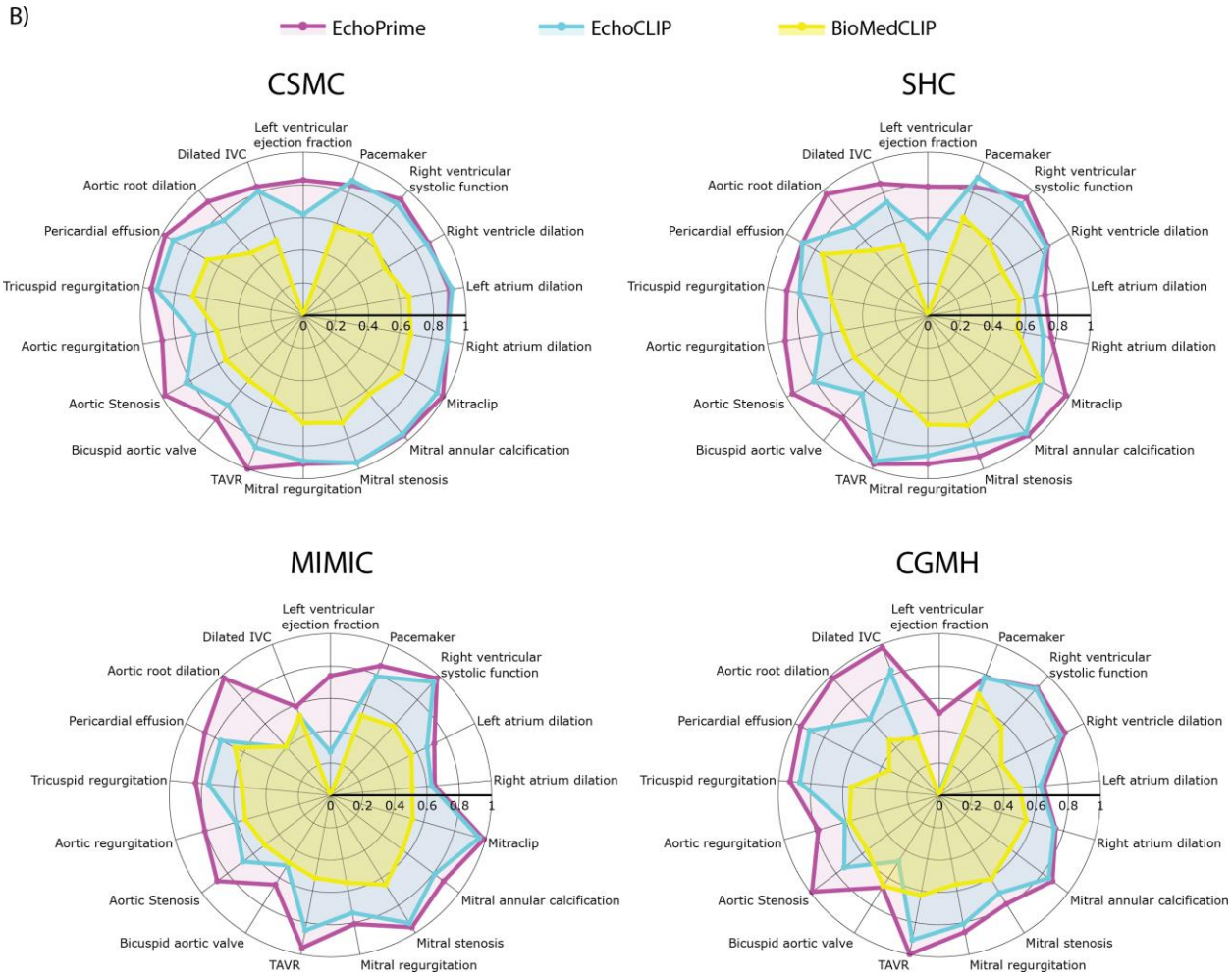


Figure 2B

In this revision, we provide more discussion of this critical point in the discussion portion of the manuscript, highlighting the need for prospective evaluation.

*In the **Discussion**:*

> *A few limitations are worth considering. Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations under diverse and variable image acquisition protocols to ensure robust and generalizable performance and to measure its practical value to physicians. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information.*

C. Data & Methodology

Video Encoder: The authors state that their video encoder is trained on a mViT checkpoint;

however, mViT dates back to 2021. Recent advancements such as BEVT [1] and VideoMAE [2] could enhance relevance and demonstrate a thorough exploration of state-of-the-art methods.

Reply: Thank you for your comment. BEVT and VideoMAE propose strategies for training foundation models using unsupervised masking, which is particularly beneficial when only videos are available. In our project, however, we have access to both videos and their corresponding reports. To leverage the additional information from the reports, we adopt a contrastive learning approach to build a joint video-text representation space.

Foundation models in the video-language domain generally fall into three categories: contrastive, masking, and generative models [1]. EchoPrime belongs to the contrastive family, and while comparisons across these three families are undoubtedly valuable, such an analysis is beyond the scope of this paper. Our focus is on demonstrating the effectiveness of the contrastive approach for this specific application. With the design choices we have implemented, EchoPrime has superior performance across a wide range of tasks and exceeds or parallels all publicly available echocardiographic AI models. Further work will continue to improve upon this model as the AI space is rapidly evolving, but we believe this is a important milestone in AI for echocardiography.

In our discussion, we now note the importance of new models that might be interesting avenues for future investigation.

[1.] Bordes, F. *et al.* An Introduction to Vision-Language Modeling. Preprint at http://arxiv.org/abs/2405.17247 (2024).

*In the **Discussion**:*

> *A few limitations are worth considering. Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations under diverse and variable image acquisition protocols to ensure robust and generalizable performance and to measure its practical value to physicians. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information.*

Text Encoder: EchoPrime supports over 500 tokens and utilizes WordPiece encoding within a BERT architecture. The authors state, “This marks a significant improvement over previous medical foundation models like BioMedCLIP and EchoCLIP, which process only single-view, individual images and handle text up to 77 tokens.” However, no ablation studies on token length are provided, which are necessary to validate the benefit of increasing the token limit. Comparisons with other models alone do not substantiate that this design choice is significantly advantageous.

Reply: Thank you for your comment. The main motivation behind extending the context length of the text encoder is to be able to fit full echocardiography reports during training. For example,

the following report requires 404 tokens by Wordpiece tokenizer, but would be truncated by CLIP's encoder (which only allows for 77 tokens):

What EchoPrime sees:

'Aorta: The aortic root is normal in size. Sinus of Valsalva: 3.3 cm. The ascending aorta is normal in size. Ascending Aorta 3.5 cm. Aortic Valve: No significant aortic stenosis or insufficiency. Trileaflet aortic valve. The aortic cusps appear mildly thickened. The peak transaortic gradient is 5 mmHg The mean transaortic gradient is 3 mmHg Atrial Septum: The interatrial septum is normal in appearance. IVC: The inferior vena cava is of normal size. The IVC diameter is 18 mm The inferior vena cava shows a normal respiratory collapse consistent with normal right atrial pressure (3 mmHg). Left Atrium: The left atrium is normal in size. Left Ventricle: Normal left ventricular size by linear cavity dimension. Normal left ventricular size by volume Mild left ventricular hypertrophy. Normal left ventricular systolic function. LV Ejection Fraction is 60 %. Mild diastolic dysfunction. There is reversal of the E to A ratio and/or prolonged deceleration time consistent with impaired left ventricular relaxation. Doppler parameters and/or lateral mitral annular (E`) velocities are consistent with normal left ventricular filling pressures. Mitral Valve: The mitral valve demonstrates normal leaflet morphology. The mitral valve demonstrates normal function with trace physiologic regurgitation. Pericardium: Normal pericardium with no pericardial effusion. Pulmonary Artery: Estimated PA Pressure is 16 mmHg. PA systolic pressure is normal. Pulmonary Veins: Pulmonary veins are normal in appearance and pulse Doppler interrogation shows normal systolic predominant flow. Pulmonic Valve: Normal pulmonic valve appearance. Normal pulmonic valve function with trace physiologic regurgitation. Resting Segmental Wall Motion Analysis: Total wall motion score is 1.00. There are no regional wall motion abnormalities Right Atrium: The right atrium is normal in size. Right Ventricle: Normal right ventricular size. Normal right ventricular systolic function. Tricuspid Valve: Normal appearance of the tricuspid valve. Est RV/RA pressure gradient is 13 mmHg. Estimated peak RVSP is 16 mmHg. There is trivial tricuspid regurgitation.'

What EchoCLIP sees:

<|startoftext|>aorta : the aortic root is normal in size . sinus of valsalva : 3 . 3 cm . the ascending aorta is normal in size . ascending aorta 3 . 5 cm . aortic valve : no significant aortic stenosis or insufficiency . trileaflet aortic valve . the aortic cusps appear mildly thickened . the

The ability to process the full echocardiography reports is a fundamental aspect of EchoPrime's design, allowing it to leverage the complete context provided by the detailed reports during training, while a much longer context is not useful if reports are too short to utilize a much longer context. While we appreciate the suggestion for ablation studies on token length, this has significant computational cost and the choice of context length was rationaly designed based on the ability to accommodating full reports. Truncating reports, as required by models with limited token capacity like CLIP, likely results in loss of performance.

Method: The evaluation of EchoPrime lacks comprehensiveness. While the authors selected two foundation models, BioMedCLIP and EchoCLIP, additional comparisons, such as with PubMedCLIP [3], PMC-CLIP [4], and ROCO [5], would provide a broader context.

Reply: Thank you for your feedback. In the original manuscript, we chose benchmarks that have highest likelihood of the strongest performance (BioMedCLIP has the strongest performance among foundation models trained on publicly available data and EchoCLIP is the most comparable echocardiographic foundation model). In this revision, we had the recommended comparison models (PubMedCLIP and PMC-CLIP) and show that they perform worse than EchoPrime (**Supplementary Tables 2)**. Of note, ROCO. [5] doesn't introduce a new machine learning model but rather is a dataset that can be used for training contrastive models (such as

PubMedCLIP and PMC-CLIP).

The performance of general medical data foundation models is poor on echocardiography tasks (likely due to the small size of echocardiography videos present in their training set). Specifically, PubMedCLIP achieves an average of 0.50 across AUROC metrics, which is essentially random performance, and PMC-CLIP performs slightly better than random with 0.532 average across AUROC metrics. BioMedCLIP achieves an average of 0.608 across AUROC metrics but still far worse than EchoCLIP (average AUC =0.866) and EchoPrime (average AUC = 0.922), foundation models trained specifically for echocardiography.

In the ***Results*** *"****Automated echocardiogram interpretation without supervised learning****"*

> *We evaluated EchoPrime on a diverse set of benchmarks for cardiac structure and function across datasets from four different institutions (****Error! Reference source not found.****). On echocardiographic data from all four geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging [17,19–21] (Figure 2B,* ***Supplementary Tables 2****-5),*

Supplementary Table 2 *Performance of foundation models for biomedical imaging on Cedars-Sinai test set*

CEDARS-SINAI TEST SET					
	PubMedCLIP	*PMC-CLIP*	*BioMed CLIP*	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-0.32 (-0.40--0.25)*	*-0.38 (-0.46--0.31)*	*-2.60 (-3.00 - -2.20)*	*0.62 (0.59-0.65)*	***0.83 (0.81-0.85)***
LV ejection fraction (MAE)	*14.68 (14.30-15.06)*	*14.03 (13.58-14.49)*	*26.93 (26.42-27.44)*	*7.00 (6.74-7.26)*	***4.79 (4.60-4.97)***
Pacemaker	*0.50 (0.50-0.50)*	*0.51 (0.47-0.54)*	*0.58 (0.55-0.62)*	***0.88 (0.85-0.91)***	*0.85 (0.82-0.88)*
RV systolic function	*0.50 (0.49-0.51)*	*0.53 (0.48-0.58)*	*0.64 (0.59-0.69)*	*0.89 (0.86-0.92)*	***0.93 (0.91-0.95)***
RV dilation	*0.50 (0.45-0.56)*	*0.59 (0.50-0.67)*	*0.58 (0.52-0.64)*	*0.88 (0.83-0.92)*	***0.89 (0.83-0.95)***
LA dilation	*0.50 (0.50-0.50)*	*0.57 (0.48-0.65)*	*0.66 (0.58-0.74)*	***0.93 (0.89-0.96)***	*0.91 (0.86-0.96)*
RA dilation	*0.50 (0.50-0.50)*	*0.58 (0.47-0.68)*	*0.67 (0.59-0.75)*	*0.90 (0.82-0.97)*	***0.90 (0.82-0.97)***
Mitraclip	*0.50 (0.50-0.50)*	*0.49 (0.43-0.55)*	*0.70 (0.63-0.76)*	*0.95 (0.92-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.50 (0.46-0.55)*	*0.59 (0.54-0.64)*	*0.63 (0.58-0.68)*	*0.95 (0.93-0.96)*	***0.96 (0.95-0.97)***
Mitral stenosis	*0.50 (0.50-0.50)*	*0.55 (0.44-0.66)*	*0.70 (0.61-0.79)*	***0.96 (0.91-0.99)***	***0.96 (0.91-0.99)***
Mitral regurgitation	*0.50 (0.50-0.50)*	*0.59 (0.55-0.63)*	*0.66 (0.63-0.69)*	*0.89 (0.87-0.91)*	***0.92 (0.91-0.94)***
TAVR	*0.50 (0.50-0.51)*	*0.51 (0.46-0.55)*	*0.54 (0.50-0.57)*	*0.86 (0.83-0.89)*	***1.00 (0.99-1.00)***
Bicuspid AV	*0.50 (0.50-0.50)*	*0.51 (0.47-0.60)*	*0.52 (0.51-0.53)*	*0.72 (0.58-0.85)*	***0.83 (0.67-0.96)***

Aortic stenosis	*0.50 (0.50-0.52)*	*0.48 (0.42-0.55)*	*0.55 (0.53-0.57)*	*0.83 (0.79-0.87)*	***0.98 (0.96-0.99)***
Aortic regurgitation	*0.50 (0.50-0.50)*	*0.50 (0.41-0.58)*	*0.54 (0.53-0.55)*	*0.68 (0.60-0.76)*	***0.88 (0.83-0.93)***
Tricuspid regurgitation	*0.49 (0.45-0.53)*	*0.57 (0.51-0.63)*	*0.69 (0.64-0.74)*	*0.91(0.89-0.93)*	***0.95 (0.93-0.97)***
Pericardial effusion	*0.54 (0.45-0.62)*	*0.50 (0.42-0.57)*	*0.68 (0.59-0.77)*	*0.92 (0.89-0.95)*	***0.98 (0.95-0.99)***
Aortic root dilation	*0.50 (0.50-0.50)*	*0.44 (0.35-0.55)*	*0.50 (0.49-0.51)*	*0.76 (0.63-0.88)*	***0.91 (0.83-0.98)***
Dilated IVC	*0.50 (0.50-0.51)*	*0.53 (0.50-0.56)*	*0.49 (0.47-0.50)*	*0.81 (0.79-0.83)*	***0.84 (0.82-0.86)***
PA pressure (R^2)	*-0.59 (-0.74--0.46)*	*-0.83 (-1.07--0.62)*	*-0.19 (-0.24 - -0.14)*	*0.35 (0.29-0.41)*	***0.43 (0.39-0.47)***
PA pressure (MAE)	*11.27 (9.98-12.62)*	*13.65 (13.04-14.28*	*10.62 (10.12-11.12)*	*7.84 (7.45- 8.23)*	***7.30 (6.95-7.67)***

Data: For the view classification task, the origin of the training dataset is unclear. Is this dataset part of the 12,124,168 echocardiography videos from CSMC as listed in Table 1? If not, please specify the data source and provide statistical details for this specific dataset.

Reply: Thank you for pointing this out. The view classification dataset consists of 77,426 echocardiograms, which is indeed a subset of the 12,124,168 training videos. This subset manually labeled by sonographers into 58 different view categories

We modify the text to be more specific about the source of the view classification dataset.

In the methods:

> We first trained a view classifier to provide a fine-grained characterization of echocardiogram views as defined by the American Society of Echocardiography guidelines[36]. We selected 77,426 echocardiogram videos, a subset of videos from the training set, for cardiac sonographers to label into 58 different view categories and trained an image-based ConvNextBase[37] view classification model on 224 x 224 images.

Presentation: The metric represented in Figure 2b is not specified. Including a description in the figure caption would clarify this. Furthermore, providing figures for data statistics would enhance understanding of the dataset's distribution and characteristics.

Reply: We clarify the metrics used in Figure 2B.

> ***(B)*** *Comparison of EchoPrime, EchoCLIP, and BioMedCLIP on a range of echocardiography interpretation tasks on cohorts from four different clinical centers. AUROC is shown for classification tasks and R2 score for regression tasks*

Figure 1A visualizes characteristics of the training dataset. Table 1 shows detailed statistics of each dataset used.

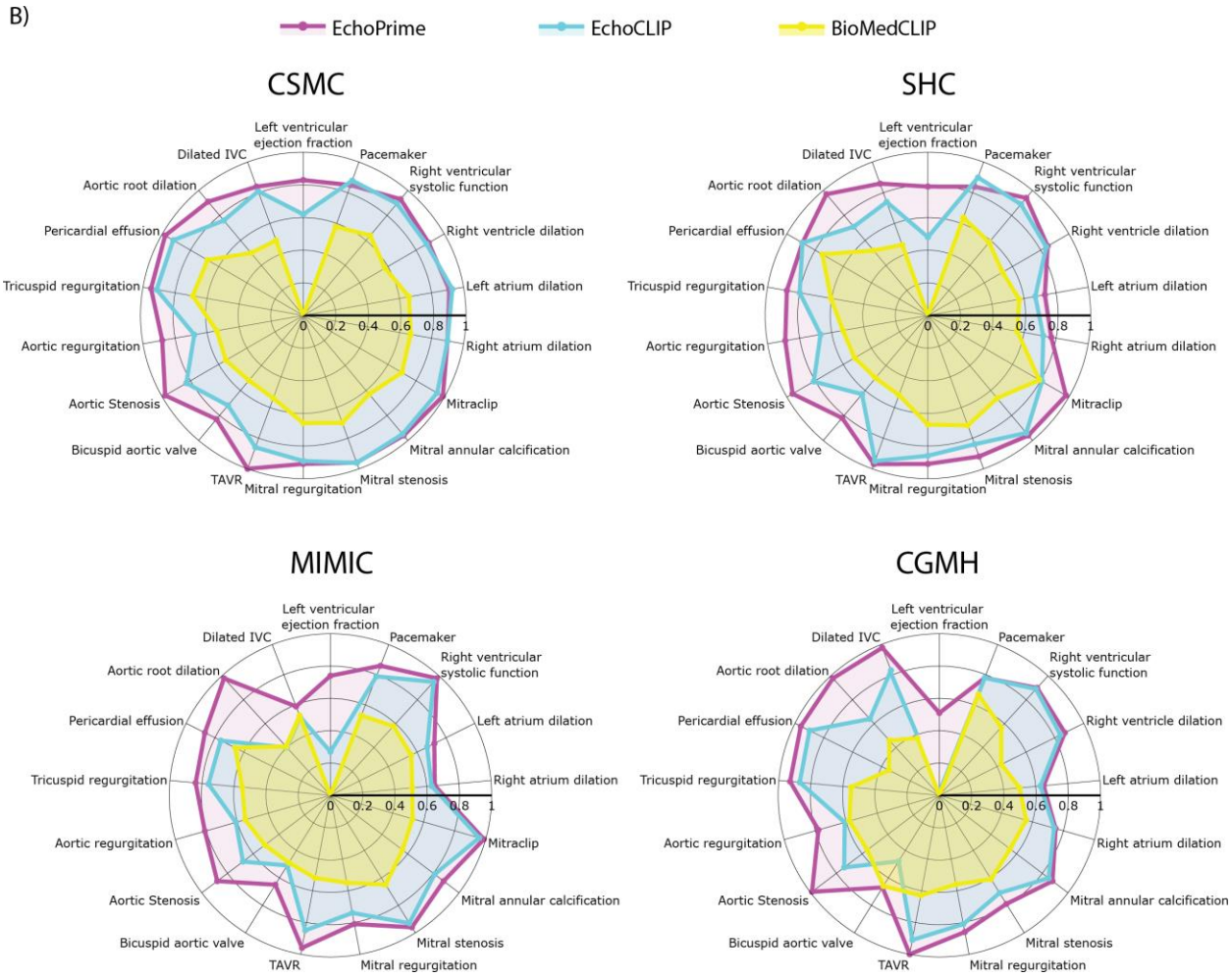


Figure 2**B**

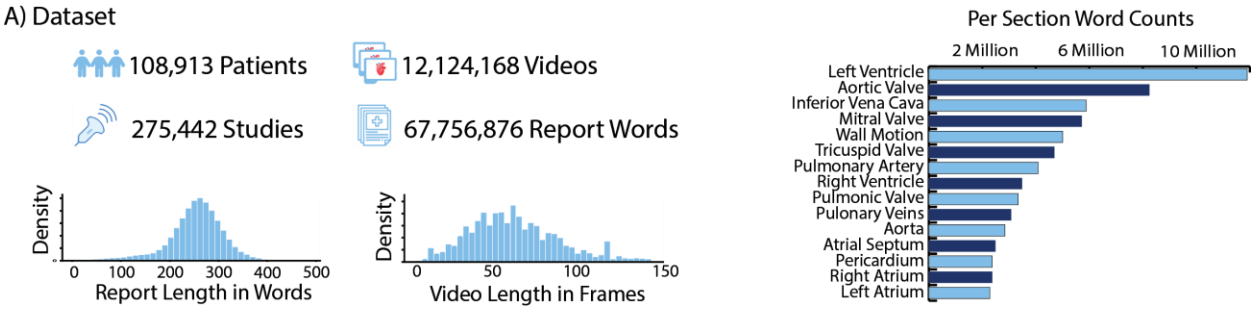


Figure 1A

Table 2: *Clinical characteristics of the study cohorts.*

	Cedars Sinai	***Stanford***	***MIMIC***	***CGMH***
Videos	*12,124,168*	*91,746*	*75,768*	*24,724*

Studies (N)	*275,442*	*1,792*	*2,431*	*1,188*
Patients	*108,913*	*1,779*	*1,767*	*1,188*
Age (mean (s.d.))	*66.31 (16.7)*	*60.42 (17.3)*	*67.67 (13.5)*	*65.87 (16.4)*
Female (%)	*117,324 (42.6)*	*843 (47.0)*	*909 (51.4)*	*583 (49.1)*
Race (%)				
Non-Hispanic White	*166,091 (60.3)*	*900 (50.2)*	*1,234 (69.8)*	*0 (0.0)*
Black	*35,624 (12.9)*	*76 (4.2)*	*343 (19.4)*	*0 (0.0)*
Hispanic	*27,194 (9.9)*	*241 (13.5)*	*68 (3.8)*	*0 (0.0)*
Asian	*20,299 (7.4)*	*301 (16.8)*	*45 (2.5)*	*1188 (100.00)*
Other	*19,832 (7.2)*	*161 (9.0)*	*42 (2.4)*	*0 (0.0)*
Unknown	*5,321 (1.9)*	*85 (4.7)*	*35 (2.0)*	*0 (0.0)*
Pacific Islander	*984 (0.4)*	*28 (1.6)*	*0 (0.0)*	*0 (0.0)*
Conditions (%)				
Hypertension	*112,919 (41.0)*	*1,055 (59.9)*	*1,510 (85.5)*	-
Heart failure	*93,875 (34.1)*	*746 (41.6)*	*996 (56.4)*	-
Atrial fibrillation	*60,448 (21.9)*	*555 (30.0)*	*510 (28.9)*	-
Chronic kidney disease	*50,302 (18.3)*	*580 (32.4)*	*755 (42.7)*	-
Diabetes mellitus	*46,656 (16.9)*	*168 (9.3)*	*486 (27.5)*	-
Pulmonary artery disease	*28,903 (10.5)*	*41 (2.3)*	*141 (8.0)*	-
Cerebrovascular accident	*18,442 (6.7)*	*104 (5.8)*	*160 (9.1)*	-
Myocardial infarction	*18,283 (6.6)*	*296 (16.5)*	*507 (28.7)*	-

D. Use of Statistics and Treatment of Uncertainties

The authors claim that EchoPrime matches or exceeds the performance of task-specific models, but they have not employed statistical methods to verify whether EchoPrime's results differ significantly from those of task-specific models. The lack of statistical measures in the paper is a

major limitation. Including confidence intervals for the results is strongly recommended to demonstrate the robustness of the performance comparisons.

Reply: Thank you for this important point. There was a substantial difference in the point estimate of performance for EchoPrime vs. benchmarks (BioMedCLIP,EchoCLIP), but we recognize the value of confidence intervals to evaluate true statistical difference. In this revision, we add confidence intervals to Supplementary Tables 2-6. Task Specific performance is reported from the cited literature and confidence intervals were included when they were reported in the original papers.

In **Methods "Evaluation and Benchmarking" section**:

> *To calculate the confidence intervals for performance metrics, we randomly sample the dataset with replacement, maintaining the same sample size as the original dataset, and compute the performance metrics. This process is repeated 10,000 times to generate a distribution of the metrics. Finally, determine the 2.5th and 97.5th percentiles to obtain the bootstrapped 95% confidence intervals.*

In Results "**Automated echocardiogram interpretation without supervised learning**" section:

> *EchoPrime outperforms or matches fully supervised models for single-task prediction[19–24] (Supplementary Table 6).*

Supplementary Table 2 *Performance of foundation models for biomedical imaging on Cedars-Sinai test set*

CEDARS-SINAI TEST SET					
	PubMedCLIP	*PMC-CLIP*	*BioMed CLIP*	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-0.32 (-0.40--0.25)*	*-0.38 (-0.46--0.31)*	*-2.60 (-3.00 - -2.20)*	*0.62 (0.59-0.65)*	***0.83 (0.81-0.85)***
LV ejection fraction (MAE)	*14.68 (14.30-15.06)*	*14.03 (13.58-14.49)*	*26.93 (26.42-27.44)*	*7.00 (6.74-7.26)*	***4.79 (4.60-4.97)***
Pacemaker	*0.50 (0.50-0.50)*	*0.51 (0.47-0.54)*	*0.58 (0.55-0.62)*	***0.88 (0.85-0.91)***	*0.85 (0.82-0.88)*
RV systolic function	*0.50 (0.49-0.51)*	*0.53 (0.48-0.58)*	*0.64 (0.59-0.69)*	*0.89 (0.86-0.92)*	***0.93 (0.91-0.95)***
RV dilation	*0.50 (0.45-0.56)*	*0.59 (0.50-0.67)*	*0.58 (0.52-0.64)*	*0.88 (0.83-0.92)*	***0.89 (0.83-0.95)***
LA dilation	*0.50 (0.50-0.50)*	*0.57 (0.48-0.65)*	*0.66 (0.58-0.74)*	***0.93 (0.89-0.96)***	*0.91 (0.86-0.96)*
RA dilation	*0.50 (0.50-0.50)*	*0.58 (0.47-0.68)*	*0.67 (0.59-0.75)*	*0.90 (0.82-0.97)*	***0.90 (0.82-0.97)***
Mitraclip	*0.50 (0.50-0.50)*	*0.49 (0.43-0.55)*	*0.70 (0.63-0.76)*	*0.95 (0.92-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.50 (0.46-0.55)*	*0.59 (0.54-0.64)*	*0.63 (0.58-0.68)*	*0.95 (0.93-0.96)*	***0.96 (0.95-0.97)***
Mitral stenosis	*0.50 (0.50-0.50)*	*0.55 (0.44-0.66)*	*0.70 (0.61-0.79)*	***0.96 (0.91-0.99)***	***0.96 (0.91-0.99)***

Mitral regurgitation	*0.50 (0.50-0.50)*	*0.59 (0.55-0.63)*	*0.66 (0.63-0.69)*	*0.89 (0.87-0.91)*	***0.92 (0.91-0.94)***
TAVR	*0.50 (0.50-0.51)*	*0.51 (0.46-0.55)*	*0.54 (0.50-0.57)*	*0.86 (0.83-0.89)*	***1.00 (0.99-1.00)***
Bicuspid AV	*0.50 (0.50-0.50)*	*0.51 (0.47-0.60)*	*0.52 (0.51-0.53)*	*0.72 (0.58-0.85)*	***0.83 (0.67-0.96)***
Aortic stenosis	*0.50 (0.50-0.52)*	*0.48 (0.42-0.55)*	*0.55 (0.53-0.57)*	*0.83 (0.79-0.87)*	***0.98 (0.96-0.99)***
Aortic regurgitation	*0.50 (0.50-0.50)*	*0.50 (0.41-0.58)*	*0.54 (0.53-0.55)*	*0.68 (0.60-0.76)*	***0.88 (0.83-0.93)***
Tricuspid regurgitation	*0.49 (0.45-0.53)*	*0.57 (0.51-0.63)*	*0.69 (0.64-0.74)*	*0.91(0.89-0.93)*	***0.95 (0.93-0.97)***
Pericardial effusion	*0.54 (0.45-0.62)*	*0.50 (0.42-0.57)*	*0.68 (0.59-0.77)*	*0.92 (0.89-0.95)*	***0.98 (0.95-0.99)***
Aortic root dilation	*0.50 (0.50-0.50)*	*0.44 (0.35-0.55)*	*0.50 (0.49-0.51)*	*0.76 (0.63-0.88)*	***0.91 (0.83-0.98)***
Dilated IVC	*0.50 (0.50-0.51)*	*0.53 (0.50-0.56)*	*0.49 (0.47-0.50)*	*0.81 (0.79-0.83)*	***0.84 (0.82-0.86)***
PA pressure (R^2)	*-0.59 (-0.74--0.46)*	*-0.83 (-1.07--0.62)*	*-0.19 (-0.24 - -0.14)*	*0.35 (0.29-0.41)*	***0.43 (0.39-0.47)***
PA pressure (MAE)	*11.27 (9.98-12.62)*	*13.65 (13.04-14.28*	*10.62 (10.12-11.12)*	*7.84 (7.45- 8.23)*	***7.30 (6.95-7.67)***

Supplementary Table 3 *Performance of foundation models for biomedical imaging on Stanford dataset*

STANFORD DATASET			
	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	-3.50 (-4.15- -3.03)	0.48 (0.40-0.55)	**0.79 (0.76-0.82)**
LV ejection fraction (MAE)	23.00 (22.68-23.79)	6.23 (5.89-6.44)	**4.14 (3.94-4.34)**
Pacemaker	0.64 (0.60-0.68)	**0.90 (0.88-0.93)**	0.84 (0.81-0.87)
RV systolic function	0.59 (0.02-0.92)	0.89 (0.73-1.00)	**0.94 (0.88-1.00)**
RV dilation	0.53 (0.49-0.58)	0.84 (0.79-0.89)	**0.85 (0.80-0.90)**
LA dilation	0.57 (0.55-0.60)	0.67 (0.65-0.70)	**0.73 (0.70-0.75)**
RA dilation	0.56 (0.52-0.59)	0.72 (0.69-0.75)	**0.77 (0.74-0.81)**
Mitraclip	0.79 (0.78-0.81)	0.81 (0.44-1.00)	**0.98 (0.94-1.00)**
Mitral annular calcification	0.66 (0.56-0.76)	0.94 (0.90-0.97)	**0.96 (0.94-0.98)**
Mitral stenosis	0.72 (0.63-0.81)	0.84 (0.75-0.92)	**0.92 (0.86-0.98)**
Mitral regurgitation	0.67 (0.57-0.71)	0.86 (0.82-0.90)	**0.91 (0.89-0.93)**
TAVR	0.52 (0.47-0.69)	0.95 (0.93-0.97)	**0.97 (0.90-1.00)**
Bicuspid AV	0.50 (0.47-0.54)	0.63 (0.53-0.73)	**0.82 (0.73-0.90)**
Aortic stenosis	0.52 (0.48-0.55)	0.81 (0.73-0.89)	**0.96 (0.92-0.99)**
Aortic regurgitation	0.53 (0.48-0.61)	0.67 (0.54-0.80)	**0.89 (0.80-0.97)**
Tricuspid regurgitation	0.61 (0.55-0.66)	0.80 (0.75-0.85)	**0.88 (0.86-0.91)**
Pericardial effusion	0.75 (0.60-0.66)	0.89 (0.78-0.98)	**0.89 (0.77-0.98)**
Aortic root dilation	0.52 (0.50-0.56)	0.71 (0.62-0.79)	**0.97 (0.94-0.99)**
Dilated IVC	0.46 (0.42-0.51)	0.75 (0.66-0.82)	**0.86 (0.81-0.91)**
PA pressure (R^2)	-0.25(-0.34- -0.17)	0.22 (0.13-0.30)	**0.36 (0.30-0.42)**
PA pressure (MAE)	10.92 (10.00-11.86)	8.80 (8.08-9.52)	**7.97 (7.35-8.66)**

Supplementary Table 4 *Performance of foundation models for biomedical imaging on MIMIC dataset from Boston*

MIMIC DATASET			
	BioMed CLIP	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-2.48 (-3.26--1.85)*	*0.27 (0.13-0.40)*	***0.74 (0.66-0.80)***
LV ejection fraction (MAE)	*26.96 (25.31-28.58)*	*10.37 (9.30-11.51)*	***6.21 (5.54-6.92)***
Pacemaker	*0.53 (0.39-0.68)*	*0.79 (0.66-0.90)*	***0.86 (0.80-0.92)***
RV systolic function	*0.58 (0.02-0.90)*	*0.95 (0.89-0.99)*	***0.98 (0.97-0.99)***
RV dilation	*-*	*-*	*-*
LA dilation	*0.56 (0.52-0.59)*	*0.67 (0.64-0.71)*	***0.72 (0.68-0.75)***
RA dilation	*0.51 (0.47-0.54)*	*0.63 (0.59-0.67)*	***0.65 (0.61-0.69)***
Mitraclip	*0.53 (0.29-0.99)*	*0.96 (0.86-1.00)*	***0.99 (0.99-1.00)***
Mitral annular calcification	*0.56 (0.51-0.60)*	*0.81 (0.77-0.85)*	***0.88 (0.85-0.91)***
Mitral stenosis	*0.65 (0.48-0.98)*	*0.93 (0.80-1.00)*	***0.96 (0.95-0.98)***
Mitral regurgitation	*0.55 (0.52-0.58)*	*0.74 (0.71-0.76)*	***0.81 (0.79-0.83)***
TAVR	*0.52 (0.44-0.60)*	*0.85 (0.79-0.90)*	***0.96 (0.94-0.98)***
Bicuspid AV	*0.49 (0.46-0.56)*	*0.51 (0.42-0.61)*	***0.65 (0.52-0.78)***
Aortic stenosis	*0.51 (0.48-0.54)*	*0.68 (0.64-0.73)*	***0.88 (0.84-0.91)***
Aortic regurgitation	*0.55 (0.49-0.61)*	*0.61 (0.54-0.67)*	***0.81 (0.75-0.86)***
Tricuspid regurgitation	*0.55 (0.52-0.58)*	*0.76 (0.73-0.78)*	***0.84 (0.82-0.86)***
Pericardial effusion	*0.66 (0.61-0.72)*	*0.76 (0.70-0.81)*	***0.87 (0.83-0.92)***
Aortic root dilation	*0.41 (0.40-0.42)*	*0.40 (0.39-0.40)*	***0.98 (0.96-0.99)***
Dilated IVC	*0.53 (0.44-0.63)*	*0.52 (0.40-0.64)*	***0.59 (0.49-0.69)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Supplementary Table 5 *Performance of foundation models for biomedical imaging on CGMH dataset from Taiwan*

CGMH DATASET			
	BioMed CLIP	*Echo CLIP*	*Echo Prime*
LV ejection fraction (R^2)	*-8.31 (-9.74--7.06)*	*0.02 (-0.13-0.15)*	***0.51 (0.44-0.57)***
LV ejection fraction (MAE)	*32.64 (31.89-33.38)*	*9.15 (8.75-9.54)*	***6.46 (6.19-6.74)***
Pacemaker	*0.67 (0.56-0.80)*	***0.78 (0.59-0.92)***	***0.78 (0.63-0.92)***
RV systolic function	*0.57 (0.41-0.72)*	*0.89 (0.84-0.94)*	***0.90 (0.79-0.98)***
RV dilation	*0.43 (0.30-0.61)*	*0.84 (0.65-0.97)*	***0.87 (0.69-0.97)***
LA dilation	*0.50 (0.47-0.53)*	*0.63 (0.60-0.65)*	***0.65 (0.62-0.67)***
RA dilation	*0.56 (0.47-0.64)*	*0.74 (0.65-0.83)*	***0.75 (0.67-0.84)***
Mitraclip	*-*	*-*	*-*
Mitral annular calcification	*0.55 (0.40-0.70)*	*0.85 (0.69-0.97)*	***0.88 (0.75-0.96)***
Mitral stenosis	*0.61 (0.48-0.79)*	*0.71 (0.51-0.92)*	***0.79 (0.59-0.98)***
Mitral regurgitation	*0.56 (0.50-0.63)*	*0.81 (0.76-0.86)*	***0.86 (0.82-0.89)***
TAVR	*0.63 (0.50-0.77)*	*0.91 (0.83-0.97)*	***1.00 (1.00-1.00)***
Bicuspid AV	*0.66 (0.43-1.00)*	*0.48 (0.35-0.77)*	***0.67 (0.45-1.00)***
Aortic stenosis	*0.56 (0.43-0.71)*	*0.74 (0.59-0.88)*	***0.99 (0.98-1.00)***
Aortic regurgitation	*0.58 (0.50-0.66)*	*0.61 (0.53-0.69)*	***0.78 (0.70-0.84)***
Tricuspid regurgitation	*0.55 (0.49-0.61)*	*0.87 (0.83-0.91)*	***0.93 (0.90-0.95)***
Pericardial effusion	*0.35 (0.16-0.65)*	*0.90 (0.69-1.00)*	***0.96 (0.91-1.00)***
Aortic root dilation	*0.46 (0.39-0.62)*	*0.64 (0.41-0.86)*	***0.98 (0.97-1.00)***
Dilated IVC	*0.38 (0.37-0.39)*	*0.83 (0.50-1.00)*	***0.98 (0.94-1.00)***
PA pressure (R^2)	*-*	*-*	*-*
PA pressure (MAE)	*-*	*-*	*-*

Supplementary Table 6 *EchoPrime vs Task Specific Models*

	ECHOPRIME INTERNAL	***TASK SPECIFIC INTERNAL***	***ECHOPRIME EXTERNAL***	***TASK SPECIFIC EXTERNAL***
LV ejection fraction (R^2)	***0.83 (0.81-0.85)***	*0.81*	***0.79 (0.76-0.82)***	*0.77*
LV ejection fraction (MAE)	*4.79 (4.60-4.97)*	***4.10***	***4.14 (3.94-4.34)***	*6.00*
Pacemaker	*0.85 (0.82-0.88)*	-	*0.84 (0.81-0.87)*	-
RV systolic function	*0.93 (0.91-0.95)*	-	*0.94 (0.88-1.00)*	-
RV dilation	*0.89 (0.83-0.95)*	-	*0.85 (0.80-0.90)*	-
LA dilation	***0.91 (0.86-0.96)***	*0.85*	*0.73 (0.70-0.75)*	-
RA dilation	*0.90 (0.82-0.97)*	-	*0.77 (0.74-0.81)*	-
Mitraclip	*0.99 (0.99-1.00)*	-	*0.98 (0.94-1.00)*	-
Mitral annular calcification	*0.96 (0.95-0.97)*	-	*0.96 (0.94-0.98)*	-
Mitral stenosis	*0.96 (0.91-0.99)*	-	*0.92 (0.86-0.98)*	-
Mitral regurgitation	***0.92 (0.91-0.94)***	***0.92 (0.90-0.93)***	*0.91 (0.89-0.93)*	***0.95 (0.92–0.97)***
TAVR	*1.00 (0.99-1.00)*	-	*0.97 (0.90-1.00)*	-
Bicuspid AV	*0.83 (0.67-0.96)*	-	*0.82 (0.73-0.90)*	-
Aortic stenosis	***0.98 (0.96-0.99)***	***0.98 (0.97- 0.99)***	***0.96 (0.92-0.99)***	*0.95 (0.94- 0.96)*
Aortic regurgitation	*0.88 (0.83-0.93)*	-	*0.89 (0.80-0.97)*	-
Tricuspid regurgitation	***0.95 (0.93-0.97)***	*0.93 (0.91- 0.94)*	*0.88 (0.86-0.91)*	***0.95 (0.94 - 0.96)***
Pericardial effusion	***0.98 (0.95-0.99)***	*0.94 (0.75- 0.99)*	*0.89 (0.77-0.98)*	-
Aortic root dilation	*0.91 (0.83-0.98)*	-	*0.97 (0.94-0.99)*	-
Dilated IVC	*0.84 (0.82-0.86)*	-	*0.86 (0.81-0.91)*	-
PA pressure (R^2)	*0.43 (0.39-0.47)*	-	*0.36 (0.30-0.42)*	-
PA pressure (MAE)	*7.30 (6.95-7.67)*	-	*7.97 (7.35-8.66)*	-

E. Conclusions: Robustness, Validity, Reliability

While the model demonstrates promising results, the robustness of these findings is limited by the absence of human evaluation, lack of prospective analysis, and exclusive reliance on quantitative metrics such as AUC. Additionally, the absence of statistical measures (e.g., confidence intervals) weakens the reliability of the conclusions. Inclusion of additional comparison methods such as PubMedCLIP [3], PMC-CLIP [4], and ROCO [5], or at least reference to them, would improve the comprehensiveness and validity of the evaluation.

Reply: *Thank you for your feedback. In the original manuscript, we chose benchmarks that have highest likelihood of the strongest performance In this revision, we add additional comparison models (PubMedCLIP and PMC-CLIP) and show that they perform worse than EchoPrime (**Supplementary Tables 2)**.*

Supplementary Table 2 Performance of foundation models for biomedical imaging on Cedars-Sinai test set

CEDARS-SINAI TEST SET					
	PubMedCLIP	PMC-CLIP	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	-0.32 (-0.40--0.25)	-0.38 (-0.46--0.31)	-2.60 (-3.00 - -2.20)	0.62 (0.59-0.65)	**0.83 (0.81-0.85)**
LV ejection fraction (MAE)	14.68 (14.30-15.06)	14.03 (13.58-14.49)	26.93 (26.42-27.44)	7.00 (6.74-7.26)	**4.79 (4.60-4.97)**
Pacemaker	0.50 (0.50-0.50)	0.51 (0.47-0.54)	0.58 (0.55-0.62)	**0.88 (0.85-0.91)**	0.85 (0.82-0.88)
RV systolic function	0.50 (0.49-0.51)	0.53 (0.48-0.58)	0.64 (0.59-0.69)	0.89 (0.86-0.92)	**0.93 (0.91-0.95)**
RV dilation	0.50 (0.45-0.56)	0.59 (0.50-0.67)	0.58 (0.52-0.64)	0.88 (0.83-0.92)	**0.89 (0.83-0.95)**
LA dilation	0.50 (0.50-0.50)	0.57 (0.48-0.65)	0.66 (0.58-0.74)	**0.93 (0.89-0.96)**	0.91 (0.86-0.96)
RA dilation	0.50 (0.50-0.50)	0.58 (0.47-0.68)	0.67 (0.59-0.75)	0.90 (0.82-0.97)	**0.90 (0.82-0.97)**
Mitraclip	0.50 (0.50-0.50)	0.49 (0.43-0.55)	0.70 (0.63-0.76)	0.95 (0.92-1.00)	**0.99 (0.99-1.00)**
Mitral annular calcification	0.50 (0.46-0.55)	0.59 (0.54-0.64)	0.63 (0.58-0.68)	0.95 (0.93-0.96)	**0.96 (0.95-0.97)**
Mitral stenosis	0.50 (0.50-0.50)	0.55 (0.44-0.66)	0.70 (0.61-0.79)	**0.96 (0.91-0.99)**	**0.96 (0.91-0.99)**
Mitral regurgitation	0.50 (0.50-0.50)	0.59 (0.55-0.63)	0.66 (0.63-0.69)	0.89 (0.87-0.91)	**0.92 (0.91-0.94)**
TAVR	0.50 (0.50-0.51)	0.51 (0.46-0.55)	0.54 (0.50-0.57)	0.86 (0.83-0.89)	**1.00 (0.99-1.00)**
Bicuspid AV	0.50 (0.50-0.50)	0.51 (0.47-0.60)	0.52 (0.51-0.53)	0.72 (0.58-0.85)	**0.83 (0.67-0.96)**
Aortic stenosis	0.50 (0.50-0.52)	0.48 (0.42-0.55)	0.55 (0.53-0.57)	0.83 (0.79-0.87)	**0.98 (0.96-0.99)**
Aortic regurgitation	0.50 (0.50-0.50)	0.50 (0.41-0.58)	0.54 (0.53-0.55)	0.68 (0.60-0.76)	**0.88 (0.83-0.93)**
Tricuspid regurgitation	0.49 (0.45-0.53)	0.57 (0.51-0.63)	0.69 (0.64-0.74)	0.91(0.89-0.93)	**0.95 (0.93-0.97)**
Pericardial effusion	0.54 (0.45-0.62)	0.50 (0.42-0.57)	0.68 (0.59-0.77)	0.92 (0.89-0.95)	**0.98 (0.95-0.99)**
Aortic root dilation	0.50 (0.50-0.50)	0.44 (0.35-0.55)	0.50 (0.49-0.51)	0.76 (0.63-0.88)	**0.91 (0.83-0.98)**
Dilated IVC	0.50 (0.50-0.51)	0.53 (0.50-0.56)	0.49 (0.47-0.50)	0.81 (0.79-0.83)	**0.84 (0.82-0.86)**

PA pressure (R^2)	-0.59 (-0.74--0.46)	-0.83 (-1.07--0.62)	-0.19 (-0.24 - -0.14)	0.35 (0.29-0.41)	**0.43 (0.39-0.47)**
PA pressure (MAE)	11.27 (9.98-12.62)	13.65 (13.04-14.28	10.62 (10.12-11.12)	7.84 (7.45- 8.23)	**7.30 (6.95-7.67)**

Referee #3 (Remarks on code availability):

It is not very clear about the checkpoint of the model in the Github repo.

Reply: Checkpoint of the model can be found under the releases in the GitHub repository under model_data/weights/echo_prime_encoder.pt. During this revision, we have clarified our README, added requirements.txt and Dockerfile to make it easier for a broader audience to access and work with our model.

Referee #1:

Dear Authors,

I appreciate your efforts in revising the manuscript and your detailed rebuttal. Here are my point-by-point comments that, to my view, are not fully addressed.

- In your rebuttal, you stated that comparing EchoPrime to an ECG-based AI model (AI-ECG) constitutes an "apples to oranges" comparison. My intention was not to equate ECG and echocardiography purely on technical grounds, but to emphasize that both might be used as screening or triage tools. ECG is often the first-line test in nearly every clinical setting, from outpatient clinics to the ICU, and an AI-based approach with high negative predictive value (NPV) can rule out certain pathologies. If EchoPrime's binary classification tasks can achieve similarly high NPV, I was suggesting it could be valuable in a similar triage context. My question remains: if EchoPrime is not intended for screening in a low pre-test probability population, then what distinct advantage does it offer in high pretest-probability (in-hospital) settings, especially if there is already a comprehensive in-person exam?

Thank you for this feedback and comparison with AI-ECG models like EchoNext which evaluates ECG for valvular and structural heart disease. We are intimately familiar with the model as we ran the model on CMSC ECGs and were a validation site. We have a three part response:

1. ECG is not the modality that diagnoses valvular pathologies, hence AI-ECG models are framed in the context of screening. Echocardiography is the modality that diagnoses cardiac valvular pathology, hence AI-Echo is presented an aid in the diagnostic process. In future work, this could be potentially fully 'autonomous', but that requires additional clinical testing and for now, believe there is value in EchoPrime providing preliminary assessment of echocardiograms similar to how sonographers do so in clinical practice.
2. Even recognizing this is not an apples-to-apples comparison, EchoPrime has stronger numerical performance than reported by EchoNext. In this revision, we report NPV across 5 validation sites in Supplementary Table 11. Specifically, for aortic regurgitation as pointed out by the reviewer, average NPV was 0.995 across sites, better than EchoNext.
3. We are currently planning a clinical trial of EchoPrime in comparison with sonographers, similar to our prior work in He et al. Nature 2025, that demonstrates of our proposed use case. The proposed clinical use-case is EchoPrime for comprehensive preliminary echocardiogram assessment in a population of all patients getting echocardiograms (not just a low risk screening population of patients receiving ECGs), and our trial will prospectively conduct blinded, randomized assessment by cardiologist to see whether AI vs. sonographer preliminary assessments are more likely to be changed in the final assessment.

In the results, we add:

In addition, EchoPrime achieves consistently high negative predictive values across all validation sites on core tasks (Supplementary Table 12), suggesting its utility in preliminary assessment of echocardiography images as well as triaging abnormal studies that require more urgent clinician review.

Supplementary Table 12 Negative predictive values of core tasks across external datasets.

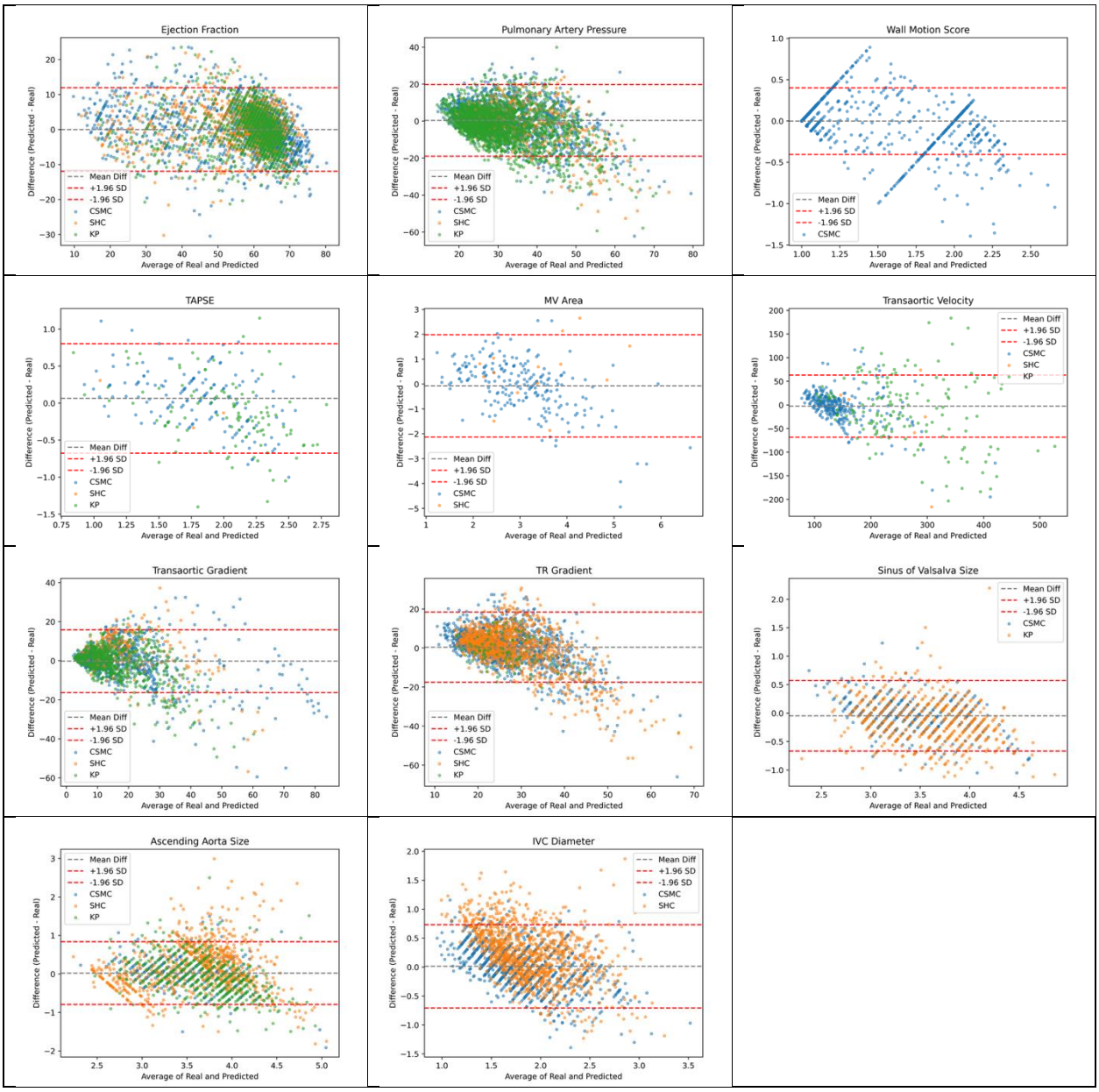
	CSMC	SHC	MIMIC	CGMH
Pacemaker	0.96	0.95	1.00	1.00
RV systolic function	1.00	1.00	1.00	1.00
RV dilation	1.00	0.99	1.00	1.00
LA dilation	1.00	0.81	0.95	0.76
RA dilation	1.00	0.95	0.95	0.99
Mitraclip	1.00	1.00	1.00	-
Mitral annular calcification	0.99	1.00	0.99	1.00
Mitral stenosis	1.00	1.00	1.00	1.00
Mitral regurgitation	0.98	1.00	0.92	0.99
TAVR	1.00	1.00	1.00	1.00
Bicuspid AV	1.00	0.99	0.99	1.00
Aortic stenosis	1.00	1.00	0.98	1.00
Aortic regurgitation	1.00	1.00	0.99	0.99
Tricuspid regurgitation	1.00	1.00	0.94	0.99
Pericardial effusion	1.00	1.00	0.99	1.00
Aortic root dilation	1.00	1.00	1.00	1.00
Dilated IVC	0.92	1.00	0.99	1.00

- I note that you have provided Mean Absolute Error (MAE) and R^2 for multiple regression tasks, but these two metrics, as they stand (which units in the MAEs? Clinical relevance?), do not fully capture the model's performance. How do we know how it behaves throughout the whole range? I suggest to address this question by also using Bland–Altman plots to analyse possible biases and agreement across the full range of measurements. For example, an MAE of ~7 mmHg for pulmonary pressures could be clinically significant or not, depending on the range of values and whether the errors skew low or high. Without Bland–Altman or explicit discussion of bias over the entire distribution, interpreting such differences remains difficult. I would also suggest to include the units to which the MAEs report to, otherwise in the metrics regarding to morphology, it could mean cms, or mm, or anything else, which makes it difficult to assess the magnitude of the findings. Additionally, you mention internal held-out sets (by the second supplementary table 7 that is provided in your rebuttal letter), yet it's unclear how consistent the regression performance is across other external cohorts. It would help to see these continuous metrics (beyond EF) over the other multiple sites explicitly. If you have performed regression analyses externally (e.g., for pulmonary pressures, aortic gradients, etc.), sharing those results—ideally with Bland–Altman plots or at least a more granular view—would greatly clarify how EchoPrime's continuous outputs hold up in diverse settings.

Thank you for this feedback and opportunity to clarify the performance of EchoPrime. In this revision:

1. We add Bland-Altman plots as a supplemental figure:

Supplementary Figure 5 Bland-Altman plots demonstrating EchoPrime's performance on regression tasks across three medical centers



2. We add units for MAE and break down performance by site. Sites that did not report certain metrics (for example wall motion score), were excluded.

Supplementary Table 8 EchoPrime's performance on regression tasks across three medical centers

Task	CSMC	KP	SHC
Ejection Fraction R^2	0.83 (0.82-0.85)	0.66 (0.62-0.71)	0.79 (0.76-0.82)
Ejection Fraction MAE (%)	4.64 (4.47-4.81)	4.51 (4.30-4.72)	4.14 (3.94-4.34)

Pulmonary Artery Pressure R^2	0.42 (0.36-0.46)	0.31 (0.27-0.35)	0.36 (0.30-0.42)
Pulmonary Artery Pressure MAE (mmHg)	7.30 (6.96-7.65)	7.39 (7.08-7.70)	7.97 (7.35-8.66)
Wall Motion Score R^2	0.77 (0.74-0.80)	-	-
Wall Motion Score MAE (unitless)	0.10 (0.10-0.11)	-	-
TAPSE R^2	0.39 (0.26-0.51)	0.28 (0.02-0.46)	0.34 (0.00–0.70)
TAPSE MAE (cm)	0.31 (0.27-0.35)	0.35 (0.30-0.40)	0.25 (0.13-0.33)
MV Area R^2	0.34 (0.19-0.44)	-	0.22 (0.16-0.28)
MV Area MAE (cm^2)	0.78 (0.69-0.88)	-	1.03 (0.62-1.44)
Transaortic velocity R^2	0.70 (0.59-0.78)	0.50 (0.34-0.61)	0.54 (0.23-0.75)
Transaortic velocity MAE (cm/s)	20.4 (18.1-22.8)	56.4 (49.7-63.7)	51.1 (30.6-75.2)
Transaortic gradient R^2	0.69 (0.64-0.74)	0.50 (0.45-0.55)	0.38 (0.33-0.53)
Transaortic gradient MAE (mmHg)	5.00 (4.62-5.40)	6.02 (5.56 – 6.51)	7.32 (6.63-8.00)
TR gradient R^2	0.36 (0.30-0.41)	0.26 (0.12-0.43)	0.28 (0.23-0.32)
TR gradient MAE (mmHg)	6.72 (6.39-7.07)	6.33 (5.41-7.26)	7.41 (6.93-7.85)
Sinus of Valsalva Size R^2	0.62 (0.55-0.67)	0.39 (0.30-0.47)	-
Sinus of Valsalva Size MAE (cm)	0.24 (0.22-0.26)	0.33 (0.31-0.35)	-
Ascending Aorta Size R^2	0.36 (0.23-0.48)	0.34 (0.23-0.44)	0.26 (0.20-0.32)
Ascending Aorta Size MAE (cm)	0.31 (0.28-0.34)	0.34 (0.32-0.36)	0.40 (0.37-0.43)
IVC Diameter R^2	0.44 (0.38-0.49)	-	0.18 (0.14-0.23)
IVC Diameter MAE (cm)	0.29 (0.28-0.30)	-	0.41 (0.39-0.43)

3. We have a separate model for quantitative measurements (EchoNet-Measurements) which would be the true comparison for quantitative measurements. While EchoPrime is capable of performing regression tasks, EchoPrime is a vision-language model and not a segmentation model like EchoNet-Measurements or Us2ai. Similar prior work have already been published which highlight the value of task-specific models particularly for quantitative measurements, and we are pleased that an alternative approach with a vision-language model has similar or better performance. We contextualize this in the discussion.

> A few limitations are worth considering. Since EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Many quantitative measurements are performed on still frames or images, therefore future versions should consider integrating the output of segmentation models[31,32] into the final report.

31. Sahashi, Y. *et al.* Artificial intelligence automation of echocardiographic measurements. Preprint at https://doi.org/10.1101/2025.03.18.25324215 (2025).
32. Tromp, J. *et al.* A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat. Commun.* **13**, 6776 (2022).

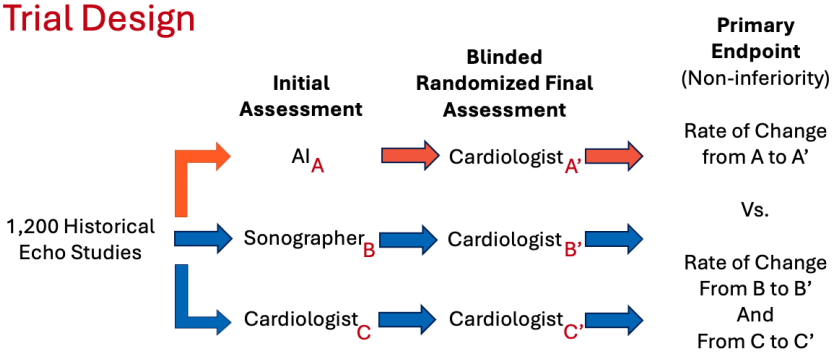
2b. In your rebuttal, you explain that the 55% EF is an "intermediate" output (closest single retrieved report), whereas 61% is the final aggregated EF from the top 50 retrieved reports. While that clarifies your pipeline, if both outputs are visible in the code or user-facing tool, users may be unsure which number is the "true" result. If users are to rely on EchoPrime, one would expect only the final aggregated EF (61%)—or some single definitive number—be displayed to avoid confusion. It would be helpful if you could clarify how such intermediate results are handled in any final production version, and whether the user-facing EF is always the aggregated figure.

Thank you for this feedback to improve the presentation in our research code. The shared code and weights on github are for the research and instructional purposes and is not meant as the clinical presentation. This is not the user interface for our upcoming clinical trial or we how we anticipate clinical use. In the clinical implementation, the end-user will only see the final aggregated EF. To clarify that final aggregated EF was used for tests in the paper, we added a comment in the code and clarify in the methods.

> "Feature logits used for EchoPrime evaluation in the paper" was added next to predict_metrics() function.

3. You argue EchoPrime is primarily for assisting cardiologists and sonographers by automatically generating preliminary reports—particularly in a higher pretest probability setting. My question is how you envision these automatically regressed measurements being used when some R^2 values are relatively low, the model's error margins remain unclear (units, clinical significance, how the model behaves throughout the whole range and in external cohorts). With this in mind, are users expected to trust or override these measurements? If they must be manually re-checked anyway, what exactly is the efficiency gain? Also, if the model excels at classification (binary predictions), it might have value in ruling out significant abnormalities in a broader population. Yet you state explicitly it is not for "clearing" low-risk patients. This leaves me uncertain how you see these two sets of outputs (classifications vs. continuous measurements) integrating into a real workflow.

Thank you for this opportunity to clarify how EchoPrime can be used in the clinical setting. We anticipate conducting a prospective clinical trial of EchoPrime shortly to directly assess affinity and acceptability to cardiologists that are reviewing the echocardiogram images. We have already acquired funding and institutional support to conduct this prospective evaluation, which will soon start. In this trial setting, we will randomize cardiologists to preliminary assessments of echocardiograms by EchoPrime vs. preliminary assessments performed by cardiologists and sonographers and track the rate of change as a second cardiologist conducts blinded randomized review.



We disagree that there is significant error in the models regression output compared the clinical variation or prior AI models. The EchoPrime MAE is comparable with prior work focused on measurements (Sahashi et al and Tromp et al. reporting MAE approximately 30-40% of mean measurement on many metrics) as well as has human variability (with MAE of approximately 20-30% on average, pmc.ncbi.nlm.nih.gov/articles/PMC11688108/). As recommended by the reviewer, in this revision we added more detail with regard to units and site specific performance with regression tasks. The absolute MAE is quite strong and comparable to clinician

performance, while R2 is dependent on the sample population (in populations without much variation in a particular measurement, there can be low R2 even with a low MAE).

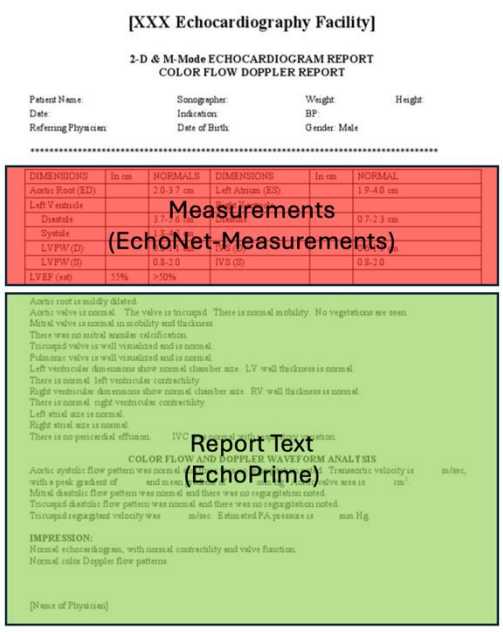
Supplementary Table 8

Task	CSMC	KP	SHC
Ejection Fraction R^2	0.83 (0.82-0.85)	0.66 (0.62-0.71)	0.79 (0.76-0.82)
Ejection Fraction MAE (%)	4.64 (4.47-4.81)	4.51 (4.30-4.72)	4.14 (3.94-4.34)
Pulmonary Artery Pressure R^2	0.42 (0.36-0.46)	0.31 (0.27-0.35)	0.36 (0.30-0.42)
Pulmonary Artery Pressure MAE (mmHg)	7.30 (6.96-7.65)	7.39 (7.08-7.70)	7.97 (7.35-8.66)
Wall Motion Score R^2	0.77 (0.74-0.80)	-	-
Wall Motion Score MAE (unitless)	0.10 (0.10-0.11)	-	-
TAPSE R^2	0.39 (0.26-0.51)	0.28 (0.02-0.46)	0.34 (0.00–0.70)
TAPSE MAE (cm)	0.31 (0.27-0.35)	0.35 (0.30-0.40)	0.25 (0.13-0.33)
MV Area R^2	0.34 (0.19-0.44)	-	0.22 (0.16-0.28)
MV Area MAE (cm^2)	0.78 (0.69-0.88)	-	1.03 (0.62-1.44)
Transaortic velocity R^2	0.70 (0.59-0.78)	0.50 (0.34-0.61)	0.54 (0.23-0.75)
Transaortic velocity MAE (cm/s)	20.4 (18.1-22.8)	56.4 (49.7-63.7)	51.1 (30.6-75.2)
Transaortic gradient R^2	0.69 (0.64-0.74)	0.50 (0.45-0.55)	0.38 (0.33-0.53)
Transaortic gradient MAE (mmHg)	5.00 (4.62-5.40)	6.02 (5.56 – 6.51)	7.32 (6.63-8.00)
TR gradient R^2	0.36 (0.30-0.41)	0.26 (0.12-0.43)	0.28 (0.23-0.32)
TR gradient MAE (mmHg)	6.72 (6.39-7.07)	6.33 (5.41-7.26)	7.41 (6.93-7.85)
Sinus of Valsalva Size R^2	0.62 (0.55-0.67)	0.39 (0.30-0.47)	-
Sinus of Valsalva Size MAE (cm)	0.24 (0.22-0.26)	0.33 (0.31-0.35)	-
Ascending Aorta Size R^2	0.36 (0.23-0.48)	0.34 (0.23-0.44)	0.26 (0.20-0.32)
Ascending Aorta Size MAE (cm)	0.31 (0.28-0.34)	0.34 (0.32-0.36)	0.40 (0.37-0.43)
IVC Diameter R^2	0.44 (0.38-0.49)	-	0.18 (0.14-0.23)
IVC Diameter MAE (cm)	0.29 (0.28-0.30)	-	0.42 (0.39-0.43)

Second, we note that echocardiographic measurement is a well-established area of study and we have our own separate model for quantitively measurements (EchoNet-Measurements, https://www.medrxiv.org/content/10.1101/2025.03.18.25324215v1) which would be the true head-to-head comparison for quantitative measurements. Prior work (Tromp et al. Nature Communications) and EchoNet-Measurements have highlighted the value of task-specific models for quantitative measurements. We add this clarification in the discussion.

> A few limitations are worth considering. Since EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Many quantitative measurements are performed on still frames or images, therefore future versions should consider integrating the output of segmentation models[31,32] into the final report.

31. Sahashi, Y. *et al.* Artificial intelligence automation of echocardiographic measurements. Preprint at https://doi.org/10.1101/2025.03.18.25324215 (2025).
32. Tromp, J. *et al.* A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat. Commun.* **13**, 6776 (2022).



Finally, we would like highlight that in the usual clinical workflow, measurements are roughly 1/4 to 1/3 of the report, while the vast majority of echo reporting is based on natural language that EchoPrime, a vision-language model generates. On the left hand side, we show a canonical example of an echo report found online (for which many other similar examples can be found), and highlight the area focused on measurements which would be most consistent with EchoNet-Measurements in red, and the natural language component of the report most consistent with EchoPrime in green. We use this example to highlight the complementary but independent nature of these two types of tasks, as well as the relative weighting of these two types of tasks.

In our anticipated clinical trial, we aim to study both EchoNet-Measurement and EchoPrime in combination as a fully automated preliminary report for echocardiograms. While we understand the importance of numerical measurements, this is a less novel task as noted by the reviewer in US2AI and EchoNet-Measurement. As the largest vision-language echo model trained on over 10-100x more data than other echo models, EchoPrime has unique characteristics, performance, and abilities that is distinct from the proposed measurement workflow.

4. I noted that US2AI can provide robust quantitative measurements—often recognized as the more challenging aspect of echo interpretation. In the rebuttal, you characterize US2AI as essentially segmentation-focused with minimal classification. However, many of these commercial tools produce key quantitative outputs (dimensions, strain, EF, volumes, etc.) that can then be thresholded or automatically labeled. If US2AI and similar solutions reliably measure the same metrics EchoPrime is attempting to regress, I believe it is still relevant to discuss whether EchoPrime's performance is near, below, or above that commercial standard—even if direct side-by-side comparisons are not feasible. Indeed, while US2AI may be a "black box," that does not preclude testing both EchoPrime and US2AI on the same de-identified dataset (with ground-truth labels) to compare final metrics such as EF, gradients, and volumes.

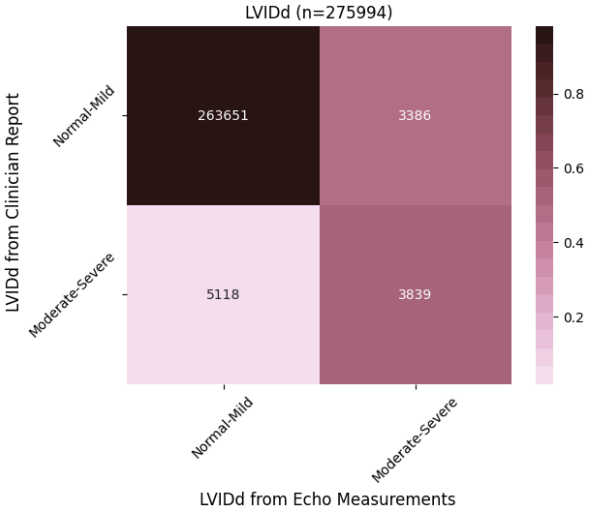
Furthermore, you highlight EchoPrime's ability to produce qualitative classification findings, whereas US2AI allegedly does not. However, in an actual echo lab, measured parameters (e.g., EF, volume, valve gradient) are regularly binned or thresholded to yield clinical categories ("mild," "moderate," etc.). Thus, the distinction between "quantitative segmentation" vs. "qualitative classification" can be blurred in real-world practice. In this sense, more clarity on how EchoPrime's performance on these key quantitative parameters compares to or complements existing commercial tools would help readers appreciate the added value.

Thank you for this opportunity to clarify how EchoPrime can be useful in the clinical setting as well as highlight limitations with measurements in the current clinical workflow. First, the key Achille's heel of echocardiography is variability, where common measurements have significant intra-observer variability in most quantitative measurements.[3–6] This variability in measurement is well recognized among clinical echocardiographers and why the American Society of Echocardiography recommends measuring things 3-5 times and averaging the measurements to create a point assessment.

3. Farsalinos, K. E. *et al.* Head-to-Head Comparison of Global Longitudinal Strain Measurements among Nine Different Vendors: The EACVI/ASE Inter-Vendor Comparison Study. *J. Am. Soc. Echocardiogr.* **28**, 1171–1181, e2 (2015).
4. Pillai, B., Salerno, M., Schnittger, I., Cheng, S. & Ouyang, D. Precision of echocardiographic measurements. *J. Am. Soc. Echocardiogr.* (2024) doi:10.1016/j.echo.2024.01.001.
5. Lang, R. M. *et al.* Recommendations for cardiac chamber quantification by echocardiography in adults: an update from the American Society of Echocardiography and the European Association of Cardiovascular Imaging. *Eur. Heart J. Cardiovasc. Imaging* **16**, 233–270 (2015).
6. Ouyang, D. *et al.* Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 580, 252–256 (2020).

This variability results in inconsistency between measurements and categorial assessments such that, in most echo labs and core labs, clinicians do not directly rely solely on measurements for categorical assessments. As an example, we demonstrate what happens when one compares the quantitative measurement of LVIDd and categorial assessment of left ventricular dilation across 275,994 historical echocardiogram studies at CSMC.

	Measurement Label	**Clinician Report**
Normal to Mild	Male && LVIDd≤6.3 Female && LVIDd≤5.6	"Normal left ventricular size by linear cavity dimension" "Mildly dilated left ventricular size by linear cavity dimension"
Moderate to Severe	Male && LVIDd>6.3 Female && LVIDd>5.6	"Moderately dilated left ventricular size by linear cavity dimension" "Severely dilated left ventricular size by linear cavity dimension"



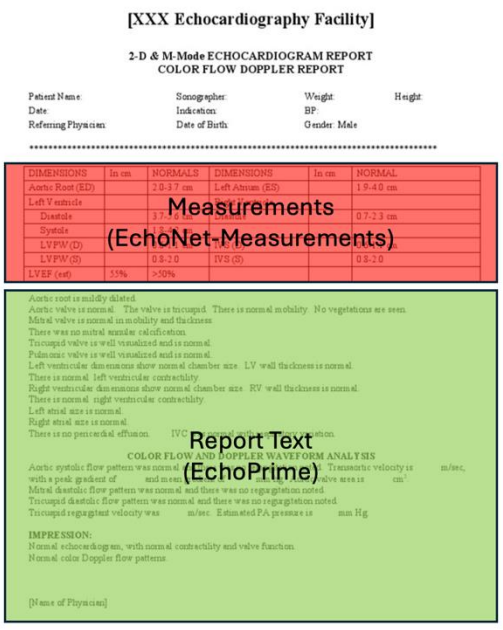
From the confusion matrix itself, one can see that there is significant overlap and inconsistency between categorial variables and measurements. IE, there are many patients with high numerical LVIDd without described LV dilation in the report and vis versa. The numerical measurement does not directly map to clinician categorial assessment, and as such, a purely measurement

driven finding workflow is limited. It is often said that echocardiography is ‘semi-quantitative’ and our prior work shows significant inconsistency when going solely from measurements. Rather, EchoPrime performs significantly better in matching the reports written by clinicians. It is an important point to note that we need to further study this area, and we do agree that prospective clinical trialing should be undertaken to understand performance and acceptability (which we will do shortly!). In this revision, in the discussion, we add:

> Although EchoPrime was evaluated at five geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in the real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime preliminary assessment. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows.

Furthermore, AI models for echocardiographic measurement is a well-established area of study and we have our own separate model for quantative measurements (EchoNet-Measurements, https://www.medrxiv.org/content/10.1101/2025.03.18.25324215v1) which would be the direct head-to-head comparison for quantitative measurements. Prior work and EchoNet-Measurements have highlighted the value of task-specific models for quantitative measurements, and in this revision, we discuss the differences and merits of each approach.

The development of EchoPrime does not displace or devalue EchoNet-Measurement of similar models, but there is a complementary nature to both approaches. In the usual clinical workflow, measurements are roughly 1/4 to 1/3 of the report, while the majority of report is based on natural language similar to what EchoPrime (a vision-language model) generates. On the right hand side, we show an example of an echo report found online and highlight the area focused on measurements and the area focused on natural language component most consistent with EchoPrime. We use this example to highlight the complementary but independent nature of these two types of tasks, as well as the relative weighting of these two types of tasks.



Finally, we would like to highlight that EchoPrime’s MAE on important quantitative tasks are better than the US2.ai model in their paper (Tromp et al. Nature Comm, see Supplementary Figure 5 for comparable metrics in TR velocity of 12cm MAE vs. 7cm MAE for EchoPrime). Many of the measurements described by the US2.ai model is not in the structured reporting, so cannot be head-to-head compared with EchoPrime, but can be compared with EchoNet-Measurements generally performs better.

In the discussion we add:

A few limitations are worth considering. Since EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Many quantitative measurements are performed on still frames or images, therefore future versions should consider integrating the output of segmentation models[31,32] into the final report.

31. Sahashi, Y. *et al.* Artificial intelligence automation of echocardiographic measurements. Preprint at https://doi.org/10.1101/2025.03.18.25324215 (2025).
32. Tromp, J. *et al.* A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat. Commun.* **13**, 6776 (2022).

5. In your rebuttal letter, you mention that EchoPrime "streamlines workflow" and "reduces inter-physician variability." These are potentially large claims. Yet I see no user-study data or operational metrics documenting time-savings, nor do I see statistics on variance reduction between cardiologists when aided by EchoPrime. In many echo labs, we already have standardized "template" reports with placeholders for key measurements. Merely producing "1000 possible statements" from a retrieval-based model may not, on its own, guarantee a more efficient or consistent process—particularly if measured values still require manual over-read or if some statements are clinically extraneous. My suggestion is, if you have any preliminary user feedback, pilot data, or even a small prospective trial that suggests the model truly improves turnaround time or consistency, that would be hugely beneficial to strengthen your claims of workflow improvement.

Thank you for this important opportunity to clarify EchoPrime's performance and compare with clinical performance. While we are currently starting our prospective clinical trial and not yet able to present prospective data in this revision, we add additional experiments to show that EchoPrime's precision with historical studies is higher than inter-physician variability in historical studies.

In this new experiment, we constructed a test cohort to evaluate variability by cardiologists and compare EchoPrime performance on this cohort. This clinical cohort comprises sequential echocardiography studies from the same patient for which a cardiologist declared there was no significant change between the two studies, and we compared the variability between the two studies in key quantitative and qualitative assessments. In this comparison, we show EchoPrime has similar or improved precision when evaluated on the same two studies, suggesting improved precision for EchoPrime compared to cardiologists:

Supplementary Table 13 EchoPrime vs Physician variability compared to Physician vs Physician variability. Reported values are balanced accuracies.

N=4607	ECHOPRIME VS PHYSICIAN	PHYSICIAN VS PHYSICIAN
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.73 (0.70-0.75)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	**4.59 (4.47-4.72)**
Pacemaker	0.78 (0.75-0.80)	**0.83 (0.80-0.85)**
RV systolic function	**0.88 (0.85-0.90)**	0.82 (0.79-0.86)
RV dilation	**0.87 (0.81-0.92)**	0.71 (0.66-0.76)
LA dilation	**0.86 (0.81-0.92)**	0.77 (0.72-0.82)

RA dilation	**0.90 (0.84-0.96)**	0.76 (0.69-0.84)
Mitraclip	**0.98 (0.95-0.99)**	0.96 (0.93-0.98)
Mitral annular calcification	**0.90 (0.87-0.92)**	0.78 (0.75-0.81)
Mitral stenosis	**0.92 (0.91-0.93)**	0.70 (0.59-0.82)
Mitral regurgitation	0.86 (0.83-0.88)	**0.87 (0.85-0.90)**
TAVR	0.98 (0.98-0.99)	**0.99 (0.98-0.99)**
Bicuspid AV	0.84 (0.70-0.96)	**0.87 (0.81-0.92)**
Aortic stenosis	**0.95 (0.93-0.96)**	0.83 (0.78-0.88)
Aortic regurgitation	**0.83 (0.78-0.88)**	0.75 (0.69-0.80)
Tricuspid regurgitation	**0.90 (0.88-0.93)**	0.86 (0.83-0.89)
Pericardial effusion	**0.91 (0.87-0.94)**	0.88 (0.81-0.94)
Aortic root dilation	**0.87 (0.79-0.94)**	0.77 (0.71-0.82)
PA pressure (R^2)	0.43 (0.39-0.47)	**0.50 (0.44-0.55)**
PA pressure (MAE)	7.30 (6.95-7.67)	**5.89 (5.70-6.09)**
Wall Motion Score (R^2)	**0.77 (0.74-0.80)**	0.76 (0.72-0.80)
Wall Motion Score (MAE)	0.10 (0.10-0.11)	**0.04 (0.04-0.05)**
MV Area (R^2)	0.34 (0.19-0.44)	**0.38 (0.13-0.59)**
MV Area (MAE)	0.78 (0.69-0.88)	**0.70 (0.60-0.81)**
Transaortic gradient (R^2)	0.69 (0.64-0.74)	**0.82 (0.78-0.86)**
Transaortic gradient (MAE)	5.00 (4.62-5.40)	**3.12 (2.97-3.27)**
TRgradient (R^2)	0.36 (0.30-0.41)	**0.50 (0.33-0.67)**
TRgradient (MAE)	6.72 (6.39-7.07)	**5.42 (5.21-5.66)**
Sinus of Valsalva Size (R^2)	**0.62 (0.55-0.67)**	0.42 (0.06-0.87)
Sinus of Valsalva Size (MAE)	0.24 (0.22-0.26)	**0.21 (0.16-0.30)**
Ascending Aorta Size (R^2)	0.36 (0.23-0.48)	**0.81 (0.77-0.85)**
Ascending Aorta Size (MAE)	0.31 (0.28-0.34)	**0.17 (0.16-0.19)**
IVC Diameter (R^2)	**0.44 (0.38-0.49)**	0.05 (-0.68-0.28)
IVC Diameter (MAE)	**0.29 (0.28-0.30)**	0.32 (0.31-0.34)

In the Methods:

> To assess inter-physician variability in echocardiography interpretations, we evaluated clinician interpretations and EchoPrime interpretations on a cohort of patients undergoing consecutive echocardiogram studies for which a cardiologist thought there was no significant difference between the two studies. This dataset comprised of 4607 paired studies for which we evaluated the degree of variation between the first and second studies of the same patient.

In the Results:

> To contextualize EchoPrime's performance and reproducibility of assessments, we assessed cardiologist inter-observer variability in a series of paired echocardiogram studies from the same patient for which cardiologists noted no significant change. We found that EchoPrime's agreement with cardiologists is comparable to the agreement between two cardiologists as EchoPrime had an average balanced accuracy of 0.89, while the cardiologists had a balanced accuracy of 0.82 (Supplementary Table 13).

In the Discussion:

> Although EchoPrime was evaluated at five geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in the real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime preliminary assessment. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration

into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate many steps in echocardiographic imaging workflow.

6. While I appreciate that the retrieval-based approach can generate a wide array of textual phrases, I believe most clinicians are far more concerned with core structured findings (e.g., EF, diastolic function, valve gradients, presence/absence of regurgitation). Without robust external validation of these measured or classified parameters—or a demonstrated added value of these 1000 phrases—the mere variety of textual statements is unlikely to impress. Many clinical settings already use semi-structured "auto-text" with blank fields. Thus, clarifying the true novelty in actual practice would strengthen the paper.

Thank you for this opportunity to improve the presentation of our results. We understand the difficulty of showing full performance of over 1000 unique text phrases, and as such, our main results (Table 2) sought to highlight a few key structured findings that comprise the most commonly seen pathologies as well as all cardiac structures. We wanted to emphasize that this is just a subset of what EchoPrime can do, because in the first revision, reviewers thought EchoPrime can *only* do these tasks.

Table 2

	CSMC	SHC	MIMIC	CGMH	KP
LV ejection fraction (R^2)	0.83 (0.81-0.85)	0.79 (0.76-0.82)	0.74 (0.66-0.80)	0.51 (0.44-0.57)	0.66 (0.62-0.71)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	4.14 (3.94-4.34)	6.21 (5.54-6.92)	6.46 (6.19-6.74)	4.51 (4.30-4.72)
Pacemaker	0.85 (0.82-0.88)	0.84 (0.81-0.87)	0.86 (0.80-0.92)	0.78 (0.63-0.92)	0.84 (0.81-0.87)
RV systolic function	0.93 (0.91-0.95)	0.94 (0.88-1.00)	0.98 (0.97-0.99)	0.90 (0.79-0.98)	0.92 (0.88-0.95)
RV dilation	0.89 (0.83-0.95)	0.85 (0.80-0.90)	-	0.87 (0.69-0.97)	0.78 (0.73-0.83)
LA dilation	0.91 (0.86-0.96)	0.73 (0.70-0.75)	0.72 (0.68-0.75)	0.65 (0.62-0.67)	0.77 (0.75-0.79)
RA dilation	0.90 (0.82-0.97)	0.77 (0.74-0.81)	0.65 (0.61-0.69)	0.75 (0.67-0.84)	0.82 (0.79-0.84)
Mitraclip	0.99 (0.99-1.00)	0.98 (0.94-1.00)	0.99 (0.99-1.00)	-	1.00 (1.00-1.00)
Mitral annular calcification	0.96 (0.95-0.97)	0.96 (0.94-0.98)	0.88 (0.85-0.91)	0.88 (0.75-0.96)	0.89 (0.86-0.91)
Mitral stenosis	0.96 (0.91-0.99)	0.92 (0.86-0.98)	0.96 (0.95-0.98)	0.79 (0.59-0.98)	0.88 (0.78-0.95)
Mitral regurgitation	0.92 (0.91-0.94)	0.91 (0.89-0.93)	0.81 (0.79-0.83)	0.86 (0.82-0.89)	0.89 (0.86-0.91)
TAVR	1.00 (0.99-1.00)	0.97 (0.90-1.00)	0.96 (0.94-0.98)	1.00 (1.00-1.00)	0.99 (0.98-0.99)
Bicuspid AV	0.83 (0.67-0.96)	0.82 (0.73-0.90)	0.65 (0.52-0.78)	0.67 (0.45-1.00)	0.80 (0.72-0.88)
Aortic stenosis	0.98 (0.96-0.99)	0.96 (0.92-0.99)	0.88 (0.84-0.91)	0.99 (0.98-1.00)	0.97 (0.96-0.98)
Aortic regurgitation	0.88 (0.83-0.93)	0.89 (0.80-0.97)	0.81 (0.75-0.86)	0.78 (0.70-0.84)	0.87 (0.83-0.92)
Tricuspid regurgitation	0.95 (0.93-0.97)	0.88 (0.86-0.91)	0.84 (0.82-0.86)	0.93 (0.90-0.95)	0.91 (0.89-0.93)
Pericardial effusion	0.98 (0.95-0.99)	0.89 (0.77-0.98)	0.87 (0.83-0.92)	0.96 (0.91-1.00)	1.00 (0.99-1.00)
Aortic root dilation	0.91 (0.83-0.98)	0.97 (0.94-0.99)	0.98 (0.96-0.99)	0.98 (0.97-1.00)	0.81 (0.78-0.85)
Dilated IVC	0.84 (0.82-0.86)	0.86 (0.81-0.91)	0.59 (0.49-0.69)	0.98 (0.94-1.00)	0.80 (0.77-0.82)
Diastolic dysfunction	0.91 (0.88-0.93)	0.86 (0.81-0.90)	0.93 (0.86-0.98)	-	0.84 (0.81-0.87)
PA pressure (R^2)	0.43 (0.39-0.47)	0.36 (0.30-0.42)	-	-	0.31 (0.27-0.35)
PA pressure (MAE)	7.30 (6.95-7.67)	7.97 (7.35-8.66)	-	-	7.39 (7.08-7.70)

To further support the generalizability of our findings, we have now included an additional 5th validation site—Kaiser Permanente— and demonstrate consistent performance across diverse healthcare systems, including now an integrated delivery system. (**Table 2**, **Supplementary Table 6**, **Supplementary Table 8** and **Figure 2**)

In the Results:

> We evaluated EchoPrime on a diverse set of benchmarks for cardiac structure and function across datasets from five different institutions (**Error! Reference source not found.**). On echocardiographic data from all five geographically diverse hospitals, EchoPrime outperformed previously developed foundation models for biomedical imaging [17,19–21] (Figure 2, Supplementary Tables 2-6), and without additional fine-tuning or training, EchoPrime outperforms or matches fully supervised models for single-task prediction[22–27] (Supplementary Table 7). EchoPrime had a mean AUC of 0.92 on CSMC, 0.89 on Stanford Healthcare (SHC), 0.85 on Beth Israel Deaconess Medical Center (MIMIC), 0.86 on Chang Gung Memoral Hospital (CGMH), and 0.88 on Kaiser Permanente (KP) healthcare system across 17 classification tasks and 11 regression tasks (Table 2, Supplementary Table 8).

Supplementary Table 6 Performance of foundation models for biomedical imaging on Kaiser Permanente dataset

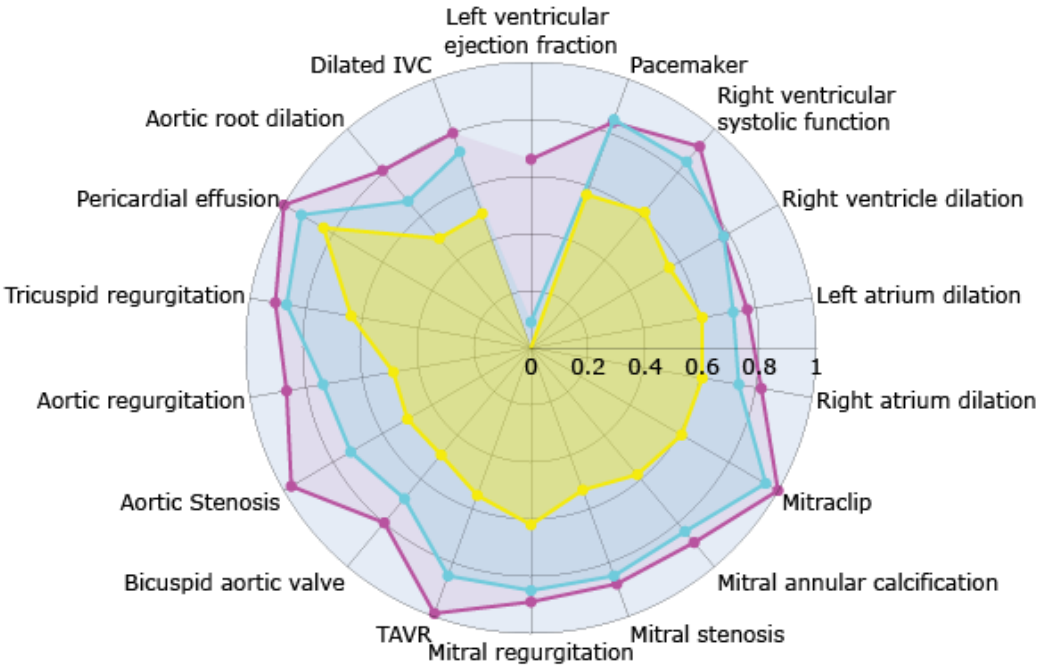
KP DATASET			
	BioMed CLIP	Echo CLIP	Echo Prime
LV ejection fraction (R^2)	*-8.31 (-9.74--7.09)*	*0.09 (-0.04-0.21)*	*0.66 (0.62-0.71)*
LV ejection fraction (MAE)	*29.67 (29.09-30.25)*	*7.26 (6.89-7.64)*	*4.51 (4.30-4.72)*
Pacemaker	*0.57 (0.52-0.61)*	*0.85 (0.81-0.88)*	*0.84 (0.81-0.87)*
RV systolic function	*0.62 (0.55-0.68)*	*0.85 (0.80-0.90)*	*0.92 (0.88-0.95)*
RV dilation	*0.56 (0.51-0.62)*	*0.78 (0.72-0.84)*	*0.78 (0.73-0.83)*
LA dilation	*0.61 (0.59-0.63)*	*0.72 (0.70-0.74)*	*0.77 (0.75-0.79)*
RA dilation	*0.61 (0.57-0.64)*	*0.74 (0.71-0.77)*	*0.82 (0.79-0.84)*
Mitraclip	*0.61 (0.47-0.74)*	*0.95 (0.84-1.00)*	*1.00 (1.00-1.00)*
Mitral annular calcification	*0.58 (0.55-0.62)*	*0.84 (0.81-0.87)*	*0.89 (0.86-0.91)*
Mitral stenosis	*0.53 (0.49-0.61)*	*0.85 (0.74-0.95)*	*0.88 (0.78-0.95)*
Mitral regurgitation	*0.62 (0.58-0.66)*	*0.85 (0.81-0.88)*	*0.89 (0.86-0.91)*
TAVR	*0.55 (0.51-0.60)*	*0.85 (0.80-0.89)*	*0.99 (0.98-0.99)*
Bicuspid AV	*0.49 (0.49-0.49)*	*0.69 (0.60-0.77)*	*0.80 (0.72-0.88)*
Aortic stenosis	*0.50 (0.49-0.51)*	*0.73 (0.69-0.76)*	*0.97 (0.96-0.98)*
Aortic regurgitation	*0.49 (0.48-0.51)*	*0.74 (0.68-0.80)*	*0.87 (0.83-0.92)*
Tricuspid regurgitation	*0.64 (0.60-0.68)*	*0.87 (0.83-0.90)*	*0.91 (0.89-0.93)*
Pericardial effusion	*0.84 (0.69-0.95)*	*0.93 (0.78-1.00)*	*1.00 (0.99-1.00)*
Aortic root dilation	*0.50 (0.48-0.52)*	*0.67 (0.63-0.72)*	*0.81 (0.78-0.85)*
Dilated IVC	*0.50 (0.49-0.52)*	*0.73 (0.70-0.76)*	*0.80 (0.77-0.82)*
PA pressure (R^2)	*-0.20 (-0.25--0.14)*	*0.17 (0.12-0.21)*	*0.31 (0.27-0.35)*
PA pressure (MAE)	*9.46 (9.05-9.89)*	*8.26 (7.92-8.60)*	*7.39 (7.08-7.70)*

Supplementary Table 8

Task	CSMC	KP	SHC
Ejection Fraction R^2	0.83 (0.82-0.85)	0.66 (0.62-0.71)	0.79 (0.76-0.82)
Ejection Fraction MAE (%)	4.64 (4.47-4.81)	4.51 (4.30-4.72)	4.14 (3.94-4.34)
Pulmonary Artery Pressure R^2	0.42 (0.36-0.46)	0.31 (0.27-0.35)	0.36 (0.30-0.42)
Pulmonary Artery Pressure MAE (mmHg)	7.30 (6.96-7.65)	7.39 (7.08-7.70)	7.97 (7.35-8.66)
Wall Motion Score R^2	0.77 (0.74-0.80)	-	-
Wall Motion Score MAE (unitless)	0.10 (0.10-0.11)	-	-
TAPSE R^2	0.39 (0.26-0.51)	0.28 (0.02-0.46)	0.34 (0.00–0.70)
TAPSE MAE (cm)	0.31 (0.27-0.35)	0.35 (0.30-0.40)	0.25 (0.13-0.33)
MV Area R^2	0.34 (0.19-0.44)	-	0.22 (0.16-0.28)
MV Area MAE (cm^2)	0.78 (0.69-0.88)	-	1.03 (0.62-1.44)
Transaortic velocity R^2	0.70 (0.59-0.78)	0.50 (0.34-0.61)	0.54 (0.23-0.75)

Transaortic velocity MAE (cm/s)	20.4 (18.1-22.8)	56.4 (49.7-63.7)	51.1 (30.6-75.2)
Transaortic gradient R^2	0.69 (0.64-0.74)	0.50 (0.45-0.55)	0.38 (0.33-0.53)
Transaortic gradient MAE (mmHg)	5.00 (4.62-5.40)	6.02 (5.56 – 6.51)	7.32 (6.63-8.00)
TR gradient R^2	0.36 (0.30-0.41)	0.26 (0.12-0.43)	0.28 (0.23-0.32)
TR gradient MAE (mmHg)	6.72 (6.39-7.07)	6.33 (5.41-7.26)	7.41 (6.93-7.85)
Sinus of Valsalva Size R^2	0.62 (0.55-0.67)	0.39 (0.30-0.47)	-
Sinus of Valsalva Size MAE (cm)	0.24 (0.22-0.26)	0.33 (0.31-0.35)	-
Ascending Aorta Size R^2	0.36 (0.23-0.48)	0.34 (0.23-0.44)	0.26 (0.20-0.32)
Ascending Aorta Size MAE (cm)	0.31 (0.28-0.34)	0.34 (0.32-0.36)	0.40 (0.37-0.43)
IVC Diameter R^2	0.44 (0.38-0.49)	-	0.18 (0.14-0.23)
IVC Diameter MAE (cm)	0.29 (0.28-0.30)	-	0.43 (0.39-0.43)

Comparison of EchoPrime vs EchoCLIP and BiomedCLIP on KP dataset is Added to Figure 2:



7. In your rebuttal to Reviewer #3, you provided a sample echocardiography report in which EchoPrime outputs statements describing "Mild diastolic dysfunction" (e.g., mentioning E-to-A reversal, prolonged deceleration time). However, it is not clear whether diastolic function was explicitly benchmarked among the model's validated outputs. If not, how should the user interpret these automatically generated statements on diastolic dysfunction? Do you have internal metrics or pilot data indicating that EchoPrime can reliably detect and characterize diastolic abnormalities, or are these statements purely illustrative at this stage? By clarifying this, readers can better assess the reliability of EchoPrime's commentary on diastolic parameters and understand whether these reported findings have been validated in a rigorous clinical or quantitative manner.

Thank you for this opportunity further validate EchoPrimes performance. In this revision, we add additional benchmarks for diastolic function to the main results (Table 2). Across all sites that reliably assess diastolic function, EchoPrime was able to show strong performance in comparison with clinical reporting. The international site, CGMH from Taiwan, did not routinely report diastology, and was excluded from this experiment.

Table 2

	CSMC	SHC	MIMIC	CGMH	KP
LV ejection fraction (R^2)	0.83 (0.81-0.85)	0.79 (0.76-0.82)	0.74 (0.66-0.80)	0.51 (0.44-0.57)	0.66 (0.62-0.71)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	4.14 (3.94-4.34)	6.21 (5.54-6.92)	6.46 (6.19-6.74)	4.51 (4.30-4.72)
Pacemaker	0.85 (0.82-0.88)	0.84 (0.81-0.87)	0.86 (0.80-0.92)	0.78 (0.63-0.92)	0.84 (0.81-0.87)
RV systolic function	0.93 (0.91-0.95)	0.94 (0.88-1.00)	0.98 (0.97-0.99)	0.90 (0.79-0.98)	0.92 (0.88-0.95)
RV dilation	0.89 (0.83-0.95)	0.85 (0.80-0.90)	-	0.87 (0.69-0.97)	0.78 (0.73-0.83)
LA dilation	0.91 (0.86-0.96)	0.73 (0.70-0.75)	0.72 (0.68-0.75)	0.65 (0.62-0.67)	0.77 (0.75-0.79)
RA dilation	0.90 (0.82-0.97)	0.77 (0.74-0.81)	0.65 (0.61-0.69)	0.75 (0.67-0.84)	0.82 (0.79-0.84)
Mitraclip	0.99 (0.99-1.00)	0.98 (0.94-1.00)	0.99 (0.99-1.00)	-	1.00 (1.00-1.00)
Mitral annular calcification	0.96 (0.95-0.97)	0.96 (0.94-0.98)	0.88 (0.85-0.91)	0.88 (0.75-0.96)	0.89 (0.86-0.91)
Mitral stenosis	0.96 (0.91-0.99)	0.92 (0.86-0.98)	0.96 (0.95-0.98)	0.79 (0.59-0.98)	0.88 (0.78-0.95)
Mitral regurgitation	0.92 (0.91-0.94)	0.91 (0.89-0.93)	0.81 (0.79-0.83)	0.86 (0.82-0.89)	0.89 (0.86-0.91)
TAVR	1.00 (0.99-1.00)	0.97 (0.90-1.00)	0.96 (0.94-0.98)	1.00 (1.00-1.00)	0.99 (0.98-0.99)
Bicuspid AV	0.83 (0.67-0.96)	0.82 (0.73-0.90)	0.65 (0.52-0.78)	0.67 (0.45-1.00)	0.80 (0.72-0.88)
Aortic stenosis	0.98 (0.96-0.99)	0.96 (0.92-0.99)	0.88 (0.84-0.91)	0.99 (0.98-1.00)	0.97 (0.96-0.98)
Aortic regurgitation	0.88 (0.83-0.93)	0.89 (0.80-0.97)	0.81 (0.75-0.86)	0.78 (0.70-0.84)	0.87 (0.83-0.92)
Tricuspid regurgitation	0.95 (0.93-0.97)	0.88 (0.86-0.91)	0.84 (0.82-0.86)	0.93 (0.90-0.95)	0.91 (0.89-0.93)
Pericardial effusion	0.98 (0.95-0.99)	0.89 (0.77-0.98)	0.87 (0.83-0.92)	0.96 (0.91-1.00)	1.00 (0.99-1.00)
Aortic root dilation	0.91 (0.83-0.98)	0.97 (0.94-0.99)	0.98 (0.96-0.99)	0.98 (0.97-1.00)	0.81 (0.78-0.85)
Dilated IVC	0.84 (0.82-0.86)	0.86 (0.81-0.91)	0.59 (0.49-0.69)	0.98 (0.94-1.00)	0.80 (0.77-0.82)
Diastolic dysfunction	0.91 (0.88-0.93)	0.86 (0.81-0.90)	0.93 (0.86-0.98)	-	0.84 (0.81-0.87)
PA pressure (R^2)	0.43 (0.39-0.47)	0.36 (0.30-0.42)	-	-	0.31 (0.27-0.35)
PA pressure (MAE)	7.30 (6.95-7.67)	7.97 (7.35-8.66)	-	-	7.39 (7.08-7.70)

I want to emphasize that I do see potential in EchoPrime, particularly given the scale of training data and the comprehensiveness of tasks. My comments aim to ensure clarity and clinical credibility. Thank you for the opportunity to review and for your time in addressing these concerns.

Thank you for this opportunity to improve upon of manuscript as well as push us to clarify our presentation. We appreciate the comments and think these suggestions have significantly improved our manuscript and results presentation.

Referee #2:

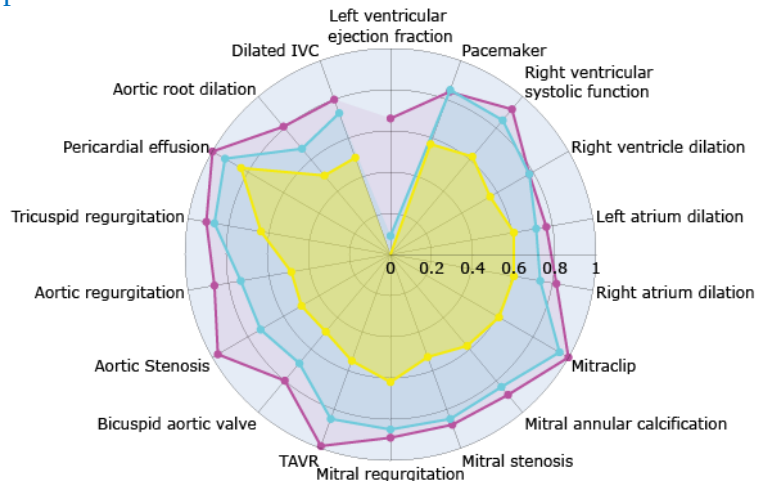
Congratulations for a significantly improved manuscript. You have expanded your testing cohort to include two more clinical sites (including one from Taiwan with somewhat different echo standards than American academic labs). You have added several new metrics for both categorical and parametric variables, which shows both the versatility and limitation of your approach.

Thank you for this opportunity to improve the presentation of our results. To further support the generalizability of our findings, we have now included an additional 5th validation site—Kaiser Permanente— and demonstrate consistent performance across diverse healthcare systems, including now an integrated delivery system.

Table 2

	CSMC	SHC	MIMIC	CGMH	KP
LV ejection fraction (R^2)	0.83 (0.81-0.85)	0.79 (0.76-0.82)	0.74 (0.66-0.80)	0.51 (0.44-0.57)	0.66 (0.62-0.71)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	4.14 (3.94-4.34)	6.21 (5.54-6.92)	6.46 (6.19-6.74)	4.51 (4.30-4.72)
Pacemaker	0.85 (0.82-0.88)	0.84 (0.81-0.87)	0.86 (0.80-0.92)	0.78 (0.63-0.92)	0.84 (0.81-0.87)
RV systolic function	0.93 (0.91-0.95)	0.94 (0.88-1.00)	0.98 (0.97-0.99)	0.90 (0.79-0.98)	0.92 (0.88-0.95)
RV dilation	0.89 (0.83-0.95)	0.85 (0.80-0.90)	-	0.87 (0.69-0.97)	0.78 (0.73-0.83)
LA dilation	0.91 (0.86-0.96)	0.73 (0.70-0.75)	0.72 (0.68-0.75)	0.65 (0.62-0.67)	0.77 (0.75-0.79)
RA dilation	0.90 (0.82-0.97)	0.77 (0.74-0.81)	0.65 (0.61-0.69)	0.75 (0.67-0.84)	0.82 (0.79-0.84)
Mitraclip	0.99 (0.99-1.00)	0.98 (0.94-1.00)	0.99 (0.99-1.00)	-	1.00 (1.00-1.00)
Mitral annular calcification	0.96 (0.95-0.97)	0.96 (0.94-0.98)	0.88 (0.85-0.91)	0.88 (0.75-0.96)	0.89 (0.86-0.91)
Mitral stenosis	0.96 (0.91-0.99)	0.92 (0.86-0.98)	0.96 (0.95-0.98)	0.79 (0.59-0.98)	0.88 (0.78-0.95)
Mitral regurgitation	0.92 (0.91-0.94)	0.91 (0.89-0.93)	0.81 (0.79-0.83)	0.86 (0.82-0.89)	0.89 (0.86-0.91)
TAVR	1.00 (0.99-1.00)	0.97 (0.90-1.00)	0.96 (0.94-0.98)	1.00 (1.00-1.00)	0.99 (0.98-0.99)
Bicuspid AV	0.83 (0.67-0.96)	0.82 (0.73-0.90)	0.65 (0.52-0.78)	0.67 (0.45-1.00)	0.80 (0.72-0.88)
Aortic stenosis	0.98 (0.96-0.99)	0.96 (0.92-0.99)	0.88 (0.84-0.91)	0.99 (0.98-1.00)	0.97 (0.96-0.98)
Aortic regurgitation	0.88 (0.83-0.93)	0.89 (0.80-0.97)	0.81 (0.75-0.86)	0.78 (0.70-0.84)	0.87 (0.83-0.92)
Tricuspid regurgitation	0.95 (0.93-0.97)	0.88 (0.86-0.91)	0.84 (0.82-0.86)	0.93 (0.90-0.95)	0.91 (0.89-0.93)
Pericardial effusion	0.98 (0.95-0.99)	0.89 (0.77-0.98)	0.87 (0.83-0.92)	0.96 (0.91-1.00)	1.00 (0.99-1.00)
Aortic root dilation	0.91 (0.83-0.98)	0.97 (0.94-0.99)	0.98 (0.96-0.99)	0.98 (0.97-1.00)	0.81 (0.78-0.85)
Dilated IVC	0.84 (0.82-0.86)	0.86 (0.81-0.91)	0.59 (0.49-0.69)	0.98 (0.94-1.00)	0.80 (0.77-0.82)
Diastolic dysfunction	0.91 (0.88-0.93)	0.86 (0.81-0.90)	0.93 (0.86-0.98)	-	0.84 (0.81-0.87)
PA pressure (R^2)	0.43 (0.39-0.47)	0.36 (0.30-0.42)	-	-	0.31 (0.27-0.35)
PA pressure (MAE)	7.30 (6.95-7.67)	7.97 (7.35-8.66)	-	-	7.39 (7.08-7.70)

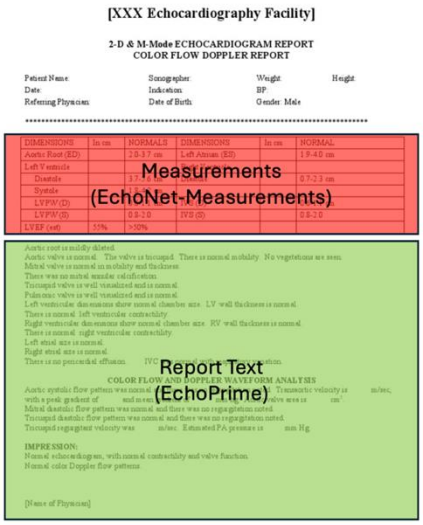
Comparison of EchoPrime vs EchoCLIP and BiomedCLIP on KP dataset is Added to Figure 2:



One limitation that is not evident on a casual reading of the paper is that you did not include in your model training any M-mode or spectral Doppler (PW & CW), which are essential to modern echocardiography (OK, maybe M-mode is a bit of a throw-back, but you can't quantify aortic stenosis without Doppler). This results in the fairly dismal results in Supplementary Table 7 for tasks like transaortic gradient, TR gradient, etc. This limitation needs to be stated more explicitly in the paper and its limitations with a discussion of how spectral Doppler might be included in the future.

Thank you for this opportunity to improve the presentation of our results. An interesting result from EchoPrime was that M-mode and spectral Doppler information was not necessary for most cardiac assessments, including Diastology which one would assume depends on such assessment. Perhaps there is additional information in B-mode videos that are not recognized by clinicians but are co-linear or complementary that can derive most of the assessments.

We have a parallel work for measurements and assessments done on M-Mode and Spectral Doppler which is our model for quantitative measurements (EchoNet-Measurements, https://www.medrxiv.org/content/10.1101/2025.03.18.25324215v1) and we anticipate will be used in parallel with EchoPrime, our vision-language model. This combination is highlighted by the fact that a full echo report combines both measurement and textual information for a holistic assessment of the ultrasound images, but for which ¾ is textual information a la EchoPrime.
We are currently planning a prospective, blinded clinical trial, which combines the use of EchoNet-Measurements and EchoPrime for a complete echocardiographic report. In this revision we discuss the differences and merits of each approach and how it can be used in parallel.



Finally, we highlight that in the usual clinical workflow, measurements are roughly 1/4 to 1/3 of the report, while the vast majority of echo reporting is based on natural language that EchoPrime, a vision-language model generates. On the left hand side, we show a canonical example of an echo report found online (for which many other similar examples can be found), and highlight the area focused on measurements which would be most consistent with EchoNet-Measurements in red, and the natural language component of the report most consistent with EchoPrime in green. We use this example to highlight the complementary but independent nature of these two types of tasks, as well as the relative weighting of these two types of tasks.

In the Discussion we add:

> A few limitations are worth considering. Since EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Many quantitative measurements are performed on still frames or images, therefore future versions should consider integrating the output of segmentation models[31,32] into the final report.

31. Sahashi, Y. *et al.* Artificial intelligence automation of echocardiographic measurements. Preprint at https://doi.org/10.1101/2025.03.18.25324215 (2025).
32. Tromp, J. *et al.* A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat. Commun.* **13**, 6776 (2022).

We would disagree that there is significant error in the regression tasks even for EchoPrime. Notably, we would like to highlight the MAE is comparable with prior AI work focused on measurements specifically (Sahashi et al and Tromp et al.) as well as has human variability (pmc.ncbi.nlm.nih.gov/articles/PMC11688108/). We would suggest focusing on comparison of the MAE as the relevant metric as the R2 is heavily affected by the distribution variance and the EchoPrime datasets were not chosen to represent the diversity of measurements.

Supplementary Table 8 EchoPrime's performance on regression tasks across three medical centers

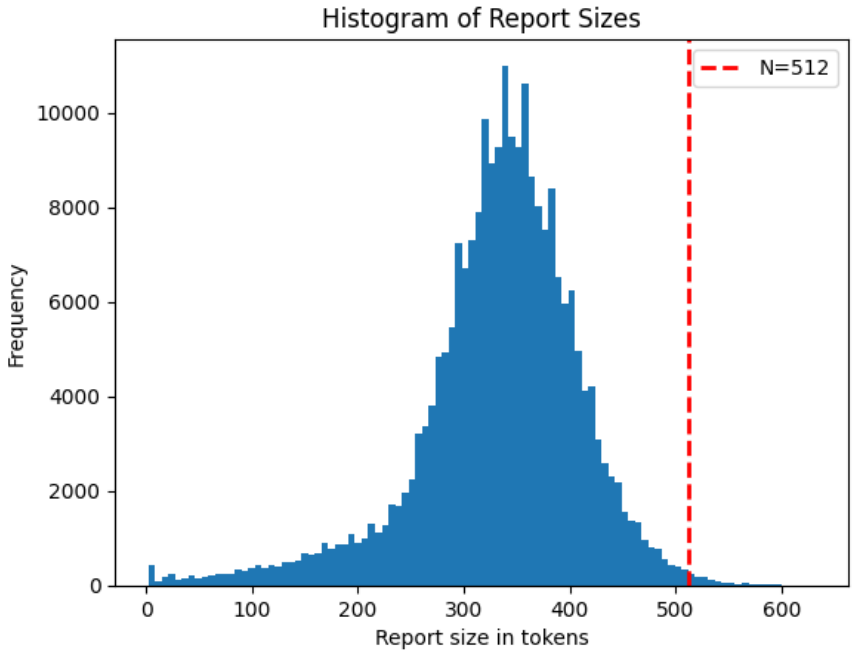
Task	CSMC	KP	SHC
Ejection Fraction R^2	0.83 (0.82-0.85)	0.66 (0.62-0.71)	0.79 (0.76-0.82)
Ejection Fraction MAE (%)	4.64 (4.47-4.81)	4.51 (4.30-4.72)	4.14 (3.94-4.34)
Pulmonary Artery Pressure R^2	0.42 (0.36-0.46)	0.31 (0.27-0.35)	0.36 (0.30-0.42)
Pulmonary Artery Pressure MAE (mmHg)	7.30 (6.96-7.65)	7.39 (7.08-7.70)	7.97 (7.35-8.66)
Wall Motion Score R^2	0.77 (0.74-0.80)	-	-
Wall Motion Score MAE (unitless)	0.10 (0.10-0.11)	-	-
TAPSE R^2	0.39 (0.26-0.51)	0.28 (0.02-0.46)	0.34 (0.00–0.70)
TAPSE MAE (cm)	0.31 (0.27-0.35)	0.35 (0.30-0.40)	0.25 (0.13-0.33)
MV Area R^2	0.34 (0.19-0.44)	-	0.22 (0.16-0.28)
MV Area MAE (cm^2)	0.78 (0.69-0.88)	-	1.03 (0.62-1.44)
Transaortic velocity R^2	0.70 (0.59-0.78)	0.50 (0.34-0.61)	0.54 (0.23-0.75)
Transaortic velocity MAE (cm/s)	20.4 (18.1-22.8)	56.4 (49.7-63.7)	51.1 (30.6-75.2)
Transaortic gradient R^2	0.69 (0.64-0.74)	0.50 (0.45-0.55)	0.38 (0.33-0.53)
Transaortic gradient MAE (mmHg)	5.00 (4.62-5.40)	6.02 (5.56 – 6.51)	7.32 (6.63-8.00)
TR gradient R^2	0.36 (0.30-0.41)	0.26 (0.12-0.43)	0.28 (0.23-0.32)
TR gradient MAE (mmHg)	6.72 (6.39-7.07)	6.33 (5.41-7.26)	7.41 (6.93-7.85)
Sinus of Valsalva Size R^2	0.62 (0.55-0.67)	0.39 (0.30-0.47)	-
Sinus of Valsalva Size MAE (cm)	0.24 (0.22-0.26)	0.33 (0.31-0.35)	-
Ascending Aorta Size R^2	0.36 (0.23-0.48)	0.34 (0.23-0.44)	0.26 (0.20-0.32)
Ascending Aorta Size MAE (cm)	0.31 (0.28-0.34)	0.34 (0.32-0.36)	0.40 (0.37-0.43)
IVC Diameter R^2	0.44 (0.38-0.49)	-	0.18 (0.14-0.23)
IVC Diameter MAE (cm)	0.29 (0.28-0.30)	-	0.44 (0.39-0.43)

In addition to these general comments, I have a few specific issues for you to consider.

Line 168: 512 tokens are indeed a significant improvement over the 77 in EchoCLIP and BioMedCLIP, but it remains a limitation when representing complex reports that may have more than 400 words and would contain the most detailed (and perhaps most useful to the model) RAI information. Is there a way to expand to a greater token length than the current 512.

Thank you for this important question. We checked 234,036 reports from our training set and 99.56% of the reports were shorter than 512 tokens. Less than 0.05% of reports were longer than 512 tokens, in which case we shuffled the text sampled during each training epoch such that all

relevant text information is presented at various epochs of training. We attach a full histogram of report lengths in tokens for your convenience.



In Methods “Contrastive Visual Language Training”

> PubMedBERT has a maximum context length of 512 tokens. The vast majority (99.56%) of the reports in our training dataset are 512 or fewer tokens, and longer reports were subsampled during each epoch to cover the entire report across multiple epochs.

Line 172-174: A typo that should be corrected.

Thank you for the opportunity to correct this typographical error.

> This marks a significant improvement over previous medical foundation models like BioMedCLIP and EchoCLIP, which process only single-view, individual images and handle text up to 77 tokens long.

Line 174-177: If I am reading this correctly, these are quite impressive results. I’m particularly struck that you could uniquely (at least top-10) find an essentially normal report (and echo going the other direction), as there must be 10’s of thousands of normal studies that look similar to each other. Did the 97-98% success extend to the many nearly identical normal echoes?

Thank you for this intriguing question. Similar to other vision-language models (like EchoCLIP and BioMedCLIP), at first glance there seems to be surprisingly good performance in uniquely identifying echo studies, even normal studies (for which we assume the degrees of freedom are less) and might be less distinguishable.

As we thought about this question, we have a few main points. First, EchoCLIP, our prior work with an image vision-language model, had similarly high retrieval with a mean cross-modal retrieval rank of 206 out of 21484 (or top 1%), while EchoPrime had a mean cross modal retrieval rank of 2-3 out of 2254 (similarly top 1%). This suggests this ability is not unique and potentially there is more uniqueness in medical imaging that clinicians are able to discern.

As Tolstoy might have put it “"All happy families [echos] are alike; each unhappy family [echo] is unhappy in its own way". However, one should consider that there is a fair amount of variation even among normal studies. For example, a study with an LVEF of 55% would be considered normal, while another study with LVEF of 65% would also be considered normal – however a clinician would be able to distinguish between two such studies. Considering that this only one element of an echo report and there are 10+ sections with many features each - a higher dimensional space might separate out differentiating factors within the span of “normal”. Due to the high level of detail in both the reports and the videos, there are very few “nearly identical normal” studies.

This is a particularly interesting question, and to better confirm this performance, we sought to assess retrieval performance across only normal and only abnormal studies. In this setting, echocardiogram studies with no significant findings and normal cardiac function were assessed for retrieval and found that this performance was worse than the set of echocardiogram studies with significant findings. This seems to confirm our intuition that retrieval tasks are easier in abnormal studies, but there is still significant information in variation within normal studies. In our EchoCLIP work, we show that this is able to distinguish between “significant” and “non-significant change” over time and highlight clinically meaningful deteriorations that have occurred to patients.

Supplementary Table 14 Retrieval metrics across normal and abnormal studies

	VIDEOS-TO-TEXT					TEXT-TO-VIDEOS				
	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10	Mean Rank	Media n Rank	Recall @1	Recall @5	Recall @10
Full Test (N=2254)	2.86	1	70%	94%	98%	3.05	1	69%	94%	97%
Normal Test (N=294)	6.02	1	56%	87%	94%	7.21	1	53%	89%	92%
Abnormal Test (N=1173)	2.38	1	72%	95%	98%	2.43	1	71%	95%	98%

We clarify this important point in the results.

In the Results:

> On videos-to-text retrieval, the correct text report appeared in the top 10 matched text reports for 98% of the studies from the test set (Recall@10), outperforming EchoCLIP by 45%. Similarly, on the text-to-videos task our model achieved Recall@10 of 97%, outperforming EchoCLIP by 35%. Retrieval was more challenging in structurally abnormal studies than structurally normal studies, although strong performance was possible in both cohorts (Supplementary Table 14).

Thank you for confirming that you have not used spectral Doppler or M-mode frames in the training or testing phases. You argue that much of their information can be obtained

from 2-D clips (for M-mode) and color Doppler clips (for spectral Doppler data, PW more than CW, given the impossibility of dealiasing the color Doppler for high-speed jets). However, this limitation does deserve explicit mention in the methods and limitations. It does help explain why your accuracy for RVSP was so much worse than EF, for example. Do you have plans to integrate these modalities into your model in future iterations?

Thank you for this opportunity to improve the presentation of our results. An interesting result from EchoPrime was that M-mode and spectral Doppler information was not necessary for most cardiac assessments, including Diastology which one would assume depends on such assessment. Perhaps there is additional information in B-mode videos that are not recognized by clinicians but are co-linear or complementary that can derive most of the assessments.

We have a parallel work for measurements and assessments done on M-Mode and Spectral Doppler which is our model for quantitative measurements (EchoNet-Measurements, https://www.medrxiv.org/content/10.1101/2025.03.18.25324215v1) and we anticipate will be used in parallel with EchoPrime, our vision-language model. We are currently planning a prospective, blinded clinical trial, which combines the use of EchoNet-Measurements and EchoPrime for a complete echocardiographic report. In this revision we discuss the differences and merits of each approach and how it can be used in parallel.

In the Discussion we add:

> A few limitations are worth considering. Since EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Many quantitative measurements are performed on still frames or images, therefore future versions should consider integrating the output of segmentation models[31,32] into the final report.

31. Sahashi, Y. *et al.* Artificial intelligence automation of echocardiographic measurements. Preprint at https://doi.org/10.1101/2025.03.18.25324215 (2025).
32. Tromp, J. *et al.* A formal validation of a deep learning-based automated workflow for the interpretation of the echocardiogram. *Nat. Commun.* **13**, 6776 (2022).

There seems to be an error with the description of the contrastive learning set up. You say all clip-report pairs are from different studies. This would lead to non-informative training as all of the pairs would be different or 0 in identity matrix. To learn joint embeddings like this some of the pairs should be matching (ie, from same study) which is the typical set up for this. They need to clarify as this might just be an error.

Thank you for this opportunity to clarify our methods. We clarify that this means that clip-report pairs were indeed from the same study, but each pair in the batch came from a different study.

We clarify the explanation in **Methods "Contrastive Visual Language Training" section**

> During training, batches of 32 video-report pairs were constructed, with each pair consisting of a video and report from the same study, and all pairs within a batch coming from different studies.

The self-supervised way that you produce anatomical attention is actually very clever. Once you have trained an image and text model that share the same embedding space, you can then look at individual sections of the text and see if which views are most "important" for predicting that particular section. You might expand a bit on this in the discussion, which really is quite brief given all the innovative techniques introduced here and doesn't really put EchoPrime into context regarding previous attempts. In addition, as you discuss how this might be applied in resource limited settings, a reference to AI methods for echo guidance would be appropriate.

Thank you for your commends. We expand our discussion section to highlight the anatomical attention module.

We revise the **Discussion** accordingly:

> Compared to prior echocardiographic models, EchoPrime has a longer temporal context, longer text context length, and integrates complementary videos with a robust view classifier and multiple instance anatomical attention.. By utilizing multiple instance learning to integrate information from multiple videos, we reproduce the relationship between echocardiographic views and the associated anatomic structures that cardiologists use to make assessments, making EchoPrime an inherently clinically interpretable model.

Thank you for noticing a potential of combining EchoPrime with AI methods for echo guidance. While ultrasound machines are widely available and inexpensive to maintain, there is often a shortage of trained personnel, and the primary bottleneck and expense of an echocardiographic exam is clinicians' time and lack of enough clinicians with expertise. We extend our discussion to mention the potential of combining EchoPrime with automatic probe guidance methods.

In the **Discussion**

> Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate many steps in echocardiographic imaging workflow.

I am still a bit confused as to how the individual level predictions were made and evaluated by EchoPrime. I understand that 50 most similar reports from training set were retrieved and it seems that the embeddings (or "features" – what features specifically?) from these reports were then averaged. How did you go from this to individual level predictions or even an entire report generation (also not clear which of these they actually did). Essentially, what is the actual output of this model?

Thank you for this opportunity to clarify our approach. After identifying the 50 most similar reports, we ensemble findings based on frequency or averaging for numerical tasks. For example, we would have 50 unique LVEFs across 50 reports, this LVEF will be averaged to produce the final EchoPrime report. For categorial findings, the final EchoPrime report displays the finding if an individual finding is present in more than half of the most similar reports. In this revision, we clarify in the methods:

Retrieval Augmented Interpretation

> Similar to retrieval augmented generation[18], we leverage the corpus of historical echocardiogram reports and videos for EchoPrime study interpretation. Using pre-trained parametric memory (EchoPrime weights) and non-parametric memory (a corpus of clinical reports), our approach leverages anatomic attention to weight the cross-modal retrieval based on each section in an echocardiogram report. Interpretations are weighted by anatomy and video to obtain comprehensive echocardiography exam interpretations. The anatomical attention module identifies the most informative views for the specified section and computes a weighted average of video embeddings based on their views, resulting in a section-specific study embedding. We use this method to generate 50 candidate interpretation reports and average features across these 50 reports to obtain final predictions. Quantitative assessments are the average of the 50 reports, while the final EchoPrime report displays the finding if an individual finding is present in more than half of the retrieved reports for categorial findings. Hyperparameter k=50 is chosen because it achieves the optimal trade-off between performance and runtime (Supplementary Figure 4). This approach does not require additional fine-tuning and can generalize to predict any feature present in the reports.

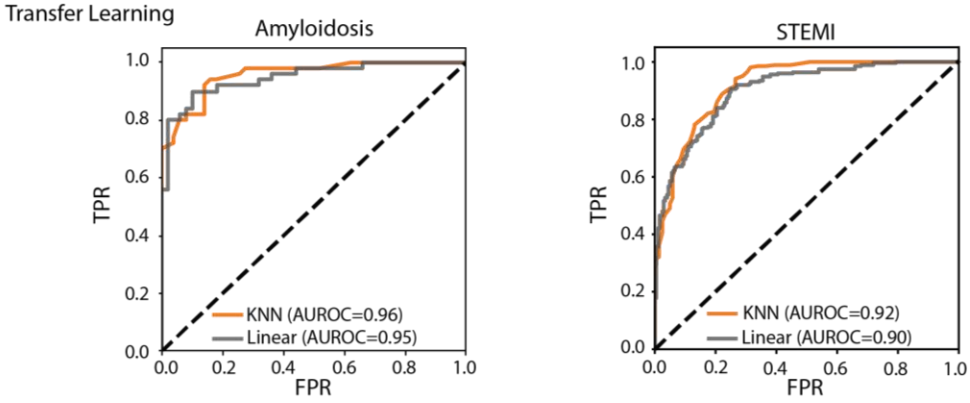
I do have concerns with the "retrieval augmented interpretation" approach and how this might work on rare findings that might not be highly represented in the training dataset. If the findings are not well represented, then you won't be able to retrieve enough relevant reports for averaging the features to generate the correct output. For example, can your model recognize something like (for example) Ebstein's anomaly or congenitally corrected TGA? Also, as novel echo methods evolve, will this need to be completely retrained rather than just fine tuned. The whole concept of a "foundational model" is that it can be used and adapted for many downstream tasks, including fine-tuning of the model. The problem I see here is that the "retrieval augmented interpretation" framework does not allow for an obvious approach to fine tuning, as retrieval is happening with the original training approach. If you could detail how this could be used downstream for fine-tuning, it would be helpful.

Thank you for this important question. Our model was trained on 12M videos from more than 275k studies at a tertiary care hospital's echo lab, so even relatively uncommon findings are captured in the training data. This dataset represents 10-100x larger dataset than other datasets used to train other echocardiography models, although we agree that there might still be some particularly rare findings that are not identified in this corpus. Furthermore, for EchoPrime training, we excluded pediatric and adult congenital echocardiogram studies from the initial

iteration, focusing only on echocardiograms interpreted by general cardiologists. We will look at congenital findings (Ebsteins, ccTGA, dTGA) in future models. In the methods, we clarify our inclusion criteria:

> *Each echocardiogram study corresponds to a unique patient during a specific examination, with different videos representing different facets of cardiac function. We utilized all adult transthoracic echocardiogram studies for EchoPrime training, excluding studies for pediatric echocardiography and adult congenital heart disease echocardiograms. For EchoPrime training, we curated a dataset of 12,124,168 2D and color Doppler echocardiography videos from 275,442 studies and corresponding medical reports from 108,913 patients collected between 2011 and 2022.*

As the initial contrastive learning was not based on RAI, there is still significant value in EchoPrime as a foundation model, and further video-text pairs from congenital heart disease cases would apply similarly. This is similar to how we could fine tune for external sites that use different vocabulary or languages (although in this manuscript, we did not fine-tune for external validation). Given the rich embedding space, there are features of diagnoses that can be identified even if not directly diagnosed through echo – for example our experiments with multi-modal disease diagnoses - STEMI and cardiac amyloidosis. Linear probing from a rich embedding is a fine-tuning approach that requires minimal time and computational resources—it can be trained in just minutes on a moderately powered computer. We employed linear probing on limited datasets to predict STEMI (N=641) and Amyloidosis (N=159), and achieved a strong performance on both (Figure 2B).

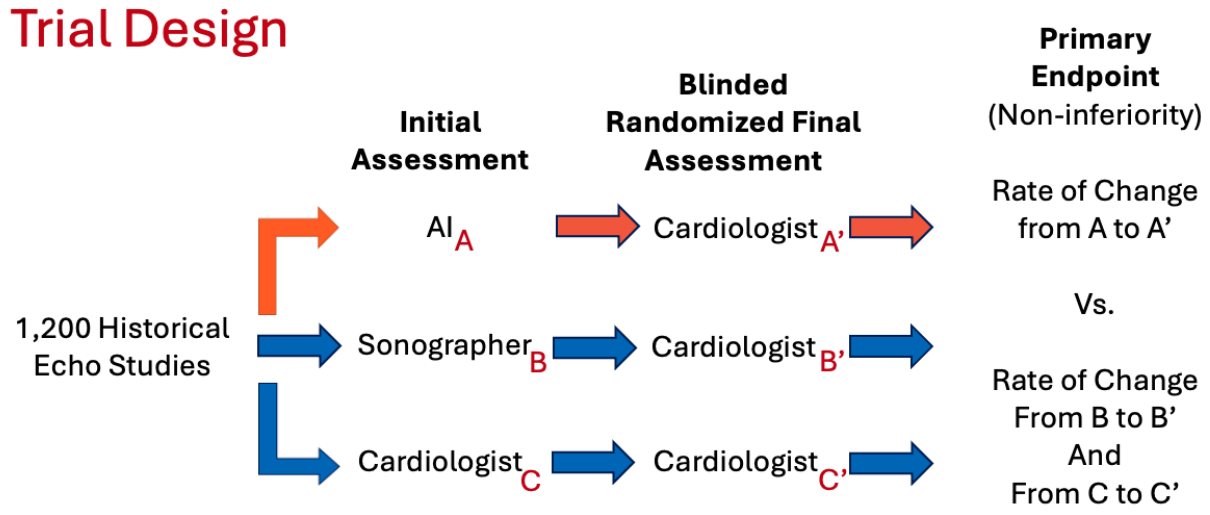


Referee #3:

The authors did not address several concerns I raised in the first round. I do not agree with their responses, which rely on discussion rather than real experiments or clinical evaluations conducted by human physicians.

Although the authors claim that EchoPrime can assist physicians in the automated preliminary assessment of comprehensive echocardiography following rigorous clinical validation, no human evaluation has been performed. The current assessment relies solely on quantitative metrics such as AUC, which limits the study's clinical relevance. Despite the dataset being comprehensive, the analysis remains entirely retrospective. Prospective experiments would provide stronger evidence of EchoPrime's applicability in real clinical settings and would offer more convincing validation of its utility. The absence of such analysis significantly undermines the study's validity and potential impact.

Thank you for this opportunity to clarify how EchoPrime can be used in the clinical setting. We anticipate conducting a prospective clinical trial of EchoPrime shortly to directly assess prospective performance and acceptability to cardiologists that are reviewing the echocardiogram images. We have already acquired funding and institutional support to conduct this prospective evaluation, which potentially speaks to our internal understanding how it would compare with current standard of care. In this trial setting, we will randomize cardiologists to preliminary assessments of echocardiograms by EchoPrime vs. preliminary assessments performed by cardiologists and sonographers and track the rate of change as cardiologist's conducted blinded review.



We recognize the importance of prospective evaluation, though such studies require substantial time and financial resources. While we are currently starting our prospective clinical trial and not yet able to present prospective data supporting EchoPrime's improved precision, in this revision, we add additional experiments to show that EchoPrime's precision across different studies from the same patient when there is no clinical change is higher than inter-physician variability in interpreting studies when there is "no significant change".

In this new experiment, we construct a new test cohort to evaluate variability by cardiologists. This clinical cohort comprises sequential echocardiography studies from the same patient for which a cardiologist declared there was no significant change between the two studies, and we compared the variability between the two studies in key quantitative and qualitative assessments. In this comparison, we show EchoPrime has similar or improved precision when evaluated on the same two studies, suggesting improved precision for EchoPrime compared to cardiologists:

Supplementary Table 13 EchoPrime vs Physician variability compared to Physician vs Physician variability. Reported values are balanced accuracies.

N=4607	ECHOPRIME VS PHYSICIAN	PHYSICIAN VS PHYSICIAN
LV ejection fraction (R^2)	**0.83 (0.81-0.85)**	0.73 (0.70-0.75)
LV ejection fraction (MAE)	4.79 (4.60-4.97)	**4.59 (4.47-4.72)**
Pacemaker	0.78 (0.75-0.80)	**0.83 (0.80-0.85)**
RV systolic function	**0.88 (0.85-0.90)**	0.82 (0.79-0.86)
RV dilation	**0.87 (0.81-0.92)**	0.71 (0.66-0.76)
LA dilation	**0.86 (0.81-0.92)**	0.77 (0.72-0.82)
RA dilation	**0.90 (0.84-0.96)**	0.76 (0.69-0.84)
Mitraclip	**0.98 (0.95-0.99)**	0.96 (0.93-0.98)
Mitral annular calcification	**0.90 (0.87-0.92)**	0.78 (0.75-0.81)
Mitral stenosis	**0.92 (0.91-0.93)**	0.70 (0.59-0.82)
Mitral regurgitation	0.86 (0.83-0.88)	**0.87 (0.85-0.90)**
TAVR	0.98 (0.98-0.99)	**0.99 (0.98-0.99)**
Bicuspid AV	0.84 (0.70-0.96)	**0.87 (0.81-0.92)**
Aortic stenosis	**0.95 (0.93-0.96)**	0.83 (0.78-0.88)
Aortic regurgitation	**0.83 (0.78-0.88)**	0.75 (0.69-0.80)
Tricuspid regurgitation	**0.90 (0.88-0.93)**	0.86 (0.83-0.89)
Pericardial effusion	**0.91 (0.87-0.94)**	0.88 (0.81-0.94)
Aortic root dilation	**0.87 (0.79-0.94)**	0.77 (0.71-0.82)
PA pressure (R^2)	0.43 (0.39-0.47)	**0.50 (0.44-0.55)**
PA pressure (MAE)	7.30 (6.95-7.67)	**5.89 (5.70-6.09)**
Wall Motion Score (R^2)	**0.77 (0.74-0.80)**	0.76 (0.72-0.80)
Wall Motion Score (MAE)	0.10 (0.10-0.11)	**0.04 (0.04-0.05)**
MV Area (R^2)	0.34 (0.19-0.44)	**0.38 (0.13-0.59)**
MV Area (MAE)	0.78 (0.69-0.88)	**0.70 (0.60-0.81)**
Transaortic gradient (R^2)	0.69 (0.64-0.74)	**0.82 (0.78-0.86)**
Transaortic gradient (MAE)	5.00 (4.62-5.40)	**3.12 (2.97-3.27)**
TRgradient (R^2)	0.36 (0.30-0.41)	**0.50 (0.33-0.67)**
TRgradient (MAE)	6.72 (6.39-7.07)	**5.42 (5.21-5.66)**
Sinus of Valsalva Size (R^2)	**0.62 (0.55-0.67)**	0.42 (0.06-0.87)
Sinus of Valsalva Size (MAE)	0.24 (0.22-0.26)	**0.21 (0.16-0.30)**
Ascending Aorta Size (R^2)	0.36 (0.23-0.48)	**0.81 (0.77-0.85)**
Ascending Aorta Size (MAE)	0.31 (0.28-0.34)	**0.17 (0.16-0.19)**
IVC Diameter (R^2)	**0.44 (0.38-0.49)**	0.05 (-0.68-0.28)
IVC Diameter (MAE)	**0.29 (0.28-0.30)**	0.32 (0.31-0.34)

While prospective experiments would provide a stronger validation of EchoPrime's effectiveness in real-world, and we are currently implementing a prospective clinical trial for EchoPrime, which we anticipate will be ready for presentation in 2026. We will formally measure time taken by cardiologists and sonographers in the study.

In the Methods:

> To assess inter-physician variability in echocardiography interpretations, we evaluated clinician interpretations and EchoPrime interpretations on a cohort of patients undergoing consecutive echocardiogram studies for which a cardiologist thought there was no

significant difference between the two studies. This dataset comprised of 4607 paired studies for which we evaluated the degree of variation between the first and second studies of the same patient.

In the Results:

To contextualize EchoPrime's performance and reproducibility of assessments, we assessed cardiologist inter-observer variability in a series of paired echocardiogram studies from the same patient for which cardiologists noted no significant change. We found that EchoPrime's agreement with cardiologists is comparable to the agreement between two cardiologists as EchoPrime had an average balanced accuracy of 0.89, while the cardiologists had a balanced accuracy of 0.82 (Supplementary Table 13).

In the Discussion:

Although EchoPrime was evaluated at five geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in the real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime preliminary assessment. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate many steps in echocardiographic imaging workflow.

Video Encoder: The authors note that their video encoder is based on a mViT checkpoint. However, mViT was introduced in 2021, and more recent methods such as BEVT [1] and VideoMAE [2] offer notable improvements. Incorporating or comparing against these recent models would strengthen the technical contribution and demonstrate a more thorough engagement with current developments.

Thank you for this careful evaluation of EchoPrime. We recognize that is this a rapidly improving and evolving field with what seems like new developments every week. In this revision, we benchmark with the comparator recommended by the reviewer, but would like to recognize that this takes extensive time/compute for each experiment (our VideoMAE comparison described below alone took three weeks) and that the strong performance of EchoPrime was based on the original architecture as described.

Table 1: Cross-modal retrieval metrics on a test set of videos-text report pairs. Lower rank is better, and higher recall is better. Best performance is bolded. EchoPrime was compared with different video encoders (VideoMAE and ResNet2+1D) trained on the same dataset.

	Videos-To-Text					Text-To-Videos				
	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10	Mean Rank	Median Rank	Recall @1	Recall @5	Recall @10
EchoPrime	**2.86**	**1**	**70%**	**94%**	**98%**	**3.05**	**1**	**69%**	**94%**	**97%**
VideoMAE	4.24	1	67%	91%	95%	4.44	1	66%	91%	95%

Resnet2+1d	5.41	1	63%	91%	95%	5.50	1	62%	90%	95%
EchoCLIP	58.71	9	20%	42%	53%	37.07	5	25%	50%	62%

In the results we describe:

> We experimented with batch size, video encoder architecture, video weighting approaches, and number of candidate reports (Supplementary Figure 4, Supplementary Tables 9-11) achieve the final EchoPrime hyperparameters and performance (**Error! Reference source not found.**).

Supplementary Table 15 Comparison of mVIT encoder and VideoMAE encoder on core metrics across five datasets

	CSMC		SHC		MIMIC		CGMH		KP	
	mVIT	Video MAE	mVIT	Video MAE	mVIT	Video MAE	mVIT	Video MAE	mVIT	Video MAE
LV ejection fraction (R^2)	**0.83**	0.80	**0.79**	0.77	0.74	**0.75**	**0.51**	0.50	0.66	**0.69**
LV ejection fraction (MAE)	**4.79**	4.98	**4.14**	4.29	**6.21**	6.24	**6.46**	6.58	4.51	**4.34**
Pacemaker	**0.85**	0.80	**0.84**	0.80	**0.86**	0.79	0.78	**0.79**	**0.84**	0.79
RV systolic function	**0.93**	0.92	**0.94**	0.89	**0.98**	0.89	0.90	**0.92**	0.92	**0.93**
RV dilation	0.89	**0.90**	**0.85**	0.83	-	-	**0.87**	0.81	0.78	**0.81**
LA dilation	**0.91**	0.90	**0.73**	0.70	**0.72**	**0.72**	**0.65**	0.64	**0.77**	**0.77**
RA dilation	0.90	**0.92**	**0.77**	**0.77**	**0.65**	0.62	**0.75**	0.72	**0.82**	0.76
Mitraclip	**0.99**	0.93	**0.98**	**0.98**	**0.99**	**0.99**	-	-	**1.00**	0.99
Mitral annular calcification	**0.96**	0.93	**0.96**	0.95	**0.88**	**0.88**	**0.88**	0.81	**0.89**	0.86
Mitral stenosis	**0.96**	**0.96**	**0.92**	0.88	0.96	**0.99**	**0.79**	**0.79**	**0.88**	**0.88**
Mitral regurgitation	**0.92**	0.90	0.91	**0.94**	**0.81**	**0.81**	0.86	**0.87**	0.89	**0.90**
TAVR	**1.00**	0.98	0.97	**0.98**	**0.96**	0.94	**1.00**	**1.00**	**0.99**	0.98
Bicuspid AV	**0.83**	0.70	**0.82**	0.75	**0.65**	0.61	**0.67**	0.66	**0.80**	0.69
Aortic stenosis	**0.98**	0.91	**0.96**	**0.96**	0.88	**0.89**	**0.99**	0.98	**0.97**	0.93
Aortic regurgitation	**0.88**	0.79	**0.89**	0.85	**0.81**	0.75	**0.78**	0.76	**0.87**	0.81
Tricuspid regurgitation	**0.95**	0.94	**0.88**	0.85	**0.84**	0.79	**0.93**	0.89	**0.91**	0.90
Pericardial effusion	**0.98**	0.96	**0.89**	0.82	**0.87**	0.84	0.96	**0.98**	**1.00**	**1.00**
Aortic root dilation	**0.91**	0.80	**0.97**	0.86	**0.98**	0.79	**0.98**	0.71	**0.81**	0.74
Dilated IVC	**0.84**	0.83	**0.86**	0.76	**0.59**	0.50	**0.98**	0.79	**0.80**	0.76
PA pressure (R^2)	**0.43**	0.40	**0.36**	0.30	-	-	-	-	0.31	**0.32**
PA pressure (MAE)	**7.30**	7.38	**7.97**	8.15	-	-	-	-	**7.39**	7.42

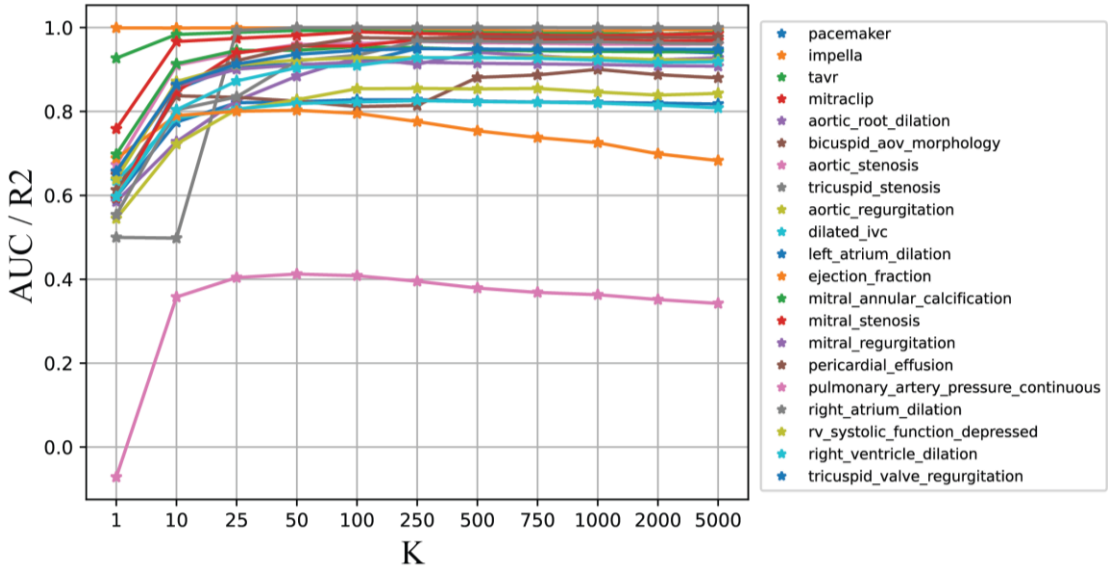
We also note that pretraining approaches are most useful in the limited data regimen and has diminishing returns as the training dataset size improves (consider Table 2 of the VideoMAE paper and the Table 5 of BEVT paper). This intuition is confirmed by our experiment with VideoMAE, in which we used a VideoMAE fine-tuned video encoder to train an EchoPrime variant with the same training dataset, as performance does not significantly improve.
It is important to understand the various trade offs and optimizations that are possible with AI models. While our academic lab is resource constrained (we can only do a set number of hyperparameter sweeps given available compute), to better understand the landscape and parameters of training Echo AI models, we conducted additional experiments to justify our hyperparameter and architectures choices.

Supplementary Table 11 EchoPrime trained on varying batch sizes

BATCH SIZE	16	32	50
LV ejection fraction (R^2)	0.81 (0.79-0.83)	0.83 (0.81-0.85)	0.80 (0.78-0.82)
LV ejection fraction (MAE)	4.87 (4.69-5.06)	4.79 (4.60-4.97)	4.94 (4.75-5.13)
Pacemaker	0.85 (0.82-0.87)	0.85 (0.82-0.88)	0.87 (0.84-0.89)
RV systolic function	0.93 (0.90-0.95)	0.93 (0.91-0.95)	0.93 (0.90-0.95)

RV dilation	0.88 (0.82-0.94)	0.89 (0.83-0.95)	0.92 (0.87-0.96)
LA dilation	0.92 (0.87-0.96)	0.91 (0.86-0.96)	0.94 (0.90-0.97)
RA dilation	0.94 (0.88-0.98)	0.90 (0.82-0.97)	0.94 (0.88-0.99)
Mitraclip	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.99-1.00)
Mitral annular calcification	0.96 (0.94-0.97)	0.96 (0.95-0.97)	0.95 (0.92-0.97)
Mitral stenosis	0.91 (0.82-0.97)	0.96 (0.91-0.99)	0.92 (0.85-0.99)
Mitral regurgitation	0.93 (0.92-0.95)	0.92 (0.91-0.94)	0.93 (0.92-0.95)
TAVR	0.99 (0.98-1.00)	1.00 (0.99-1.00)	0.99 (0.98-1.00)
Bicuspid AV	0.83 (0.67-0.96)	0.83 (0.67-0.96)	0.83 (0.67-0.96)
Aortic stenosis	0.98 (0.96-0.99)	0.98 (0.96-0.99)	0.98 (0.96-0.99)
Aortic regurgitation	0.87 (0.82-0.92)	0.88 (0.83-0.93)	0.90 (0.85-0.94)
Tricuspid regurgitation	0.95 (0.94-0.97)	0.95 (0.93-0.97)	0.96 (0.94-0.97)
Pericardial effusion	0.99 (0.98-0.99)	0.98 (0.95-0.99)	0.97 (0.93-0.99)
Aortic root dilation	0.92 (0.83-0.99)	0.91 (0.83-0.98)	0.91 (0.82-0.98)
Dilated IVC	0.83 (0.81-0.85)	0.84 (0.82-0.86)	0.83 (0.81-0.85)
PA pressure (R^2)	0.42 (0.37-0.46)	0.43 (0.39-0.47)	0.42 (0.38-0.47)
PA pressure (MAE)	7.31 (6.94-7.68)	7.30 (6.95-7.67)	7.26 (6.89-7.62)

Sweeping the number of candidates during retrieval:



Supplementary Figure 4 *Performance of Retrieval Based Interpretation as we increase the number of candidates reports to average across. For binary tasks the metric is AUROC and for regression tasks R2 score.*

We anticipate future iterations will continue to improve upon EchoPrime, but EchoPrime should be considered in the landscape of other AI in medicine models, for which we believe it compares highly favorably. The performance of EchoPrime is with the choices made currently and we believe this performance is state-of-the-art for echocardiography AI at this time. The architecture and design choices we had for EchoPrime were deliberate and thought-out, however we recognize that this is a rapidly evolving field and there continues to be room for improvement.

Referee #1 (Remarks to the Author):

I appreciate the substantial revisions, particularly the additional continuous-metric analyses:

a) for aortic-root dimensions the updated site-specific MAE and R^2 values are now reported (Supplementary Table 8);
b) for diastolic function you have added an explicit AUROC row (Table 2);
c) the new Bland-Altman plots (Supplementary Fig. 5) are helpful, but visual inspection reveals residual proportional bias and heteroscedasticity in several tasks (e.g., pulmonary-artery pressure, trans-aortic velocity/gradient, TR gradient, IVC diameter). Differences become increasingly negative at higher true values (under-prediction) and the error spread widens, particularly for trans-aortic gradient. In addition, centre-specific colour-coding suggests distribution shifts across sites (e.g., IVC diameter). Please quantify these effects (slope ± 95 % CI of difference vs. mean and limits-of-agreement by value decile), report whether slopes differ from zero, and clarify whether any statistical recalibration (overall or site-specific) has been applied or is planned. If proportional bias is significant, a simple linear-correction table or isotonic recalibration would let readers judge the clinical impact.

REPLY: Thank you for your careful review. In our prior R2 response, we described that we think EchoNet-Measurements is the right comparison model for measurements while EchoPrime is focused on text reports. Given the parallel approaches with different pros and cons, a complementary combination approach is likely most useful in future clinical use.

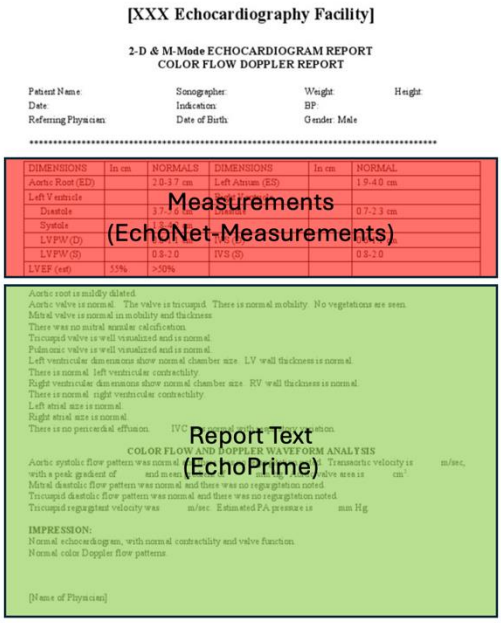
A comparison of the models:

EchoPrime	**PanEcho**	**EchoNet-Measurements**
Contrastive Vision-Language Model, both video encoder CNN and text encoder (BERT)	Supervised Training, Convolutional Neural Network	Supervised Training, Convolutional Neural Network
Outputs Natural Language Text Reports	Outputs classifications and numbers	Outputs segmentation annotations and associated numbers
Can do all echo text tasks as it generates natural language report language	Performs 39 specific tasks	Performs 18 specific tasks
Trained on 12M videos	Trained on 1.2M videos	Trained on 887k videos
View informed inference	Not view informed inference	single view inference
External validation at 4 geographically independent sites with comprehensive echo videos from full studies	External validation on 2 public single view datasets (EchoNet-Dynamic, EchoNet-LVH) and one geographically distinct site	External validation at 1 geographically independent site

	that include the full echocardiogram studies (RVENet).	

Per the suggestion of the reviewer, the evaluated PanEcho on the tasks we already prespecified in the manuscript. Among tasks we evaluated, EchoPrime is superior to PanEcho in measuring RVSP and LVEF (Supplementary Table 2 – 6) across all 5 validation sites, given our shortened timelines, we did not perform additional regression tasks (as we think the right comparison is with the EchoNet-Measurements paper and the comparisons we already did show EchoPrime is better than PanEcho for the tasks we assessed).

In the usual clinical workflow, measurements are roughly 1/4 to 1/3 of the report, while the vast majority of echo reporting is based on natural language that EchoPrime, a vision-language model generates. On the left hand side, we show a canonical example of an echo report found online (for which many other similar examples can be found), and highlight the area focused on measurements which would be most consistent with EchoNet-Measurements in red, and the natural language component of the report most consistent with EchoPrime in green. We use this example to highlight the complementary but independent nature of these two types of tasks, as well as the relative weighting of these two types of tasks.



My remaining questions centre on the incremental value of EchoPrime in light of PanEcho (JAMA 2025):

1. Benchmarking vs. PanEcho: Please provide a direct side-by-side comparison (classification AUROC, regression MAE/R^2, external-site count, compute) so readers can judge whether EchoPrime offers measurable performance gains or complementary tasks.

REPLY: Thank you for your careful review. Since the last revision, an important comparison model has come out (PanEcho in JAMA), for which we provide benchmark comparison across each of our each test datasets (both external and internal). In these experiments, we show EchoPrime has comparable or better performance across most tasks compared to PanEcho. **Best performing model in bold.** PanEcho performed similar or better on chamber size tasks, while EchoPrime performed better in all other tasks. We speculate this is because PanEcho uses an image encoder and transformer downstream for temporal integration of information, while EchoPrime uses a video encoder, but this is a ripe area for future research.

Given PanEcho is 'view agnostic' while EchoPrime uses anatomic attention, we see the most benefit in limited resource settings where optimal view selection improves model performance (Supplementary Table 9).

Supplementary Table 2 *Performance of foundation models for biomedical imaging on Cedars-Sinai test set*

CEDARS-SINAI TEST SET						
	PubMedCLIP	*PMC-CLIP*	*BioMedCLIP*	*EchoCLIP*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	*-0.32 (-0.40--0.25)*	*-0.38 (-0.46--0.31)*	*-2.60 (-3.00--2.20)*	*0.62 (0.59-0.65)*	***0.83 (0.81-0.85)***	*0.52 (0.50-0.54)*
LV ejection fraction (MAE)	*14.7 (14.3-15.1)*	*14.0 (13.6-14.5)*	*26.93 (26.4-27.4)*	*7.00 (6.7-7.3)*	***4.8 (4.6-5.0)***	*7.46 (7.16-7.78)*
Pacemaker	*0.50 (0.50-0.50)*	*0.51 (0.47-0.54)*	*0.58 (0.55-0.62)*	***0.88 (0.85-0.91)***	*0.85 (0.82-0.88)*	-
RV systolic function	*0.50 (0.49-0.51)*	*0.53 (0.48-0.58)*	*0.64 (0.59-0.69)*	*0.89 (0.86-0.92)*	***0.93 (0.91-0.95)***	*0.89 (0.87-0.92)*
RV dilation	*0.50 (0.45-0.56)*	*0.59 (0.50-0.67)*	*0.58 (0.52-0.64)*	*0.88 (0.83-0.92)*	***0.89 (0.83-0.95)***	*0.86 (0.79-0.91)*
LA dilation	*0.50 (0.50-0.50)*	*0.57 (0.48-0.65)*	*0.66 (0.58-0.74)*	***0.93 (0.89-0.96)***	*0.91 (0.86-0.96)*	*0.89 (0.86-0.92)*
RA dilation	*0.50 (0.50-0.50)*	*0.58 (0.47-0.68)*	*0.67 (0.59-0.75)*	*0.90 (0.82-0.97)*	*0.90 (0.82-0.97)*	***0.93 (0.89-0.96)***
Mitraclip	*0.50 (0.50-0.50)*	*0.49 (0.43-0.55)*	*0.70 (0.63-0.76)*	*0.95 (0.92-1.00)*	***0.99 (0.99-1.00)***	-
Mitral annular calcification	*0.50 (0.46-0.55)*	*0.59 (0.54-0.64)*	*0.63 (0.58-0.68)*	*0.95 (0.93-0.96)*	***0.96 (0.95-0.97)***	-
Mitral stenosis	*0.50 (0.50-0.50)*	*0.55 (0.44-0.66)*	*0.70 (0.61-0.79)*	***0.96 (0.91-0.99)***	***0.96 (0.91-0.99)***	***0.96 (0.93-0.98)***
Mitral regurgitation	*0.50 (0.50-0.50)*	*0.59 (0.55-0.63)*	*0.66 (0.63-0.69)*	*0.89 (0.87-0.91)*	***0.92 (0.91-0.94)***	*0.86 (0.84-0.89)*
TAVR	*0.50 (0.50-0.51)*	*0.51 (0.46-0.55)*	*0.54 (0.50-0.57)*	*0.86 (0.83-0.89)*	***1.00 (0.99-1.00)***	-
Bicuspid AV	*0.50 (0.50-0.50)*	*0.51 (0.47-0.60)*	*0.52 (0.51-0.53)*	*0.72 (0.58-0.85)*	*0.83 (0.67-0.96)*	***0.95 (0.89-0.98)***
Aortic stenosis	*0.50 (0.50-0.52)*	*0.48 (0.42-0.55)*	*0.55 (0.53-0.57)*	*0.83 (0.79-0.87)*	***0.98 (0.96-0.99)***	*0.93 (0.90-0.95)*
Aortic regurgitation	*0.50 (0.50-0.50)*	*0.50 (0.41-0.58)*	*0.54 (0.53-0.55)*	*0.68 (0.60-0.76)*	***0.88 (0.83-0.93)***	*0.77 (0.71-0.83)*
Tricuspid regurgitation	*0.49 (0.45-0.53)*	*0.57 (0.51-0.63)*	*0.69 (0.64-0.74)*	*0.91(0.89-0.93)*	***0.95 (0.93-0.97)***	*0.91 (0.88-0.94)*
Pericardial effusion	*0.54 (0.45-0.62)*	*0.50 (0.42-0.57)*	*0.68 (0.59-0.77)*	*0.92 (0.89-0.95)*	***0.98 (0.95-0.99)***	*0.93 (0.89-0.96)*
Aortic root dilation	*0.50 (0.50-0.50)*	*0.44 (0.35-0.55)*	*0.50 (0.49-0.51)*	*0.76 (0.63-0.88)*	***0.91 (0.83-0.98)***	*0.66 (0.56-0.76)*
Dilated IVC	*0.50 (0.50-0.51)*	*0.53 (0.50-0.56)*	*0.49 (0.47-0.50)*	*0.81 (0.79-0.83)*	***0.84 (0.82-0.86)***	-
PA pressure (R^2)	*-0.59 (-0.74--0.46)*	*-0.83 (-1.07--0.62)*	*-0.19 (-0.24 - -0.14)*	*0.35 (0.29-0.41)*	***0.43 (0.39-0.47)***	*0.27 (0.23-0.30)*
PA pressure (MAE)	*11.3 (10.0-12.6)*	*13.7 (13.0-14.3)*	*10.6 (10.1-11.12*	*7.8 (7.5- 8.2)*	***7.3 (7.0-7.7)***	*8.27 (7.88-8.68)*

Supplementary Table 3 *Performance of foundation models for biomedical imaging on Stanford dataset*

STANFORD DATASET				
	BioMedCLIP	*EchoCLIP*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	*-3.50 (-4.15- -3.03)*	*0.48 (0.40-0.55)*	***0.79 (0.76-0.82)***	*0.39 (0.36-0.42)*
LV ejection fraction (MAE)	*23.00 (22.68-23.79)*	*6.23 (5.89-6.44)*	***4.14 (3.94-4.34)***	*6.41 (6.03-6.79)*
Pacemaker	*0.64 (0.60-0.68)*	***0.90 (0.88-0.93)***	*0.84 (0.81-0.87)*	-
RV systolic function	*0.59 (0.02-0.92)*	*0.89 (0.73-1.00)*	***0.94 (0.88-1.00)***	*0.93 (0.80-0.99)*
RV dilation	*0.53 (0.49-0.58)*	*0.84 (0.79-0.89)*	*0.85 (0.80-0.90)*	***0.87 (0.83-0.90)***
LA dilation	*0.57 (0.55-0.60)*	*0.67 (0.65-0.70)*	*0.73 (0.70-0.75)*	***0.81 (0.79-0.83)***
RA dilation	*0.56 (0.52-0.59)*	*0.72 (0.69-0.75)*	*0.77 (0.74-0.81)*	***0.84 (0.81-0.87)***
Mitraclip	*0.79 (0.78-0.81)*	*0.81 (0.44-1.00)*	***0.98 (0.94-1.00)***	-
Mitral annular calcification	*0.66 (0.56-0.76)*	*0.94 (0.90-0.97)*	***0.96 (0.94-0.98)***	-
Mitral stenosis	*0.72 (0.63-0.81)*	*0.84 (0.75-0.92)*	*0.92 (0.86-0.98)*	***0.96 (0.94-0.98)***
Mitral regurgitation	*0.67 (0.57-0.71)*	*0.86 (0.82-0.90)*	***0.91 (0.89-0.93)***	*0.85 (0.79-0.90)*
TAVR	*0.52 (0.47-0.69)*	*0.95 (0.93-0.97)*	***0.97 (0.90-1.00)***	---
Bicuspid AV	*0.50 (0.47-0.54)*	*0.63 (0.53-0.73)*	***0.82 (0.73-0.90)***	*0.73 (0.64-0.82)*
Aortic stenosis	*0.52 (0.48-0.55)*	*0.81 (0.73-0.89)*	*0.96 (0.92-0.99)*	***0.97 (0.95-0.98)***
Aortic regurgitation	*0.53 (0.48-0.61)*	*0.67 (0.54-0.80)*	***0.89 (0.80-0.97)***	*0.83 (0.74-0.90)*
Tricuspid regurgitation	*0.61 (0.55-0.66)*	*0.80 (0.75-0.85)*	***0.88 (0.86-0.91)***	*0.85 (0.81-0.88)*
Pericardial effusion	*0.75 (0.60-0.66)*	*0.89 (0.78-0.98)*	***0.89 (0.77-0.98)***	*0.87 (0.73-0.96)*
Aortic root dilation	*0.52 (0.50-0.56)*	*0.71 (0.62-0.79)*	***0.97 (0.94-0.99)***	*0.80 (0.73-0.85)*
Dilated IVC	*0.46 (0.42-0.51)*	*0.75 (0.66-0.82)*	***0.86 (0.81-0.91)***	-
PA pressure (R^2)	*-0.25(-0.34- -0.17)*	*0.22 (0.13-0.30)*	***0.36 (0.30-0.42)***	*0.07 (0.00-0.14)*
PA pressure (MAE)	*10.92 (10.00-11.86)*	*8.80 (8.08-9.52)*	***7.97 (7.35-8.66)***	*9.12 (8.32-9.96)*

Supplementary Table 4 *Performance of foundation models for biomedical imaging on MIMIC dataset*

	MIMIC DATASET			
	BioMedCLIP	*EchoCLIP*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	*-2.48 (-3.26--1.85)*	*0.27 (0.13-0.40)*	***0.74 (0.66-0.80)***	*0.25 (0.20-0.30)*
LV ejection fraction (MAE)	*26.96 (25.31-28.58)*	*10.37 (9.30-11.51)*	***6.21 (5.54-6.92)***	*10.95 (9.91-12.01)*
Pacemaker	*0.53 (0.39-0.68)*	*0.79 (0.66-0.90)*	***0.86 (0.80-0.92)***	-
RV systolic function	*0.58 (0.02-0.90)*	*0.95 (0.89-0.99)*	***0.98 (0.97-0.99)***	*0.85 (0.78-0.93)*
RV dilation	-	-	-	-
LA dilation	*0.56 (0.52-0.59)*	*0.67 (0.64-0.71)*	***0.72 (0.68-0.75)***	*0.68 (0.65-0.71)*
RA dilation	*0.51 (0.47-0.54)*	*0.63 (0.59-0.67)*	*0.65 (0.61-0.69)*	***0.70 (0.66-0.73)***
Mitraclip	*0.53 (0.29-0.99)*	*0.96 (0.86-1.00)*	***0.99 (0.99-1.00)***	-
Mitral annular calcification	*0.56 (0.51-0.60)*	*0.81 (0.77-0.85)*	***0.88 (0.85-0.91)***	-
Mitral stenosis	*0.65 (0.48-0.98)*	*0.93 (0.80-1.00)*	***0.96 (0.95-0.98)***	*0.85 (0.66-0.99)*
Mitral regurgitation	*0.55 (0.52-0.58)*	*0.74 (0.71-0.76)*	***0.81 (0.79-0.83)***	*0.73 (0.71-0.76)*
TAVR	*0.52 (0.44-0.60)*	*0.85 (0.79-0.90)*	***0.96 (0.94-0.98)***	-
Bicuspid AV	*0.49 (0.46-0.56)*	*0.51 (0.42-0.61)*	*0.65 (0.52-0.78)*	***0.66 (0.50-0.81)***
Aortic stenosis	*0.51 (0.48-0.54)*	*0.68 (0.64-0.73)*	***0.88 (0.84-0.91)***	*0.80 (0.77-0.84)*
Aortic regurgitation	*0.55 (0.49-0.61)*	*0.61 (0.54-0.67)*	***0.81 (0.75-0.86)***	*0.69 (0.63-0.76)*
Tricuspid regurgitation	*0.55 (0.52-0.58)*	*0.76 (0.73-0.78)*	***0.84 (0.82-0.86)***	*0.76 (0.73-0.78)*
Pericardial effusion	*0.66 (0.61-0.72)*	*0.76 (0.70-0.81)*	***0.87 (0.83-0.92)***	*0.80 (0.75-0.86)*
Aortic root dilation	*0.41 (0.40-0.42)*	*0.40 (0.39-0.40)*	***0.98 (0.96-0.99)***	*0.60 (0.29-0.91)*
Dilated IVC	*0.53 (0.44-0.63)*	*0.52 (0.40-0.64)*	***0.59 (0.49-0.69)***	-
PA pressure (R^2)	-	-	-	-
PA pressure (MAE)	-	-	-	-

Supplementary Table 5 *Performance of foundation models for biomedical imaging on CGMH dataset*

	CGMH DATASET			
	BioMedCLIP	*EchoCLIP*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	*-8.31 (-9.74--7.06)*	*0.02 (-0.13-0.15)*	***0.51 (0.44-0.57)***	*0.12 (0.04-0.19)*
LV ejection fraction (MAE)	*32.64 (31.89-33.38)*	*9.15 (8.75-9.54)*	***6.46 (6.19-6.74)***	*9.02 (8.69-9.35)*
Pacemaker	*0.67 (0.56-0.80)*	***0.78 (0.59-0.92)***	***0.78 (0.63-0.92)***	-
RV systolic function	*0.57 (0.41-0.72)*	*0.89 (0.84-0.94)*	*0.90 (0.79-0.98)*	***0.93 (0.86-0.98)***
RV dilation	*0.43 (0.30-0.61)*	*0.84 (0.65-0.97)*	***0.87 (0.69-0.97)***	*0.69 (0.50-0.87)*
LA dilation	*0.50 (0.47-0.53)*	*0.63 (0.60-0.65)*	*0.65 (0.62-0.67)*	***0.72 (0.69-0.76)***
RA dilation	*0.56 (0.47-0.64)*	*0.74 (0.65-0.83)*	*0.75 (0.67-0.84)*	***0.87 (0.81-0.92)***
Mitraclip	-	-	-	-
Mitral annular calcification	*0.55 (0.40-0.70)*	*0.85 (0.69-0.97)*	***0.88 (0.75-0.96)***	-
Mitral stenosis	*0.61 (0.48-0.79)*	*0.71 (0.51-0.92)*	*0.79 (0.59-0.98)*	***0.88 (0.79-0.95)***
Mitral regurgitation	*0.56 (0.50-0.63)*	*0.81 (0.76-0.86)*	***0.86 (0.82-0.89)***	*0.76 (0.70-0.82)*
TAVR	*0.63 (0.50-0.77)*	*0.91 (0.83-0.97)*	***1.00 (1.00-1.00)***	-
Bicuspid AV	*0.66 (0.43-1.00)*	*0.48 (0.35-0.77)*	***0.67 (0.45-1.00)***	*0.47 (0.20-0.80)*
Aortic stenosis	*0.56 (0.43-0.71)*	*0.74 (0.59-0.88)*	***0.99 (0.98-1.00)***	*0.92 (0.86-0.97)*
Aortic regurgitation	*0.58 (0.50-0.66)*	*0.61 (0.53-0.69)*	***0.78 (0.70-0.84)***	*0.77 (0.70-0.83)*
Tricuspid regurgitation	*0.55 (0.49-0.61)*	*0.87 (0.83-0.91)*	***0.93 (0.90-0.95)***	*0.88 (0.84-0.91)*
Pericardial effusion	*0.35 (0.16-0.65)*	*0.90 (0.69-1.00)*	*0.96 (0.91-1.00)*	***0.99 (0.98-1.00)***
Aortic root dilation	*0.46 (0.39-0.62)*	*0.64 (0.41-0.86)*	***0.98 (0.97-1.00)***	*0.72 (0.58-0.84)*
Dilated IVC	*0.38 (0.37-0.39)*	*0.83 (0.50-1.00)*	***0.98 (0.94-1.00)***	-
PA pressure (R^2)	-	-	-	-
PA pressure (MAE)	-	-	-	-

Supplementary Table 6 *Performance of foundation models for biomedical imaging on Kaiser-Permanente dataset*

	KP DATASET			
	BioMedCLIP	*EchoCLIP*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	*-8.31 (-9.74--7.09)*	*0.09 (-0.04-0.21)*	***0.66 (0.62-0.71)***	*0.28 (0.23-0.33)*
LV ejection fraction (MAE)	*29.67 (29.09-30.25)*	*7.26 (6.89-7.64)*	***4.51 (4.30-4.72)***	*6.66 (6.37-6.97)*
Pacemaker	*0.57 (0.52-0.61)*	***0.85 (0.81-0.88)***	*0.84 (0.81-0.87)*	---
RV systolic function	*0.62 (0.55-0.68)*	*0.85 (0.80-0.90)*	***0.92 (0.88-0.95)***	*0.86 (0.82-0.90)*
RV dilation	*0.56 (0.51-0.62)*	***0.78 (0.72-0.84)***	***0.78 (0.73-0.83)***	*0.77 (0.72-0.82)*
LA dilation	*0.61 (0.59-0.63)*	*0.72 (0.70-0.74)*	***0.77 (0.75-0.79)***	*0.75 (0.73-0.77)*
RA dilation	*0.61 (0.57-0.64)*	*0.74 (0.71-0.77)*	*0.82 (0.79-0.84)*	***0.84 (0.82-0.87)***
Mitraclip	*0.61 (0.47-0.74)*	*0.95 (0.84-1.00)*	***1.00 (1.00-1.00)***	-
Mitral annular calcification	*0.58 (0.55-0.62)*	*0.84 (0.81-0.87)*	***0.89 (0.86-0.91)***	-
Mitral stenosis	*0.53 (0.49-0.61)*	*0.85 (0.74-0.95)*	*0.88 (0.78-0.95)*	***0.92 (0.87-0.96)***
Mitral regurgitation	*0.62 (0.58-0.66)*	*0.85 (0.81-0.88)*	***0.89 (0.86-0.91)***	*0.76 (0.72-0.80)*
TAVR	*0.55 (0.51-0.60)*	*0.85 (0.80-0.89)*	***0.99 (0.98-0.99)***	---

Bicuspid AV	*0.49 (0.49-0.49)*	*0.69 (0.60-0.77)*	***0.80 (0.72-0.88)***	*0.72 (0.63-0.79)*
Aortic stenosis	*0.50 (0.49-0.51)*	*0.73 (0.69-0.76)*	***0.97 (0.96-0.98)***	*0.87 (0.85-0.90)*
Aortic regurgitation	*0.49 (0.48-0.51)*	*0.74 (0.68-0.80)*	***0.87 (0.83-0.92)***	*0.73 (0.68-0.78)*
Tricuspid regurgitation	*0.64 (0.60-0.68)*	*0.87 (0.83-0.90)*	***0.91 (0.89-0.93)***	*0.87 (0.84-0.90)*
Pericardial effusion	*0.84 (0.69-0.95)*	*0.93 (0.78-1.00)*	***1.00 (0.99-1.00)***	*0.98 (0.94-1.00)*
Aortic root dilation	*0.50 (0.48-0.52)*	*0.67 (0.63-0.72)*	***0.81 (0.78-0.85)***	*0.58 (0.54-0.63)*
Dilated IVC	*0.50 (0.49-0.52)*	*0.73 (0.70-0.76)*	***0.80 (0.77-0.82)***	*-*
PA pressure (R^2)	*-0.20 (-0.25--0.14)*	*0.17 (0.12-0.21)*	***0.31 (0.27-0.35)***	*0.16 (0.13-0.19)*
PA pressure (MAE)	*9.46 (9.05-9.89)*	*8.26 (7.92-8.60)*	***7.39 (7.08-7.70)***	*8.07 (7.73-8.41)*

Supplementary Table 9 *EchoPrime performance in low-quality resource-constrained setting*

	FULL TEST		***LOW-QUALITY***		***SINGLE-VIEW***	
	EchoPrime	*PanEcho*	*EchoPrime*	*PanEcho*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	***0.83 (0.81-0.85)***	*0.52 (0.50-0.54)*	***0.80 (0.75-0.84)***	*0.42 (0.37-0.46)*	***0.78 (0.75-0.80)***	*0.34 (0.29-0.38)*
LV ejection fraction (MAE)	***4.79 (4.60-4.97)***	*7.46 (7.16-7.78)*	***5.30 (4.86-5.76)***	*8.98 (8.22-9.77)*	***5.22 (5.02-5.43)***	*8.49 (8.12-8.87)*
Pacemaker	***0.85 (0.82-0.88)***	*---*	***0.89 (0.84-0.93)***	*---*	***0.83 (0.80-0.85)***	*---*
RV systolic function	***0.93 (0.91-0.95)***	*0.89 (0.87-0.92)*	***0.91 (0.85-0.96)***	*0.85 (0.77-0.92)*	***0.92 (0.89-0.94)***	*0.83 (0.79-0.86)*
RV dilation	***0.89 (0.83-0.95)***	*0.86 (0.79-0.91)*	*0.85 (0.72-0.96)*	***0.88 (0.78-0.95)***	***0.92 (0.87-0.95)***	*0.76 (0.68-0.83)*
LA dilation	***0.91 (0.86-0.96)***	*0.89 (0.86-0.92)*	***0.95 (0.89-0.99)***	*0.88 (0.77-0.97)*	***0.92 (0.87-0.96)***	*0.80 (0.75-0.85)*
RA dilation	*0.90 (0.82-0.97)*	***0.93 (0.89-0.96)***	***0.89 (0.70-1.00)***	*0.85 (0.75-0.95)*	***0.90 (0.82-0.97)***	*0.88 (0.80-0.93)*
Mitraclip	***0.99 (0.99-1.00)***	*---*	***1.00 (0.99-1.00)***	*---*	***0.97 (0.94-0.99)***	*---*
Mitral annular calcification	***0.96 (0.95-0.97)***	*---*	***0.97 (0.94-0.98)***	*---*	***0.94 (0.91-0.96)***	*---*
Mitral stenosis	***0.96 (0.91-0.99)***	***0.96 (0.93-0.98)***	*0.97 (0.90-1.00)*	***0.99 (0.97-1.00)***	***0.94 (0.87-0.98)***	*0.79 (0.69-0.87)*
Mitral regurgitation	***0.92 (0.91-0.94)***	*0.86 (0.84-0.89)*	***0.91 (0.84-0.96)***	*0.85 (0.75-0.93)*	***0.89 (0.87-0.91)***	*0.77 (0.74-0.81)*
TAVR	***1.00 (0.99-1.00)***	*---*	***1.00 (0.99-1.00)***	*---*	***0.90 (0.87-0.93)***	*---*
Bicuspid AV	*0.83 (0.67-0.96)*	***0.95 (0.89-0.98)***	*0.72 (0.44-1.00)*	***0.84 (0.66-1.00)***	***0.78 (0.62-0.92)***	*0.76 (0.65-0.86)*
Aortic stenosis	***0.98 (0.96-0.99)***	*0.93 (0.90-0.95)*	***0.93 (0.83-1.00)***	*0.92 (0.81-0.98)*	*0.80 (0.75-0.85)*	***0.86 (0.82-0.90)***
Aortic regurgitation	***0.88 (0.83-0.93)***	*0.77 (0.71-0.83)*	***0.78 (0.38-0.99)***	*0.73 (0.44-0.97)*	***0.75 (0.68-0.82)***	*0.73 (0.66-0.80)*
Tricuspid regurgitation	***0.95 (0.93-0.97)***	*0.91 (0.88-0.94)*	***0.98 (0.97-0.99)***	*0.95 (0.90-0.99)*	***0.92 (0.88-0.94)***	*0.84 (0.80-0.88)*
Pericardial effusion	***0.98 (0.95-0.99)***	*0.93 (0.89-0.96)*	***0.94 (0.91-0.97)***	*0.92 (0.83-0.98)*	***0.93 (0.89-0.97)***	*0.80 (0.73-0.86)*
Aortic root dilation	***0.91 (0.83-0.98)***	*0.66 (0.56-0.76)*	***0.96 (0.91-1.00)***	*0.51 (0.23-0.74)*	***0.81 (0.70-0.91)***	*0.67 (0.54-0.79)*
Dilated IVC	***0.84 (0.82-0.86)***	*---*	***0.82 (0.77-0.87)***	*---*	***0.80 (0.78-0.83)***	*---*
PA pressure (R^2)	***0.43 (0.39-0.47)***	*0.27 (0.23-0.30)*	***0.32 (0.17-0.44)***	*0.16 (0.05-0.25)*	***0.36 (0.31-0.40)***	*0.03 (-0.04-0.10)*
PA pressure (MAE)	***7.30 (6.95-7.67)***	*8.27 (7.88-8.68)*	***7.76 (6.96-8.63)***	*8.91 (8.05-9.81)*	***7.67 (7.29-8.06)***	*9.55 (9.10-10.03)*

2. Free-text generation as an additive value:
Current echo workflows already use structured "pre-text" templates into which sonographers or fellows drop quantitative values. The manuscript argues that only ~¼–⅓ of a report is numeric and the remainder is narrative, but the added clinical utility of automatically generated prose remains hypothetical:
– Could you quantify editing time, keystrokes or click reduction for EchoPrime-drafted reports versus a standard template populated with EchoPrime (or EchoNet-Measurements) numbers? Even a small pilot (e.g., first 100 cases) would help substantiate the claim.

REPLY: Thank you for your careful review. We anticipate conducting a prospective clinical trial of EchoPrime shortly to directly assess prospective performance and acceptability to cardiologists that are reviewing the echocardiogram images. We have already acquired funding and institutional support to conduct this prospective evaluation, which potentially speaks to our internal understanding how it would compare with current standard of care. In this trial setting, we will randomize cardiologists to preliminary assessments of echocardiograms by EchoPrime vs. preliminary assessments performed by cardiologists and sonographers and track the rate of change as cardiologist's conducted blinded review. For this manuscript, we do not have this prospective evaluation and in discussion with the Nature editors, it was felt this was out of scope of this current paper.

We recognize the importance of prospective evaluation, though such studies require substantial time and financial resources. We clarify that this will be tested in a clinical trial:

In the Discussion:

> *Although EchoPrime was evaluated at five geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in the real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime preliminary assessment. We currently are working on a funded, prospective, blinded clinical trial of EchoPrime to assess it's impact on clinical practice. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate many steps in echocardiographic imaging workflow.*

– Please clarify what safeguards exist against hallucination. Given the acknowledged difficulty of validating more than 1 000 narrative phrases, how will users be alerted if the model outputs text that is not factually supported? For example, could common phenotypic descriptors—such as "granular sparkling," "features suggestive of ARVC," or "hyper-trabeculation" appear in the report? If so, what mechanisms ensure that such statements are appropriate for the current study and not hallucinated? Are these phrases constrained by explicit rule sets or diagnostic-criteria checks, or subjected to expert review before sign-off? Please provide any phrase-level validation metrics (e.g., precision and recall against blinded cardiologist adjudication) and explain how the interface flags sentences that lack sufficient evidentiary support. Clarifying these safeguards would help demonstrate the added value of EchoPrime's free-text output over a standard template.

REPLY: Thank you for your careful review. We consider the use of RAI and anatomic attention to be critical in avoiding hallucinations. The use of comparison with 50 closest reports ensures only consensus likely reporting is resulted from EchoPrime and we have not had any demonstrated problems with significant hallucination.

The use of anatomic attention also ensures the model prioritizes videos where the relevant cardiac structure is present so there is not assessment of the RV in the A2C view for example. Put in another way, hallucination is just incorrect output, and would show in decreased performance when there is insufficient information in the videos. The PanEcho paper actually show this phenomena in Figure 4, where videos that do not contain the relevant structure results in much worse performance (for example the LV function is not seen in the subcostal view, so results in much worse performance). By being 'view agnostic', PanEcho actually incorporates these hallucinated results in the model and averages incorrect predictions, while EchoPrime uses anatomic attention and avoids using the subcostal view in this example.

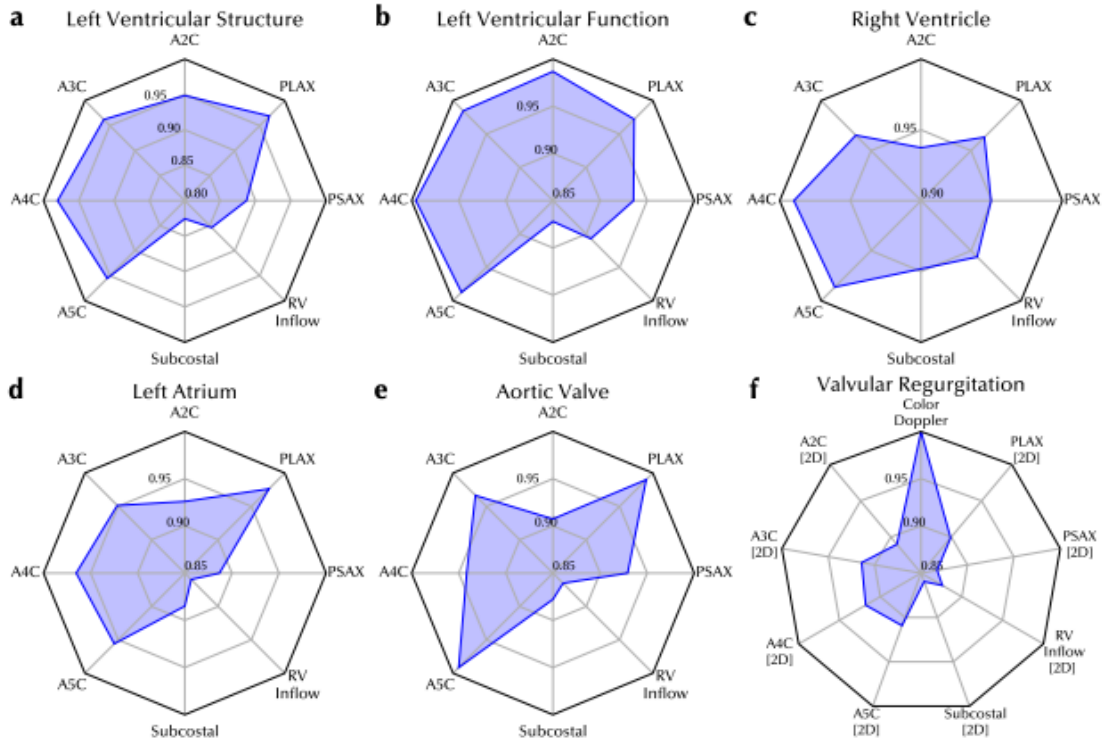


Fig. 4 | Task-specific view relevance. Radar plots depicting the relative importance of each echocardiographic view for a given aspect of cardiovascular diagnosis. For each task, a normalized view relevance score is computed, where 1 indicates the most relevant view; presented view relevance scores are then averaged over all tasks falling under a given category in each subplot. Analysis is performed on the internal YNHHS test set using up to three videos from a given echocardiographic view per study.

3. I acknowledge the planned randomised clinical trial slated to read out in 2026. Until at least preliminary results are available, I would encourage tempering workflow-efficiency claims or explicitly positioning them as future work.

REPLY: Thank you for your careful review. We recognize the importance of prospective evaluation, though such studies require substantial time and financial resources. We clarify that this will be tested in a clinical trial and explicitly position this as future work.

In the Discussion:

> *Although EchoPrime was evaluated at five geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in the real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime preliminary assessment. We currently are working on a funded, prospective, blinded clinical trial of EchoPrime to assess it's impact on clinical practice. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate many steps in echocardiographic imaging workflow.*

In short, while the revision addresses several of my prior measurement requests, the new Bland-Altman plots suggest that non-trivial proportional bias and heteroscedasticity persist in key regression tasks. Consequently, both the calibration of the continuous outputs and the distinctive advantages of the free-text generation still need empirical validation. Demonstrating either (a) superior diagnostic performance relative to PanEcho, or (b) proven reductions in reporting time/error through prospective or pilot data, would strengthen the manuscript.

REPLY: Thank you for your careful review. Since the last revision, we provide benchmark comparison across each of our test datasets (both external and internal). In these experiments, we show EchoPrime has comparable or better performance across most tasks compared to PanEcho. We acknowledge it would be ideal to have prospective data, which we anticipate having in 2026.

Referee #2 (Remarks to the Author):

In this second revision to your foundational model of echocardiography interpretation, the authors have addressed many of my concerns about the manuscript. Overall, the manuscript does a really nice job of translating a key principle of AI research into the clinical domain: namely that some high dimensional space can be mapped to a low dimension (learned) manifold. In this case, an echocardiographic study (consisting of many videos each with many pixels, i.e. very many dimensions!) can be mapped to a lower vector space that can be clustered. You further take this one step further by doing report retrieval to basically translate that vector space back into clinician speak (i.e. embedding vector space into clinical natural language).

You better acknowledge the fundamental limitation that this current approach does not utilize any standard echo measurements from still images (Doppler, M-mode, or 2D still images). A pitfall of doing this is that any retrieved measurements are basically just guesses by the foundational algorithm. Patients that are similar to "xyz" will generally have measurements close to a certain distribution. However, this "guessing" is really only as good as the model's ability to learn given the data at hand (which lacks actual measurements).

REPLY: Thank you for your careful review. We consider the use of RAI and anatomic attention to be critical in avoiding hallucinations. The use of comparison with 50 closest reports ensures only consensus likely reporting is resulted from EchoPrime and we have not had any demonstrated problems with significant hallucination or guessing without actual measurements at hand. Actually, to clarify, that is also the approach for PanEcho as that is not a segmentation or measurement model, however without the benefit of RAI or anatomic attention.

The use of anatomic attention also ensures the model prioritizes videos where the relevant cardiac structure is present so there is not assessment of the RV in the A2C view for example. Put in another way, hallucination is just incorrect output, and would show in decreased performance when there is insufficient information in the videos. The PanEcho paper actually show this phenomena in Figure 4, where videos that do not contain the relevant structure results in much worse performance (for example the LV function is not seen in the subcostal view, so results in much worse performance). By being 'view agnostic', PanEcho actually incorporates these hallucinated results in the model and averages incorrect predictions, while EchoPrime uses anatomic attention and avoids using the subcostal view in this example.

Remarkably, in your rebuttal to me (but not in the manuscript), you note an on-line pre-print "EchoNet-Measurements", https://www.medrxiv.org/content/10.1101/2025.03.18.25324215v1, in which you appear to overcome this limitation. You demonstrate accuracy in a host of 1-D, 2-D, and Doppler measurements similar to Tromp et al (Lancet Dig Health 2022). Unfortunately, this capability is not integrated into Echo-Prime in the current manuscript, which thus remains a limitation to the current report. There is no question that the authors have developed a significant advance in AI interpretation of echocardiograms, and with the multiple revisions and broader validation, this has become a high impact publication. However, the lack of integration of quantitative measurements into the echo reports remains a significant limitation. I fully appreciate the challenge of performing this in the context of the current manuscript, given word count limitations and the divergent methods of the two approaches. I would be supportive of the current approach alone, but you should emphasize more the need to integrate measurements into the automated reports in the future. If it is your policy to cite preprint publications, I would suggest doing that to emphasize the importance of this capability. I suspect it won't be long till we see a manuscript from your group that will combine quantitative measurements with this video-encoding foundational model, which will be a truly remarkable achievement.

REPLY: Thank you for your careful review. We do think measurements are very important and a critical part of reporting, it just happens to be a complementary and not duplicative approach to EchoPrime. In this revision, we provide head to head comparison with PanEcho to show that EchoPrime is likely the best echo specific model, supported by more comprehensive validation across more validation sites and a larger training corpus.

A comparison of key echo models are as follows:

EchoPrime	**PanEcho**	**EchoNet-Measurements**
Contrastive Vision-Language Model, both video encoder CNN and text encoder (BERT)	Supervised Training, Convolutional Neural Network	Supervised Training, Convolutional Neural Network
Outputs Natural Language Text Reports	Outputs classifications and numbers	Outputs segmentation annotations and associated numbers
Can do all echo text tasks as it generates natural language report language	Performs 39 specific tasks	Performs 18 specific tasks
Trained on 12M videos	Trained on 1.2M videos	Trained on 887k videos
View informed inference	Not view informed inference	single view inference
External validation at 4 geographically independent	External validation on 2 public single view datasets (EchoNet-Dynamic, EchoNet-LVH) and	External validation at 1 geographically independent site

sites with comprehensive echo videos from full studies	one geographically distinct site that include the full echocardiogram studies (RVENet).	

A big part of our Revision 2 response was highlighting that text generation is very different from the regression tasks of numerical result generation, and in many ways this is complementary (see next figure highlighting where text and where numerical outputs are in different places and both needed in a standard echocardiography report).

As another reference point to how cardiologists understand these to be distinct and unique tasks, last week we found out both EchoPrime and EchoNet-Measurements are headline presentations for the September 2025 American Society of Echocardiography annual meeting in the Late Breaking Science session. This session presents arguably the best science in echocardiography in the past year and the organizers (generally all clinical cardiologists) felt comfortable with the presentation of EchoPrime and EchoNet-Measurements side-by-side in the same session. EchoNet-Measurements will also be a simultaneous publication in JACC which highlight the complementary comparison of these two models.

They are not duplicative, and a big part of our R2 response was also showing with evidence that measurements do not naturally flow into good text categories (because they are not the same thing and there is variability in echocardiography). Furthermore, EchoNet-Measurement are traditional CNNs that are performing a segmentation task while EchoPrime is trained on 15x more data and is a foundation vision-language model. EchoPrime has significant technical novelty that we think is unparalleled either by PanEcho or EchoNet-Measurements.

Other than this broad observation, I believe you have adequately addressed my concerns from the prior revision.

Referee #3 (Remarks to the Author):

I have carefully reviewed the comments from all reviewers as well as the authors' rebuttal. The absence of prospective clinical validation substantially limits the clinical credibility of EchoPrime. Without real-world data demonstrating clinical efficacy and physician acceptance, the model's practical utility remains speculative.

Moreover, the manuscript lacks empirical evidence of clinician interaction with EchoPrime. Demonstrating actual clinician engagement, workflow integration, and efficiency gains is essential, especially for a tool intended to function in real clinical settings. Even preliminary user studies or pilot trials would considerably strengthen the authors' claims.

REPLY: Thank you for your careful review. We anticipate conducting a prospective clinical trial of EchoPrime shortly to directly assess prospective performance and acceptability to cardiologists that are reviewing the echocardiogram images. We have already acquired funding and institutional support to conduct this prospective evaluation, which potentially speaks to our internal understanding how it would compare with current standard of care. In this trial setting, we will randomize cardiologists to preliminary assessments of echocardiograms by EchoPrime vs. preliminary assessments performed by cardiologists and sonographers and track the rate of change as cardiologist's conducted blinded review. For this manuscript, we do not have this prospective evaluation and in discussion with the Nature editors, it was felt this was out of scope of this current paper.

We recognize the importance of prospective evaluation, though such studies require substantial time and financial resources. We clarify that this will be tested in a clinical trial:

In the Discussion:

> *Although EchoPrime was evaluated at five geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in the real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime preliminary assessment. We currently are working on a funded, prospective, blinded clinical trial of EchoPrime to assess it's impact on clinical practice. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows. Point-of-care ultrasound is one high-value use case such that AI can provide rapid expert evaluation of frontline diagnostic information. Combined with AI-based probe guidance[33,34], EchoPrime has the potential to automate many steps in echocardiographic imaging workflow.*

More critically, the proposed vision-language model is not built upon any

established large or small language model backbone (Llama 3.2, InternVL, QWen, etc). Given the absence of general-purpose language pretraining, its generative language capabilities are inherently constrained. This limitation raises concerns regarding its ability to generalize across diverse reporting tasks. Furthermore, the model's performance should be directly compared with recent state-of-the-art alternatives, such as GPT-4o, Gemini 2.5 Pro, and domain-specific models like MedGemma, Llava-med. In light of these considerations, I no longer believe the proposed AI approach represents the best available solution for this task.

REPLY: Thank you for your careful review. EchoPrime uses a video encoder 9mVIT) pretrained on Kinetics and a pretrained BERT text encoder (from huggingface and pretrained on BookCorpus and English Wikipedia). Furthermore, we benchmark EchoPrime on both general purpose medical and biomedical CLIP models as well as echo models (EchoCLIP and PanEcho) and show superior performance by EchoPrime.

We note some of the models suggested by the reviewer are closed source. Unfortunatley, our data use agreement does not allow uploading to 3rd party private servers, but our general experience is such models do quite poorly for echo interpretation tasks as they are OOD and not typical. In fact, we actually show these comparisons using Qwen and finetuning on general video tasks and show they perform poorly on echocardiography (https://arxiv.org/abs/2504.14391). These experiments in combination as well as results from PubMedCLIP, PMC-CLIP, and BioMedCLIP show general purpose models do not perform well on the tasks described.

Supplementary Table 2 *Performance of foundation models for biomedical imaging on Cedars-Sinai test set*

CEDARS-SINAI TEST SET						
	PubMedCLIP	*PMC-CLIP*	*BioMedCLIP*	*EchoCLIP*	*EchoPrime*	*PanEcho*
LV ejection fraction (R^2)	*-0.32 (-0.40--0.25)*	*-0.38 (-0.46--0.31)*	*-2.60 (-3.00--2.20)*	*0.62 (0.59-0.65)*	***0.83 (0.81-0.85)***	*0.52 (0.50-0.54)*
LV ejection fraction (MAE)	*14.7 (14.3-15.1)*	*14.0 (13.6-14.5)*	*26.93 (26.4-27.4)*	*7.00 (6.7-7.3)*	***4.8 (4.6-5.0)***	*7.46 (7.16-7.78)*
Pacemaker	*0.50 (0.50-0.50)*	*0.51 (0.47-0.54)*	*0.58 (0.55-0.62)*	***0.88 (0.85-0.91)***	*0.85 (0.82-0.88)*	*-*
RV systolic function	*0.50 (0.49-0.51)*	*0.53 (0.48-0.58)*	*0.64 (0.59-0.69)*	*0.89 (0.86-0.92)*	***0.93 (0.91-0.95)***	*0.89 (0.87-0.92)*
RV dilation	*0.50 (0.45-0.56)*	*0.59 (0.50-0.67)*	*0.58 (0.52-0.64)*	*0.88 (0.83-0.92)*	***0.89 (0.83-0.95)***	*0.86 (0.79-0.91)*
LA dilation	*0.50 (0.50-0.50)*	*0.57 (0.48-0.65)*	*0.66 (0.58-0.74)*	***0.93 (0.89-0.96)***	*0.91 (0.86-0.96)*	*0.89 (0.86-0.92)*
RA dilation	*0.50 (0.50-0.50)*	*0.58 (0.47-0.68)*	*0.67 (0.59-0.75)*	*0.90 (0.82-0.97)*	*0.90 (0.82-0.97)*	***0.93 (0.89-0.96)***
Mitraclip	*0.50 (0.50-0.50)*	*0.49 (0.43-0.55)*	*0.70 (0.63-0.76)*	*0.95 (0.92-1.00)*	***0.99 (0.99-1.00)***	*-*
Mitral annular calcification	*0.50 (0.46-0.55)*	*0.59 (0.54-0.64)*	*0.63 (0.58-0.68)*	*0.95 (0.93-0.96)*	***0.96 (0.95-0.97)***	*-*
Mitral stenosis	*0.50 (0.50-0.50)*	*0.55 (0.44-0.66)*	*0.70 (0.61-0.79)*	***0.96 (0.91-0.99)***	***0.96 (0.91-0.99)***	***0.96 (0.93-0.98)***
Mitral regurgitation	*0.50 (0.50-0.50)*	*0.59 (0.55-0.63)*	*0.66 (0.63-0.69)*	*0.89 (0.87-0.91)*	***0.92 (0.91-0.94)***	*0.86 (0.84-0.89)*
TAVR	*0.50 (0.50-0.51)*	*0.51 (0.46-0.55)*	*0.54 (0.50-0.57)*	*0.86 (0.83-0.89)*	***1.00 (0.99-1.00)***	*-*
Bicuspid AV	*0.50 (0.50-0.50)*	*0.51 (0.47-0.60)*	*0.52 (0.51-0.53)*	*0.72 (0.58-0.85)*	*0.83 (0.67-0.96)*	***0.95 (0.89-0.98)***
Aortic stenosis	*0.50 (0.50-0.52)*	*0.48 (0.42-0.55)*	*0.55 (0.53-0.57)*	*0.83 (0.79-0.87)*	***0.98 (0.96-0.99)***	*0.93 (0.90-0.95)*
Aortic regurgitation	*0.50 (0.50-0.50)*	*0.50 (0.41-0.58)*	*0.54 (0.53-0.55)*	*0.68 (0.60-0.76)*	***0.88 (0.83-0.93)***	*0.77 (0.71-0.83)*
Tricuspid regurgitation	*0.49 (0.45-0.53)*	*0.57 (0.51-0.63)*	*0.69 (0.64-0.74)*	*0.91(0.89-0.93)*	***0.95 (0.93-0.97)***	*0.91 (0.88-0.94)*
Pericardial effusion	*0.54 (0.45-0.62)*	*0.50 (0.42-0.57)*	*0.68 (0.59-0.77)*	*0.92 (0.89-0.95)*	***0.98 (0.95-0.99)***	*0.93 (0.89-0.96)*
Aortic root dilation	*0.50 (0.50-0.50)*	*0.44 (0.35-0.55)*	*0.50 (0.49-0.51)*	*0.76 (0.63-0.88)*	***0.91 (0.83-0.98)***	*0.66 (0.56-0.76)*
Dilated IVC	*0.50 (0.50-0.51)*	*0.53 (0.50-0.56)*	*0.49 (0.47-0.50)*	*0.81 (0.79-0.83)*	***0.84 (0.82-0.86)***	*-*
PA pressure (R^2)	*-0.59 (-0.74--0.46)*	*-0.83 (-1.07--0.62)*	*-0.19 (-0.24 - -0.14)*	*0.35 (0.29-0.41)*	***0.43 (0.39-0.47)***	*0.27 (0.23-0.30)*
PA pressure (MAE)	*11.3 (10.0-12.6)*	*13.7 (13.0-14.3)*	*10.6 (10.1-11.12*	*7.8 (7.5- 8.2)*	***7.3 (7.0-7.7)***	*8.27 (7.88-8.68)*

In summary, the technical approach underlying the AI method for the general healthcare domain lacks novelty and is, in fact, somewhat outdated at this point. I do not believe that any further rebuttal can adequately address this major concern.

REPLY: Thank you for your careful review. We respectfully disagree as there is limited evidence general purpose AI models do well on these tasks and our benchmarking for echo specific models show that our model has the best performance of all publicly available echo models. Our empiric results show that EchoPrime is likely the best echo specific model based on head to head comparison as well as training and validation characteristics.

A comparison of key echo models are as follows:

EchoPrime	**PanEcho**	**EchoNet-Measurements**
Contrastive Vision-Language Model, both video encoder CNN and text encoder (BERT)	Supervised Training, Convolutional Neural Network	Supervised Training, Convolutional Neural Network
Outputs Natural Language Text Reports	Outputs classifications and numbers	Outputs segmentation annotations and associated numbers
Can do all echo text tasks as it generates natural language report language	Performs 39 specific tasks	Performs 18 specific tasks
Trained on 12M videos	Trained on 1.2M videos	Trained on 887k videos
View informed inference	Not view informed inference	single view inference
External validation at 4 geographically independent sites with comprehensive echo videos from full studies	External validation on 2 public single view datasets (EchoNet-Dynamic, EchoNet-LVH) and one geographically distinct site that include the full echocardiogram studies (RVENet).	External validation at 1 geographically independent site

Referee #3 (Remarks on code availability):

It is not well-maintained and has not been updated for an extended period of time.

REPLY: Thank you for your careful review. The codebase was released in October 2024 with the initial preprint. Reviewers 1 and Reviewers 2 were able to use the codebase successfully with comparisons on their own datasets. We anticipate updating once the manuscript is published, and changes are on hold prior to publication.

Referee #1 (Remarks to the Author):

This revision presents extensive benchmarking of EchoPrime across five institutional datasets, with consistent advantages in key regression tasks such as LVEF and RVSP. However, several foundational issues remain unresolved, and multiple original reviewer concerns, particularly regarding generalizability, calibration, and validation methodology, have not been adequately addressed in the rebuttal.

Reply: We thank the reviewers for their feedback and providing context in this review. EchoPrime is the only model evaluated in multiple external sites, trained on 10-100x data the comparators, and have been shown to be superior in head to head comparisons described below (both in internal and external validation cohorts).

First, the use of Cedars-Sinai as a validation site is methodologically problematic. From prior publications and model descriptions, it is clear that Cedars was one of EchoPrime's primary training sources. Yet the manuscript presents Cedars results as external validation, where EchoPrime also performs best—winning 13 of 15 tasks against PanEcho, including the largest performance margins in LVEF and RVSP estimation. This undermines the claim of generalizability. The authors must provide a detailed breakdown of which sites contributed to model training and which were held out entirely for testing, alongside a clear statement on whether any patient overlap occurred across datasets. Without this, readers cannot interpret the Cedars performance as independent.

Reply: In our revision, we show performance at Cedars, Kaiser, CMGH, MIMIC, and Stanford. The training dataset was constructed exclusively from Cedars, and all other sites (Stanford, MIMIC, CGMH and Kaiser Permanente) were completely withheld for external validation and show strong superior performance compared to previous models including PanEcho.

Among 59 evaluated metrics, 36 demonstrated statistically significant differences after Bonferroni correction ($p < 0.05 / 59$), clearly demonstrating a trend towards improved performance with EchoPrime.

Extended Data Table 3 Aggregated comparison of EchoPrime and PanEcho on all external validation sites. Asterisk (*) denote statistically significant differences ($p<0.05/59$ adjusted for multiplicity). The DeLong test was used for comparing AUCs, and bootstrapping was used for MAE and R2 metrics.

	STANFORD		MIMIC		CGMH		KP	
	EchoPrime	PanEcho	EchoPrime	PanEcho	EchoPrime	PanEcho	EchoPrime	PanEcho
LV ejection fraction (R^2)	**0.79 (0.76-0.82)***	0.39 (0.36-0.42)	**0.74 (0.66-0.80)***	0.25 (0.20-0.30)	**0.51 (0.44-0.57)***	0.12 (0.04-0.19)	**0.66 (0.62-0.71)***	0.28 (0.23-0.33)
LV ejection fraction (MAE)	**4.14 (3.94-4.34)***	6.41 (6.03-6.79)	**6.21 (5.54-6.92)***	10.95 (9.91-12.01)	**6.46 (6.19-6.74)***	9.02 (8.69-9.35)	**4.51 (4.30-4.72)***	6.66 (6.37-6.97)
Pacemaker	0.84 (0.81-0.87)	-	**0.86 (0.80-0.92)**	-	**0.78 (0.63-0.92)**	-	0.84 (0.81-0.87)	---
RV systolic function	**0.94 (0.88-1.00)**	0.93 (0.80-0.99)	**0.98 (0.97-0.99)***	0.85 (0.78-0.93)	0.90 (0.79-0.98)	**0.93 (0.86-0.98)**	**0.92 (0.88-0.95)***	0.86 (0.82-0.90)
RV dilation	0.85 (0.80-0.90)	**0.87 (0.83-0.90)**	-	-	**0.87 (0.69-0.97)***	0.69 (0.50-0.87)	**0.78 (0.73-0.83)**	0.77 (0.72-0.82)
LA dilation	0.73 (0.70-0.75)	**0.81 (0.79-0.83)***	**0.72 (0.68-0.75)***	0.68 (0.65-0.71)	0.65 (0.62-0.67)	**0.72 (0.69-0.76)***	**0.77 (0.75-0.79)***	0.75 (0.73-0.77)
RA dilation	0.77 (0.74-0.81)	**0.84 (0.81-0.87)***	0.65 (0.61-0.69)	**0.70 (0.66-0.73)***	0.75 (0.67-0.84)	**0.87 (0.81-0.92)***	0.82 (0.79-0.84)	**0.84 (0.82-0.87)**
Mitraclip	**0.98 (0.94-1.00)**	-	**0.99 (0.99-1.00)**	-	-	-	**1.00 (1.00-1.00)**	-
Mitral annular calcification	**0.96 (0.94-0.98)**	-	**0.88 (0.85-0.91)**	-	**0.88 (0.75-0.96)**	-	**0.89 (0.86-0.91)**	-
Mitral stenosis	0.92 (0.86-0.98)	**0.96 (0.94-0.98)**	**0.96 (0.95-0.98)**	0.85 (0.66-0.99)	0.79 (0.59-0.98)	**0.88 (0.79-0.95)**	0.88 (0.78-0.95)	**0.92 (0.87-0.96)**
Mitral regurgitation	**0.91 (0.89-0.93)**	0.85 (0.79-0.90)	**0.81 (0.79-0.83)***	0.73 (0.71-0.76)	**0.86 (0.82-0.89)***	0.76 (0.70-0.82)	**0.89 (0.86-0.91)***	0.76 (0.72-0.80)
TAVR	**0.97 (0.90-1.00)**	---	**0.96 (0.94-0.98)**	-	**1.00 (1.00-1.00)**	-	**0.99 (0.98-0.99)**	---
Bicuspid AV	**0.82 (0.73-0.90)***	0.73 (0.64-0.82)	0.65 (0.52-0.78)	**0.66 (0.50-0.81)**	**0.67 (0.45-1.00)**	0.47 (0.20-0.80)	**0.80 (0.72-0.88)***	0.72 (0.63-0.79)
Aortic stenosis	0.96 (0.92-0.99)	**0.97 (0.95-0.98)**	**0.88 (0.84-0.91)***	0.80 (0.77-0.84)	**0.99 (0.98-1.00)**	0.92 (0.86-0.97)	**0.97 (0.96-0.98)***	0.87 (0.85-0.90)
Aortic regurgitation	**0.89 (0.80-0.97)**	0.83 (0.74-0.90)	**0.81 (0.75-0.86)***	0.69 (0.63-0.76)	**0.78 (0.70-0.84)**	0.77 (0.70-0.83)	**0.87 (0.83-0.92)***	0.73 (0.68-0.78)
Tricuspid regurgitation	**0.88 (0.86-0.91)**	0.85 (0.81-0.88)	**0.84 (0.82-0.86)***	0.76 (0.73-0.78)	**0.93 (0.90-0.95)***	0.88 (0.84-0.91)	**0.91 (0.89-0.93)***	0.87 (0.84-0.90)
Pericardial effusion	**0.89 (0.77-0.98)**	0.87 (0.73-0.96)	**0.87 (0.83-0.92)***	0.80 (0.75-0.86)	0.96 (0.91-1.00)	**0.99 (0.98-1.00)**	**1.00 (0.99-1.00)**	0.98 (0.94-1.00)
Aortic root dilation	**0.97 (0.94-0.99)**	0.80 (0.73-0.85)	**0.98 (0.96-0.99)**	0.60 (0.29-0.91)	**0.98 (0.97-1.00)***	0.72 (0.58-0.84)	**0.81 (0.78-0.85)***	0.58 (0.54-0.63)
Dilated IVC	**0.86 (0.81-0.91)**	-	**0.59 (0.49-0.69)**	-	**0.98 (0.94-1.00)**	-	**0.80 (0.77-0.82)**	-
PA pressure (R^2)	**0.36 (0.30-0.42)***	0.07 (0.00-0.14)	-	-	-	-	**0.31 (0.27-0.35)***	0.16 (0.13-0.19)
PA pressure (MAE)	**7.97 (7.35-8.66)***	9.12 (8.32-9.96)	-	-	-	-	**7.39 (7.08-7.70)***	8.07 (7.73-8.41)

In this revision, we clarify that all training data comes from CSMC and other sites are external validation.

In the Methods,

> *For EchoPrime training, we curated a dataset of 12,124,168 2D and color Doppler echocardiography videos from 275,442 studies and corresponding medical reports from 108,913 patients collected between 2011 and 2022. DICOM images were queried from the clinical data storage system, converted to AVI video files, and de-identified before model training and inference. Data were split by patient into training, validation, and internal test datasets. The training data contained 11,984,170 videos derived from 272,256 studies across 107,663 patients from CSMC, and CSMC was the only data used to train the model.*

In the Methods, External Validation Datasets

> *For external validation, we obtained echocardiographic data from four geographically distinct international sites to evaluate EchoPrime's performance. From Stanford Healthcare, we collected 91,746 videos from 1,792 transthoracic echocardiography (TTE) studies across 1,792 patients. From Beth Israel Deaconess Medical Center (MIMIC dataset), we used 75,768 videos from 2,431 TTE studies with corresponding echocardiography reports across 1,767 patients. From CGMH Hospital in Taiwan, we acquired 24,724 videos with reports from 1,188 TTE studies, each from a unique patient. CGMH had 21 videos per study on average, notably smaller from the three US sites (31-51). From Kaiser-Permanente medical center, we collected 201,752 videos from 4891 studies obtain in September 2023. All video data from external datasets underwent the same preprocessing workflow, including RGB conversion, cropping to the ultrasound region, and resizing to 224×224. Study labels were extracted in quantitative form from reports using regex with institution-specific custom phrases. All data from external validation sites was used exclusively for validation and was not accessible to the model during training.*

Second, while EchoPrime outperforms PanEcho on many tasks, the manuscript lacks formal statistical comparison. No paired significance tests (such as DeLong for AUROC or bootstrap methods for R^2 and MAE) are presented. Given that PanEcho outperforms EchoPrime on specific chamber size and morphology tasks (e.g., LA dilation at Stanford 0.81 vs 0.73, RA dilation 0.84 vs 0.77, BAV at Cedars 0.95 vs 0.83), this absence weakens the comparative analysis. The authors should pre-specify a primary endpoint (e.g., LVEF MAE), conduct paired tests across all shared metrics, and consider reporting macro-averaged performance across tasks to support any claim of superiority.

Reply: We thank the reviewer for this feedback and in this revision, add p-values for all our comparisons with PanEcho to show superiority. Among 59 evaluated metrics, 36 demonstrated statistically significant differences after Bonferroni correction ($p < 0.05 / 59$), demonstrating clear superiority of EchoPrime in these comparisons. This was not in original manuscript, because we note that 1) PanEcho did not have weights available at the time of initial submission of our manuscript and 2)

these experiments were post-hoc requested by the reviewer. As such, it is not possible to pre-specify, however the overwhelming trend is superiority of EchoPrime.

Third, the calibration of EchoPrime's regression outputs remains unquantified. Referee #1 requested Bland–Altman analyses to assess proportional bias and heteroscedasticity, including slope estimates, statistical testing, and decile-wise limits of agreement. Despite this, the authors provide only qualitative plots and do not report slope ±95% CI, p-values, or recalibration experiments. Visual inspection suggests clear underestimation at higher values and increased error spread for several outputs, including RVSP and TR gradient. This demands quantification. At minimum, the authors should report slope and LoA by value decile for key continuous variables and provide recalibrated performance using simple linear or isotonic correction to evaluate the clinical relevance of these biases.

Reply: Thank you for the comment. As requested, we performed 33 additional Bland–Altman analyses covering 11 regression tasks across three medical centers (**Supplementary Figure 1**). Each plot includes the mean bias and corresponding 95% confidence intervals, allowing visualization of agreement and variability across value ranges. We believe these analyses sufficiently demonstrate EchoPrime's calibration and zero-shot performance on continuous prediction tasks.

Fourth, the manuscript's core claim that EchoPrime is "superior across most tasks" is overstated. The benchmarking data show mixed results: PanEcho matches or outperforms EchoPrime in several chamber size assessments, particularly at Stanford and CGMH, and RVSP is not reported at two of the five sites. A more accurate framing would acknowledge that EchoPrime performs comparably or better across most, but not all—tasks, with particularly strong gains in LVEF and RVSP, while PanEcho maintains advantages in certain structural phenotypes.

Reply: We appreciate this comment and agree that it is important to interpret comparative performance carefully. As shown in Extended Data Table 3, EchoPrime outperforms PanEcho on 11 of 16 tasks at Stanford, 11 of 13 at MIMIC, 9 of 14 at CGMH, and 14 of 16 at Kaiser Permanente. Among 59 evaluated metrics, 36 demonstrated statistically significant differences after Bonferroni correction ($p < 0.05 / 59$), with EchoPrime outperforming PanEcho in 31 of those 36 cases. EchoPrime's superior performance is even more evident in low quality and low view count images (Extended Data Table 4). We believe that this consistent pattern of improvement across four geographically and institutionally distinct cohorts provides strong evidence of generalizability. In addition, Referee #2 supported this conclusion, noting that "Overall, I have no doubt that EchoPrime is the superior model both in terms of performance and novelty."

Fifth, the narrative generation feature of EchoPrime remains speculative. Although technically promising, there is no validation of the generated free-text reports, no check of potential hallucinated phrases, and no quantitative data on editing burden or time saved compared to template-based workflows. Given the acknowledged risk of inappropriate descriptors (e.g., "granular sparkling," "hyper-trabeculation," "features suggestive of ARVC"), some form of validation is essential. This could take the form of a pilot study measuring keystrokes, edit time, or phrase-level precision and recall against expert adjudication. In the absence of such data, the claims regarding clinical utility and workflow efficiency should be clearly presented as future

directions, not as current strengths.

Reply: Thank you for your feedback. We clarify in the discussion that it is difficult to comprehensively test all outputs of a vision-language model, and further work is required. We have secured funding and institutional support to conduct a prospective trial.

In the discussion:

> *Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime assessment. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials and identifying optimal integration into clinical workflows.*

Finally, the discussion does not address the confounding impact of data scale. EchoPrime was trained on 12 million echo videos: ten times more than PanEcho. While this may partly explain the observed performance advantage, the scale differential introduces a bias that should be acknowledged when interpreting head-to-head comparisons. The model's superiority may reflect access to larger training resources as much as architectural innovation.

Reply: Thank you for your feedback. Irregardless of reason why EchoPrime is better than PanEcho, whether it is due to architecture alone or the overall workflow aggregating 10x more data, we believe our analyses confirm this observed performance advantage. Similar to how LLMs like ChatGPT/Claude/Gemini are compared, the ultimate comparison is evaluation on standardized metrics, and in this case, EchoPrime has strong performance compared to all public models with code. EchoPrime is a foundation model trained in a self-supervised contrastive manner, whereas PanEcho is a supervised model that requires explicit task-specific labels. The larger dataset available to EchoPrime reflects this difference in training paradigm. Namely, our approach allows for learning directly from unstructured echocardiography reports without manual labeling. Consequently, the scale of training data is intrinsic to foundation model design and represents a core architectural advantage.

In summary, while EchoPrime is an impressive model with promising results, its evaluation design needs greater transparency, its calibration should be quantitatively assessed, and its claims moderated in light of performance trade-offs and unresolved reviewer concerns.

Reply: Thank you for your feedback. We clarify in the discussion that it is difficult to comprehensively test all outputs of a vision-language model, and further work is required. We have secured funding and institutional support to conduct a prospective trial, which we anticipate conducting in 2026.

In the discussion:

> *Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime assessment. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials, and identifying optimal integration into clinical workflows.*

Referee #2 (Remarks to the Author):

For this review, I was asked to focus primarily on the relative merits of EchoPrime vs PanEcho, which you benchmark thoroughly in the supplement.

Overall, I have no doubt that EchoPrime is the superior model both in terms of performance and novelty. The authors' rebuttal gives good points of comparison between the two (3 with EchoNet-Measurements) models on the first page of your rebuttal to Reviewer 1. I did not see this in the main manuscript or the supplement, but you should consider including it to better amplify the differences in architecture and output among the 3 models.

Reply: Thank you for your careful review. As suggested, we have included a table highlighting the key architectural and output differences among the three models as Supplementary Table 10.

In Discussion:

> *Parallel work in direct annotation of echocardiography images can help inform assessment of echocardiographic measurements[31] (Supplementary Table 10).*

Supplementary Table 10 *Summary of key differences between EchoPrime, PanEcho, and EchoNet-Measurements.*

EchoPrime	***PanEcho***	***EchoNet-Measurements***
Contrastive Vision-Language Model, both video encoder CNN and text encoder (BERT)	*Supervised Training, Convolutional Neural Network*	*Supervised Training, Convolutional Neural Network*
Outputs Natural Language Text Reports	*Outputs classifications and numbers*	*Outputs segmentation annotations and associated numbers*
Can do all echo text tasks as it generates natural language report language	*Performs 39 specific tasks*	*Performs 18 specific tasks*
Trained on 12M videos	*Trained on 1.2M videos*	*Trained on 887k videos*
View informed inference	*Not view informed inference*	*single view inference*
External validation at 4 geographically independent sites with comprehensive echo videos from full studies	*External validation on 2 public single view datasets (EchoNet-Dynamic, EchoNet-LVH) and one geographically distinct site that include the full echocardiogram studies (RVENet).*	*External validation at 1 geographically independent site*

Overall, Supplemental Tables 2-6 show EchoPrime with superior performance in most tasks and especially outperforming in tasks that reviewers have previously criticized it for poor performance. In particular, EchoPrime's performance for LVEF (perhaps the most fundamental

parameter from an echo) is far superior to PanEcho (MAE 4.8 vs 7.5, p highly significant, though not stated). It is actually surprising how poorly PanEcho works for EF given that there are many AI algorithms that used similar supervised training with a CNN to get much better results (e.g., Asch et al, DOI: 10.1161/CIRCIMAGING.120.012293 and Tromp et al, doi.org/10.1038/s41467-022-34245-1). This holds true for virtually all parameters and diagnoses, and even in those few diagnoses for which the PanEcho AUROC point estimate is higher than EchoPrime's (e.g., RA dilation and bicuspid AV in the Cedars Sinai testset), the 95% C.I.'s overlap, showing no statistical significance. This becomes particularly striking for the many diagnoses that PanEcho did not train on, such as mitral annular calcification, MitraClip, and TAVR, but that EchoPrime, with its more holistic training (and on 10x data than PanEcho) was able to address, giving AUROCs for these diagnoses between 0.96 and 1.00.

EchoPrime is also more novel in its approach, whereas PanEcho is a standard ML model, using ~1 million echo videos to train a CNN by supervised training to perform specific tasks. EchoPrime does a nice job translating a key principle of AI research into the clinical domain: namely that some high dimensional space can be mapped to a low dimension (learned) manifold. In this case, an echocardiographic study (consisting of many videos each with many pixels, i.e. very many dimensions) can be mapped to a lower vector space that can be clustered. You take this one step further by doing report retrieval to basically translate that vector space back into clinician speak (i.e. embedding vector space into clinical natural language). One salutary consequence of this is that EchoPrime "thinks" like an echocardiographer, that is, it learns which views are likely to contribute to specific diagnoses, and which ones aren't (e.g., don't look at the apical 2-chamber view for any information on the RV).

Reply: Thank you for your thoughtful evaluation of our comparative results and for recognizing both the novelty and state-of-the-art accuracy of EchoPrime.

My principal critique of EchoPrime all along has been that it does not extract specific quantitative information from the echo, in particular that it does not use PW or CW Doppler to characterize diastolic function, aortic valve gradient, right ventricular systolic pressure, etc. However, I don't think PanEcho makes any effective use of these data either. PanEcho's r^2 for RVSP is only 0.27, significantly less than EchoPrime's (admittedly marginal) 0.43. Given that you have already submitted EchoNET-Measurements in preprint form, which demonstrates your ability to extract these measurements from Doppler, M-mode, and still images with an accuracy similar to sonographers. It is entirely logical to combine EchoPrime with EchoNet-Measurements to further finetune the report generation.

Reply: Thank you for recognizing our ongoing efforts to expand quantitative extraction capabilities. As noted, our EchoNet-Measurements framework demonstrates accurate Doppler, M-mode, and still-image quantification, and that will be integrated with EchoPrime in future proposed clinical workflow. We describe this in more detail in the discussion.

> *A few limitations are worth considering. Since EchoPrime relies on a video encoder, still images such as spectral Doppler and M-mode were excluded from the training set. Even though much of their information is already present in B-mode and Color Doppler videos, future iterations might benefit from including these image types. Parallel work in direct annotation of*

echocardiography images can help inform assessment of echocardiographic measurements[31] (Supplementary Table 10).

My only suggestion would be to consider being a bit more expansive in discussing PanEcho in the main manuscript to compare and contrast with EchoPrime, perhaps putting the 3-model comparison from your rebuttal in the supplement and directing the readers' attention to specific findings in Supplemental Tables 2-6.

Thank you for your suggestion. We aggregate comparison results of EchoPrime and PanEcho across all external validation sites in Extended Data Table 3, and edit the manuscript:

In the Results:

Compared to other echocardiography AI models, EchoPrime has superior performance (Supplementary Tables 1-5, Extended Data Table 3), particularly driven by differences in low quality and low view count images when informed by anatomical attention (Extended Data Table 4).

In Discussion:

Parallel work in direct annotation of echocardiography images can help inform assessment of echocardiographic measurements[31] (Supplementary Table 10).

Supplementary Table 10 *Summary of key differences between EchoPrime, PanEcho, and EchoNet-Measurements.*

EchoPrime	***PanEcho***	***EchoNet-Measurements***
Contrastive Vision-Language Model, both video encoder CNN and text encoder (BERT)	*Supervised Training, Convolutional Neural Network*	*Supervised Training, Convolutional Neural Network*
Outputs Natural Language Text Reports	*Outputs classifications and numbers*	*Outputs segmentation annotations and associated numbers*
Can do all echo text tasks as it generates natural language report language	*Performs 39 specific tasks*	*Performs 18 specific tasks*
Trained on 12M videos	*Trained on 1.2M videos*	*Trained on 887k videos*
View informed inference	*Not view informed inference*	*single view inference*
External validation at 4 geographically independent sites with comprehensive echo videos from full studies	*External validation on 2 public single view datasets (EchoNet-Dynamic, EchoNet-LVH) and one geographically distinct site that include the full echocardiogram studies (RVENet).*	*External validation at 1 geographically independent site*

Referee #3 (Remarks to the Author):

I have carefully reviewed the comments from all reviewers as well as the authors' rebuttal. In their response, the authors state that comparisons with third-party models are infeasible due to data privacy constraints. However, this justification is not entirely persuasive. Numerous state-of-the-art large language models (LLMs)—including LLaVA-3.2, LLaVA-Med, and Qwen-3—are openly available and amenable to fine-tuning in secure, local environments without violating data governance requirements. Within the field of multimodal report generation, fine-tuning LLMs on domain-specific datasets has become standard practice. Given that the authors possess a 12-million-sample dataset, the omission of benchmarking against fine-tuned open-source LLMs constitutes a significant gap. A direct comparison would provide a more rigorous and informative assessment of EchoPrime's relative performance.

From a methodological standpoint, the core architecture employed in EchoPrime (i.e., a convolutional video encoder paired with a BERT-based text module) does not reflect contemporary standards in multimodal learning. The absence of a foundation model backbone pretrained on large-scale instruction-following or generative tasks (e.g., via causal language modeling or reinforcement learning with human feedback) limits the model's expressiveness and its capacity for generalization across diverse clinical reporting tasks. Although the authors present an extensive retrospective evaluation, the underlying model design is technically dated and lacks the flexibility, reasoning capability, and transferability that characterize current-generation multimodal architectures.

Reply: Thank you for your feedback. Based on your feedback, we trained multiple iterations of the model with varying batch sizes (Supplementary Table 7) and encoder architectures (Supplementary Table 8), both of which had minimal impact on performance, confirming that our original design achieved the best overall accuracy.

Regarding the newly raised point about large language models, we agree that fine-tuned multimodal LLMs are an exciting area of research. However, our prior experience applying such models to echocardiography indicates that they remain limited in domain-specific understanding and temporal reasoning. As shown in our recent analysis (https://arxiv.org/abs/2504.14391)

Finally, we respectfully note that EchoPrime's methodology continues to be viewed by the community as both novel and impactful. Its open-source release has attracted active engagement, with multiple external contributors extending its capabilities. Most recently, a pull request (merged on October 12, 2025) added support for standardized reporting in Italian language (https://github.com/echonet/EchoPrime/pull/11).

Moreover, as emphasized by other reviewers, the absence of prospective clinical validation—whether through randomized trials, user studies, or real-world workflow integration—substantially diminishes the translational significance of the work. Claims regarding improvements in clinical efficiency and report quality remain hypothetical in the absence of empirical evidence. If we waved the need for clinical testing of this model, the contribution of the

proposed AI solution is outdated, limited, and not SOTA, thus not promising and novel.

In its current form, the manuscript offers limited advancement in both AI algorithmic innovation and clinical applicability. Substantive revision is warranted, including comparative evaluation against modern domain-adapted LLMs and incorporation of prospective or clinician-facing validation, in order to meet the threshold for publication in Nature.

Reply: Thank you for your feedback. We clarify in the discussion that it is difficult to comprehensively test all outputs of a vision-language model, and further work is required. We have secured funding and institutional support to conduct a prospective trial.

In the discussion:

> *Although EchoPrime was evaluated at four geographically distinct academic medical centers, all data was collected retrospectively. Future experiments should include prospective evaluations in real-world settings to understand human-computer interaction as well as accuracy and acceptability of EchoPrime assessment. Working within a complex ecosystem such as healthcare requires understanding the human-computer interaction through clinical trials and identifying optimal integration into clinical workflows.*

We respectfully disagree with the assessment that the work lacks novelty or state-of-the-art performance. As noted by other reviewers, EchoPrime represents a comprehensive and innovative framework for full echocardiographic interpretation, covering all major tasks in a unified architecture. Its performance, summarized in Supplementary Tables 1–5, demonstrates consistent superiority over existing models across diverse benchmarks. Furthermore, our experiments applying such models to echocardiography indicates that they remain limited in domain-specific understanding and temporal reasoning. As shown in our recent analysis (https://arxiv.org/abs/2504.14391)