nature portfolio

Peer Review File

Contrastive Language Image Pretraining for a Cardiac Magnetic Resonance Image Embedding with Zero-shot Capabilities

Corresponding Author: Dr David Chen

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

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
Summary
The paper introduces CMRCLIP, a contrastive vision-language model for Cardiac Magnetic Resonance (CMR) imaging. The approach treats CMR studies as video sequences and aligns them with radiologists' reports using a modified space-time transformer architecture. Trained on 13,786 CMR studies, the model demonstrates zero-shot and few-shot capabilities for tasks like disease classification, image/text retrieval, and cardiac function analysis. Evaluations on internal and external datasets (ACDC) show superior performance over generalist (OpenAI CLIP) and biomedical-specific (BiomedicalCLIP) models. The framework aims to reduce reliance on labeled data and improve efficiency in CMR interpretation.
Strengths:
1. This work adapts video-language pretraining to CMR, addressing the modality's temporal and multi-view complexity.
2. This work leverages 13k+ unlabeled CMR studies, mitigating data scarcity in medical AI.
3. This work outperforms baselines by 35–42% in zero-shot tasks and achieves state-of-the-art performance in few-shot and supervised settings.
4. This work evaluates diverse tasks (classification, retrieval, segmentation) and external validation (ACDC), demonstrating generalizability.
5. This work combines CINE and LGE image types, enhancing semantic alignment and diagnostic coverage.

Weaknesses:
1. The external evaluation on ACDC dataset is limited. The public ACDC dataset only includes short-axis CINE images, failing to test the model's robustness to other CMR sequences (e.g., LGE, 3-chamber views). LGE and 3-chamber views are long-axis views of heart.
2. In the first table of supplementary material (zero-shot classification tasks), some zero-shot AUCs are low (e.g., 0.551 for LGE detection), raising doubts about clinical reliability. The proposed CMRCLIP can generalize well on the short axis view but cannot perform very well on the long axis one, suggesting its limited generalizability.
3. The high variance observed in few-shot settings could indicate instability in the model's performance when applied to small datasets. For instance, it achieves 89% on left ventricle dysfunction classification but achieves 59.1% on aortic dilation classification in the 32 few-shot setting.
4. I am concerned about the bias and generalizability of CMRCLIP. No discussion of dataset biases (e.g., vendor differences, demographic skew) or handling of incomplete/missing studies. It would be better if the author could provide more information about the dataset.
5. This work mainly focuses on embeddings but neglects generative tasks (e.g., report generation), limiting direct clinical applicability. It would be better if the author could leverage the image encoder and text encoder to conduct more experiments on report generation and image caption tasks to evaluate the utility of CMRCLIP in different generative tasks.
6. While the manuscript compares CMRCLIP to generalist models (e.g., OpenAI CLIP, BiomedicalCLIP), there is no comparison to CMR-specific models or task-specific deep learning models that might already exist for cardiomyopathies, segmentation, or ejection fraction prediction.
7. The utility of CMRCLIP is not benchmarked against human cardiologists or radiologists. Comparing the model's performance to expert interpretations would provide stronger evidence of its clinical applicability.
8. The quality of this paper is very poor. All the provided figures are very blurry. In Figure 1, the figure is too small to read.

The paragraphs in the section (named CMR Interpretation without supervision) do not align with both sides. The tables in supplemental materials do not include the table caption.
9. The CMRCLIP model is useful to the medical AI community. Although they provide source code to the paper, the weights of CMRCLIP are not provided in their repository, which is the most important stuff.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Dear Editor,

I read with great interest the original research entitled “CMR-CLIP: Contrastive Language Image Pretraining for a Cardiac Magnetic Resonance Image Embedding with Zero-shot Capabilities.” The manuscript addresses a highly relevant and timely topic, focusing on self-supervised learning for cardiac MRI, and proposes a novel model trained on a large internal dataset.The paper is well-done, interesting and offers valuable points of reflection. However, several aspects of the manuscript would benefit from clarification and refinement to enhance its rigor, readability, and impact.
Analytical comments are as follows:
The objectives and rationale of the study are not clearly articulated in the Abstract and Introduction. A clearer statement of aims would improve the focus of the paper.
The authors mention "weak alignment" This term should be defined more clearly to ensure readers fully understand its meaning in the context of this study.
The sentence “We show that the proposed CMR-CLIP achieved remarkable performance in real-world clinical tasks, such as CMR image retrieval and diagnostic report retrieval in our internal held-out test set” would benefit from including specific performance metrics or results to give readers a clearer sense of the model's effectiveness.
The quality of the figures is too low and should be improved to ensure proper visualization and interpretation of results.
The Introduction would benefit from further elaboration on how the proposed model could impact clinical workflows and decision-making.
The Introduction suggests that existing models cannot handle CMR well, but it does not sufficiently explain why CMR is uniquely challenging compared to other imaging modalities. This point should be expanded for context.
Ensure that all acronyms (e.g., ACDC) are spelled out the first time they are mentioned in the abstract or introduction.
A clearer and more detailed description of inclusion and exclusion criteria would improve reproducibility and transparency
More comprehensive information on the dataset should be provided, including sequences used, and disease types represented.
The description of the CMR imaging protocol currently mixes technical and clinical aspects, which may confuse readers. I suggest separating technical details (e.g., scanner specifications, sequences) from clinical indications for better clarity.
Did the authors observe differences in model performance based on the scanner used (e.g., 1.5T vs. 3.0T)? This could be an important factor influencing generalizability.

More details are needed regarding the radiology report data used. For instance, did the authors focus only on the "impression" section? How did they handle variability in language across reports?
There is no mention of ethical approval, data privacy protections, or patient consent procedures. These should be explicitly addressed.
While technical aspects are well discussed, the potential clinical utility and real-world applications of the proposed model are not sufficiently elaborated. Expanding on this would increase the impact and relevance of the work.

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The author addressed all the concerns well.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
Review – NCOMMS-25-09811-T
CMR-CLIP: Contrastive Language–Image Pretraining for a Cardiac MR Embedding with Zero-shot Capabilities

Summary

This manuscript introduces CMR-CLIP, a vision–language foundation model that treats a complete CMR study as a video-like sequence and aligns it with the impression section of the clinical report via contrastive self-supervision. The authors pretrain on a substantial single-system cohort (13.8k studies), evaluate on an internal hold-out and an external site, compare against a video baseline and a task-specific amyloidosis model, test zero-shot/few-shot classification, cross-modal retrieval, transfer learning, robustness to perturbations, and report generation (retrieval-based and autoregressive).
I appreciate the replies to prior reviews: the added external site, the stability experiments, the generative evaluation, and the baseline comparisons materially strengthen the paper. In my view the manuscript is already mature and potentially impactful, the modeling choice fits CMR's multi-view temporal nature, and the scale is noteworthy. To reach its full potential (generalizability, clinical utility, and reproducibility), I recommend the following focused additions/clarifications.

Explicitly confirm patient-level splits for pretraining/val/test and all downstream tasks.

For comparison cohorts (amyloidosis/HCM), confirm that no images or reports from those patients entered pretraining to avoid implicit text leakage. A simple data-flow diagram would help.

The revised prompting improved LGE AUC, but LGE remains label-sensitive. Please consider ading subanalysis with curated LGE labels

Specify the provenance of disease labels (final diagnosis vs report impression) and how potential mislabels were handled.

I would sugget to publish the full prompt list. Provide a prompt-sensitivity analysis (synonyms, negations, minor wording changes) and consider prompt ensembling/logit averaging.

I suggest adding in appendix Brier scores and calibration curves. If possible, include Decision Curve Analysis to quantify net benefit for key tasks.

Complement ROUGE/BLEU with clinician-blinded factuality on key fields (e.g., EF category, LVH presence, LGE pattern) and quantify hallucination rates (are there hallucination in the experiment?).

Include a brief privacy audit for text: de-identification steps and a membership-inference/nearest-neighbor check to rule out verbatim reuse of training impressions.

Excellent that weights are released. Could be usefult to add a model card with full training configs (hyperparameters, seeds, optimizer, batch size, temperature), preprocessing (Impression parser), sampling of the 64 frames, and 3+ reruns with different seeds reporting mean±SD.

Editorial consistency: correct minor slips ("CMR-CONVIRT", "aggravating" → "aggregating", "compared of" → "comprised of").

Figures/Tables: Good that images are now ≥300 dpi. Please standardize font sizes/scales, axis labels, and units. Expand abbreviations at first occurrence in Table 1 and captions (example T = Tesla, Y = years…=

Ethics/Privacy: The IRB statement is now present; briefly restate de-identification, and exactly what is released (weights, code; no data).

Conclusion

The manuscript is solid and already mature, and the authors' responses have substantially improved scope and rigor. With targeted additions the work would convincingly demonstrate generalizability and clinical utility.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #3

(Remarks to the Author)
I have reviewed the revised manuscript and the authors' point by point responses. The authors have thoroughly addressed all concerns. Key improvements include clearer patient level data allocation, explicit label sources, a comprehensive model card, a complete list of prompts, and new analyses (calibration curves, nearest neighbor checks, and factual verification of generated reports). The responses are thoughtful, and the manuscript has been significantly strengthened. I believe this paper is now suitable for publication in Nature Communications.
Thank you for the opportunity to review this work.

(Remarks on code availability)

Response to Reviewers: NCOMMS-25-09811-T

CMR-CLIP: Contrastive Language Image Pretraining for a Cardiac Magnetic Resonance Image Embedding with Zero-shot Capabilities

We thank the editors and reviewers for their careful and insightful review and the opportunity to further address them. Most significantly, we have provided added experimentation around domain specific, focused models as requested by the editor and Reviewer 1. We further added significant results from data taken from a separate site from our enterprise, showing generalizability to unseen data.

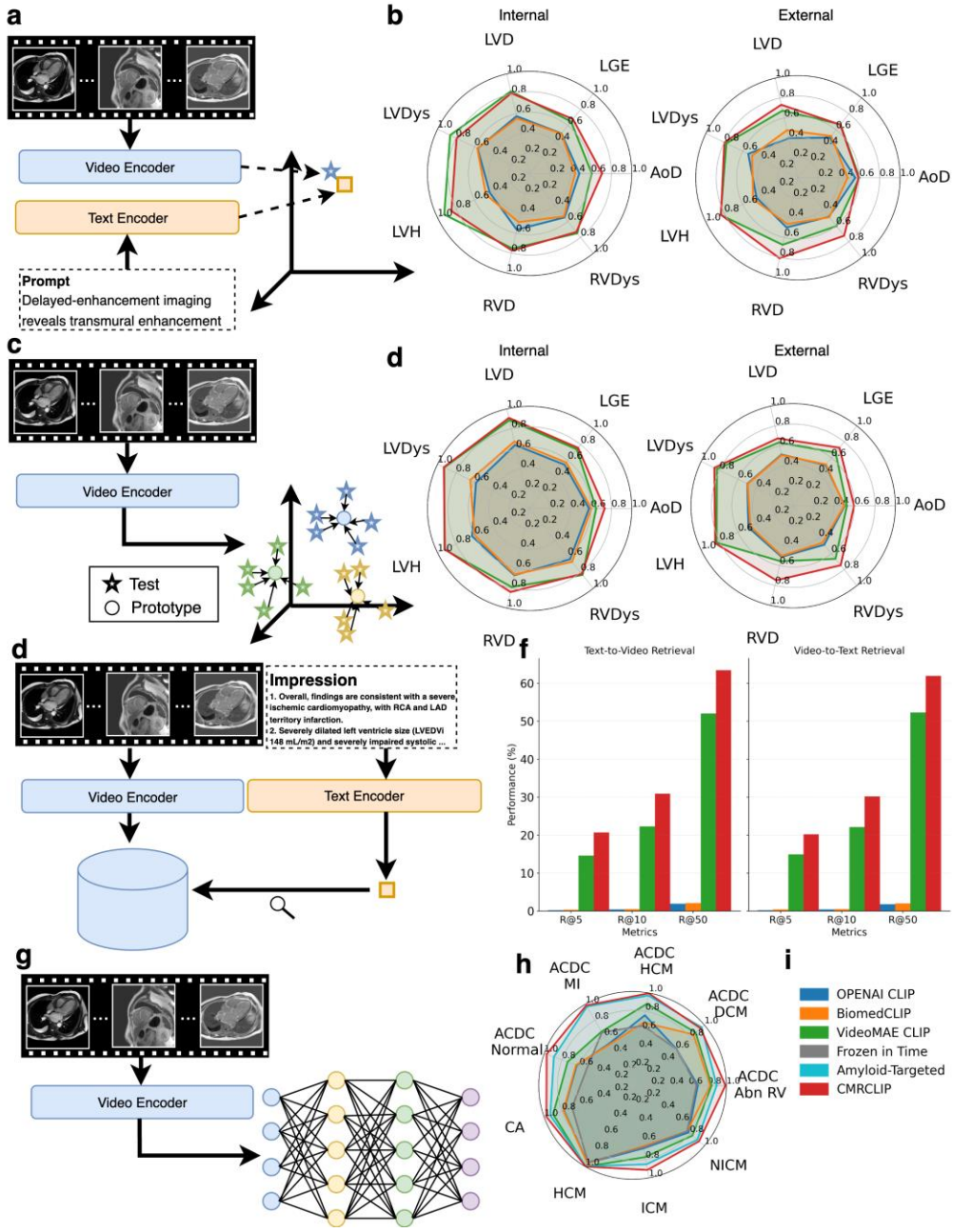
Note: Reviewers' specific critiques are written in bold and the response provided below.

Specific response to reviewer's critiques:

Reviewer 1:

R.1.1. The external evaluation on ACDC dataset is limited. The public ACDC dataset only includes short-axis CINE images, failing to test the model's robustness to other CMR sequences (e.g., LGE, 3-chamber views). LGE and 3-chamber views are long-axis views of heart.

We agree with the reviewer that the public dataset is limited, particularly with respect to the breadth of available data. We have since attempted to address this issue by evaluating our model in a separate site without our extended enterprise (Florida site) which does not share readers, vendors, or patient population with our training site. There is a small reduction in performance with respect to zero-shot and few-shot classification, however, our model achieves better results compared to other open-domain models. We copied the new results as shown in the updated Figure 2 below:



We have also added discussion of the general lack of public CMR data in the “Discussion” section, reproduced below for convenience.

> “The ACDC dataset used for public reproducibility is also limited both in volume and breadth of data, containing only SAX Cine images. This reflects a general trend in the lack of available public advanced cardiac imaging data. EchoNet-dynamic represents the largest publicly available dataset (10k echocardiograms), although it is similarly limited with respect to available view (only apical-4-chamber). Pace et al released a 3D CMR dataset for whole heart segmentation[37] although it represents only patients with congenital heart disease and does not include standard 2D views and image types”

R.1.2. In the first table of supplementary material (zero-shot classification tasks), some zero-shot AUCs are low (e.g., 0.551 for LGE detection), raising doubts about clinical reliability. The proposed CMRCLIP can generalize well on the short axis view but cannot perform very well on the long axis one, suggesting its limited generalizability.

We appreciate the reviewer's concern about the relatively low zero-shot AUC for LGE detection. Identifying late-gadolinium enhancement is a notoriously challenging task. Expert cardiologists show significant inter-observer variability in evaluating extent and even presence of LGE, particularly for non-ischemic cardiomyopathies (Flett *et al. 2011. JACC Cardiov Imag*). In the previously submitted version of our study, we grouped any mention of LGE in the report as LGE positive (as stated in the Methods). Unfortunately, this doesn't capture the nuance of LGE pattern nor does it capture the real importance of LGE as a diagnostic/prognostic tool. Specifically, large focal LGE scars which reflect ischemic disease may be very clear and consistently read by humans are grouped with diffuse enhancement indicative of non-ischemic processes.

To deal with the variability, we changed prompt to identify patients without LGE as that is more consistently reported. With these adjustments, our LGE detection AUC increased from 0.551 to 0.680 in internal cohort. We have updated results in the manuscript to reflect the new prompting strategy.

Also of note, although the model does not identify binary presence or absence of LGE highly accurately, it does identify diseases which require accurate LGE diagnosis clearly. Specifically, in disease classification tasks, CMR-CLIP enables high AUC in differentiating HCM, cardiac amyloidosis, ischemic cardiomyopathy, and non-ischemic cardiomyopathy. Each of these diseases are defined by LGE pattern and CMR-CLIP was able to achieve over 0.9 AUC for each disease, pointing to its ability to learn different patterns of LGE (Figure 2h).

R.1.3. The high variance observed in few-shot settings could indicate instability in the model's performance when applied to small datasets. For instance, it achieves 89% on left ventricle dysfunction classification but achieves 59.1% on aortic dilation classification in the 32 few-shot setting.

We agree with the reviewer that highly variable results can indicate model instability. However, in this case the source of that instability may be more representative of the capabilities of underlying imaging modality and limitations around using imaging reports as our ground truth rather than the actual model performance. As the reviewer noticed, the model performs relatively worse in right ventricular (RV) disease compared to left ventricular (LV) disease. However, the evaluation of RV disease and RV biomarkers are inherently more difficult to interpret due to the complex anatomy and limited through plane spatial resolution offered by CMR [*Kjaergaard et al. 2006. Eur J of Echo*]. This complexity causes even discrete measures of RV function such as RVEF to show far more variability between scans compared to LVEF, resulting in more human variability in interpretation/documentation.

However, we agree with the reviewer that model stability is an important characteristic that needs to be evaluated. We added experiments to assess model instability to through their sensitivity to perturbations in the data such as using slightly different slices or missing frames and compared the embedding distance with average distance between randomly selected studies. We found

that CMR-CONVIRT the distance between studies were much greater than intra-study distance. In contrast, other models were much more sensitive to intra-study noise.

> “We further evaluate model stability by comparing embedding variability in response to different perturbations in the input data compared to average cosine distances between studies. This is intended to demonstrate that the developed models capture features important to disease detection but not sensitive to subtle changes in the input data. We focus on two kinds of input perturbations: changes in the included frames (linear shift and random swapping), and missing data (masking and random sub-sampling). The results of CMR-CLIP stability compared with other models are shown in Fig. 4c, where CMR-CLIP demonstrates the smallest change in perturbed embedding compared to other models. This is particularly noticeable compared to public foundation models where perturbations in the input data can exceed potential differences between studies. Changes in the included data such as linear shifts or random swapping has a lower impact on embedding compared to missing data. VideoMAE-CLIP exhibited significant sensitivity to both shifts which is substantially higher than that of other foundation models. In contrast CMR-CLIP greatly reduces the embedding model’s sensitivity to both shifts in the included frames and missingness of data.“

We also want to address some of the variability of the model is likely due to inherent limitations of the underlying imaging. For example, aortic dilation is inherently challenging to detect from standard Cine and LGE sequences as their slice locations do not adequately cover the aorta. Clinicians typically rely on a dedicated 3D sequence to accurately assess aortic dimensions. Although readers will comment on aortic dimensions based on incidental coverage in the localizers, it is not necessary accurate. We have added some discussion of this limitation to the Discussion section.

> “First, CMR-CLIP contains only a fraction of the data used in a full protocol. … The limited context window may have contributed to some of the lower accuracy such as aortic dilation, where CMR-CONVIRT does not include the necessary views to consistently and accurately identify this pathology.”

R.1.4. I am concerned about the bias and generalizability of CMRCLIP. No discussion of dataset biases (e.g., vendor differences, demographic skew) or handling of incomplete/missing studies. It would be better if the author could provide more information about the dataset.

We agree with the reviewer that readers would benefit for more context of our claims of generalizability and site differences. We have added a table detailing the demographics, scanners, and general indications for CMRs at different sites. The further break down results based on subgroups such as by vendor and by field strength to show there are not large differences between sites and vendors. As stated previously, we have also added experiments on stability given data missingness. We summarized these findings in the results section as follows:

> *“*CMR-CLIP demonstrated robust zero-shot generalizability across scanner vendors and field strengths (Supplemental Table 12-15). On internal Philips data (N=2,726), AUCs of the findings ranged from 0.681-0.850, while a small internal Siemens subset (N=32) yielded similar AUC (0.617-0.926) with greater variance. In an external Siemens cohort (N=428), AUC (0.594-0.836)

remained consistent, indicating resilience to domain shift. Stratification by field strength on the internal cohort showed comparable performance at 1.5T (N=1,844; AUC 0.694-0.851) and 3.0T (N=914; AUC 0.652-0.854), and on the external cohort at 1.5 T (N=421; AUC 0.605–0.830), with a very small external 3.0T subgroup (N=7) showing wider metric variability."

R.1.5. This work mainly focuses on embeddings but neglects generative tasks (e.g., report generation), limiting direct clinical applicability. It would be better if the author could leverage the image encoder and text encoder to conduct more experiments on report generation and image caption tasks to evaluate the utility of CMRCLIP in different generative tasks.

We thank the reviewer for this suggestion. We have added experiments with respect to generative applications of CMRCLIP around report drafting. We take two approaches:

- Retrieval-based report drafting. We used the embedding space to retrieve the most relevant, structured report templates from training data.
- CMR-TARGET. By feeding CMRCLIP's image embeddings and indications into a Transformer-based generator, we produced fully synthetic reports.

The results of these additional experiments are detailed in a new subsection of the results *Performance on Generative Tasks* of the revised manuscript. The results illustrate CMR-CLIP's utility not only for classification and retrieval but also for end-to-end report generation. The text is copied below for convenience:

"To evaluate CMR-CLIP's utility beyond classification and retrieval, we implemented two approaches for automated report generation. First, we designed a retrieval-based generation that leverages the learned embedding space to select the most relevant impression from the training set. The images of each CMR study is encoded into embeddings by the frozen CMR-CLIP vision encoder as a key for the previously written report. When querying new studies, we compute the cosine similarity between the new study with the stored ones during training. The impression associated with the highest-scoring embedding is retrieved. Next, we also trained a Transformer-based AutoRegressive Generator for Text (CMR-TARGET) for end-to-end synthetic report generation. CMR-TARGET takes the image embedding from CMR-CLIP concatenated with the written indication for the CMR exam. Each token is generated one after the other, with new tokens taking into account the image embeddings, indications, and all previously generated words.

Both methods were evaluated on a test set of 2,758 impressions. The retrieval-based method achieved a ROUGE-L score of 0.215 and a BLEU-4 score of 0.072, whereas CMR-TARGET achieved a ROUGE-L of 0.39 and a BLEU-4 of 0.236. Qualitatively, CMR-TARGET demonstrated greater NLP evaluation metrics. Example outputs illustrating the contrast between retrieval-based drafts and CMR-TARGET generations are shown in Fig. 3c. These experiments confirm that CMR-CLIP's learned vision embeddings can be repurposed not only for classification and static retrieval, but also for generating coherent, clinically useful impressions in both template-based and generative frameworks."

R.1.6. While the manuscript compares CMRCLIP to generalist models (e.g., OpenAI CLIP, BiomedicalCLIP), there is no comparison to CMR-specific models or task-specific deep

learning models that might already exist for cardiomyopathies, segmentation, or ejection fraction prediction.

We agree that a direct comparison to a CMR-specific baseline strengthens our evaluation. Accordingly, we implemented a VideoMAE-CLIP, the approach baseline inspired by the ViTa and InternVideo framework. We first pretrain a masked autoencoder on all 64 frames of Cine and LGE CMR images, following the VideoMAE protocol. And we take the pretrained video encoder and perform CLIP-style contrastive learning. As with our CMR-CLIP model, Bio+ClinicalBERT is used as the text encoder. As shown in the revised Figure 2 of the main text, CMR-CLIP consistently outperforms VideoMAE-CLIP across nearly all downstream tasks ranging from zero-shot classification to retrieval.

We also compared a task-specific model, Amyloid-Targeted (Cockrum et al. 2025 JACC: Cardiovascular Imaging), for the disease classification task. The Amyloid-Targeted model's average AUC was 0.931, while CMR-CLIP's average AUC was 0.967.

We did not explore segmentation or ejection fraction given there are multiple highly functional commercial options.

R.1.7. The utility of CMRCLIP is not benchmarked against human cardiologists or radiologists. Comparing the model's performance to expert interpretations would provide stronger evidence of its clinical applicability.

We completely agree with the reviewer that comparison to human readers would provider stronger evidence of the clinical utility. To address this point, we provide two more comparisons using the targeted amyloidosis prediction task. This dataset includes both manual extraction of the initial clinical read as well as final diagnosis. We have added Figure 3a to emphasize this human comparison and associated discussion on results. The results are included below for convenience.

> "We furthered compared the performance of the classification models in a cohort with final diagnosis in addition to the reader interpretation. Also included in the labels is reader diagnostic confidence. This cohort included cardiac amyloidosis, hypertrophic cardiomyopathy, and other patients referred for evaluation for infiltrative cardiomyopathies. In this dataset, the readers correctly diagnosed amyloidosis in 62.9% of cases when they were strongly confident. When aggravating all mentions of amyloidosis, readers were found to be accurate in 77.3% of cases. CMRCLIP reached an accuracy of 82.5% in this dataset. For hypertrophic cardiomyopathy, both human readers and CMRCLIP achieved an accuracy exceeded 89%. Finally, readers achieved only 53.1% when confident and 70.3% when aggravating all mentions, compared to 87.5% achieved by CMRCLIP."

R.1.8. The quality of this paper is very poor. All the provided figures are very blurry. In Figure 1, the figure is too small to read. The paragraphs in the section (named CMR Interpretation without supervision) do not align with both sides. The tables in supplemental materials do not include the table caption.

We thank the reviewer for these helpful formatting and presentation suggestions. We have revised both the main manuscript and the supplemental file as follows:

1. High-resolution figures
    - All figures have been replaced with high-resolution versions (≥ 300 dpi).
2. Figure 1 legibility
    - Figure 1 has been enlarged to span the full text width and the font sizes for all labels have been increased to ensure easy reading.
3. Text alignment in “CMR Interpretation without supervision”
    - The entire section has been reformatted so that paragraphs align on both left and right margins.
4. Supplemental table captions
    - Descriptive captions have now been added to all tables in the supplemental materials.

R.1.9. The CMRCLIP model is useful to the medical AI community. Although they provide source code to the paper, the weights of CMRCLIP are not provided in their repository, which is the most important stuff.

We thank the reviewer for highlighting the importance of weight availability. To ensure full reproducibility and community adoption, we have now uploaded the pretrained CMR-CLIP weights to HuggingFace.

Reviewer 2:

R.2.1. We respond to the following critiques together given their closely related to the presentation of the significance and impact of our manuscript:

The objectives and rationale of the study are not clearly articulated in the Abstract and Introduction. A clearer statement of aims would improve the focus of the paper.

The Introduction would benefit from further elaboration on how the proposed model could impact clinical workflows and decision-making.

The Introduction suggests that existing models cannot handle CMR well, but it does not sufficiently explain why CMR is uniquely challenging compared to other imaging modalities. This point should be expanded for context.

The authors mention "weak alignment" This term should be defined more clearly to ensure readers fully understand its meaning in the context of this study.

We thank the reviewer for pointing out lack of clarity around the significance and potential impact of this work. We have modified the Abstract and Introduction to emphasize the technical difference between images and video embeddings for foundation models and the overall impact of foundation models on the implementation/dissemination of AI technologies to healthcare. Specifically, the breadth of potential tasks around use of CMR and lack of data to support these tasks is the major problem for applying AI to CMR. We also realize that “weak alignment” is ambiguous and potentially not optimal for the readership. The intention of this term is to present the idea that impression sentences rarely refer to a single image, but instead a combination of all views and image types (e.g. “LGE pattern and thickened ventricles are indicative of non-ischemic cardiomyopathy”). Instead, we change the terminology to “singular image and sentence pairs” which we hope is clear to a clinical audience. The change in the abstract is copied below for the reviewer’s convenience.

> *“*Developing task-specific artificial intelligence (AI) models to address individual tasks in healthcare is not scalable given both the large number of potential clinical tasks and the workload required to adequate volume of labels for each task. Foundation models trained using self-supervised learning are crucial to broad dissemination of AI technologies by reducing the dependency on large volumes of labeled data. However, conventional multi-modal self-supervised training approaches that rely on precise vision-language alignment are not always feasible in complex clinical imaging modalities, such as cardiac magnetic resonance (CMR). CMR provides a comprehensive visualization of cardiac anatomy, physiology, and microstructure. This requires the reader to synthesize information from long sequences of images taken a different spatial locations of the heart and representing different tissue characteristics. Therefore, there is rarely a one-to-one mapping between individual sentences and singular frames. The lack of strong alignment between individual sentences and images means foundation models trained using highly specific captions and images typical of multi-modal biomedical foundation models typically have poor downstream generalizability in clinical applications.*”*

The change in the introduction is copied below for convenience.

> “However, conventional multi-modal self-supervised training approaches that rely on precise image-text pairing are not always feasible in complex clinical imaging modalities, such as cardiac magnetic resonance (CMR). CMR provides a comprehensive visualization of cardiac anatomy, physiology, and microstructure. The reader is required to synthesize information from long

sequences of images taken at different locations of the heart and visualizes different tissue characteristics. Therefore, there is rarely a one-to-one pairing between individual sentences and singular frames in real world CMR data, resulting in poor downstream generalizability of typical biomedical foundation models."

We have also changed the introduction to reflect lack of singular paired sentences and images:

"Such vision-language frameworks have been applied to biomedical data as well including chest X-ray, single view echocardiography, and pathology slides[13–15] to improve domain specific performance beyond natural domain models. However, many of these data types are limited in data variety. For example, chest X-rays are usually limited to a single frontal and lateral view during an exam. Similarly, the work by Christensen *et al* in echocardiography was limited to only 4 chamber view. In contrast, typical CMR study is compared of over 1000 2D images with both temporal and volumetric relationships. Findings in CMR impressions rarely refer to features contained in single images (e.g. viability is often correlated with wall motion abnormalities or myocardial morphology), making it difficult to map individual sentences and images. This limits the efficacy of pretraining using pairs of singular image and text, akin to poor motion understanding of many multi-modal models."

2.2. While technical aspects are well discussed, the potential clinical utility and real-world applications of the proposed model are not sufficiently elaborated. Expanding on this would increase the impact and relevance of the work.

We thank the reviewer for pointing out the lack of discussion around clinical utility. We further elaborate on the clinical utility and real-world use case for supporting CMR interpretation in both the Introduction and Discussion sections. Here, we cite some of our prior work examining the long documentation/interpretation time for CMR. Current analytic aids can certainly ease reader analytic burden by extracting common imaging biomarkers but stop short of diagnose, particularly in the long tail of rare diseases. Furthermore, we note that these rare diseases are often mis-diagnosed. We have added a paragraph to the Discussion to emphasize these potential applications of our model. This is reproduced below for convenience:

"We demonstrate that the learned model holds immense potential for a wide variety of clinical applications beyond generalist foundation models or single image-based foundation models, and on par with some specialist models. Specifically, CMR-CLIP exceeded other models in identifying common pathologies such as myocardial fibrosis and left ventricular hypertrophy. It also exceeded other models in common computational tasks including retrieval of CMR studies or radiology reports and downstream disease classification tasks."

*"*CMR is an information dense imaging modality due to its ability to capture cardiac anatomy, function, and tissue characteristics. AI decision support tools to aid in CMR interpretation is particularly important to increasing access given the lengthy interpretation process[3] and the lack of readers[19]. Current commercial tools to aid in interpretation are focused on automating biomarker extraction. Such tools, in combination with intelligently engineered data pipelines can drastically reduce CMR documentation times[~r]. However, standard biomarkers are not adequate to accurately diagnose many cardiac diseases, particularly in the long tail of rare diseases[~r]. These diseases (e.g. cardiac amyloidosis or Fabry disease) are frequently mis-diagnosed because subtle image features are not immediately apparent and not well differentiated in standard imaging biomarkers. Unfortunately, developing decision aids for such diseases is one of the more difficult tasks in healthcare because the total volume of CMR is 1/100th of that of other cardiac imaging modalities[20,21], and largely limited to specialized cardiac care centers[21]. A foundation model developed specifically for CMR can have outsized influence in enabling AI development for such

diseases by vastly lowering the volume of data necessary, instead relying on the emergent abilities encoded in the model during self-supervised training as shown by this work.”

R.2.3. The sentence “We show that the proposed CMR-CLIP achieved remarkable performance in real-world clinical tasks, such as CMR image retrieval and diagnostic report retrieval in our internal held-out test set” would benefit from including specific performance metrics or results to give readers a clearer sense of the model's effectiveness.

Thank you for the suggestion. We have updated the manuscript to include retrieval metrics alongside our qualitative statement, as reproduced below.

> “CMR-CLIP achieved remarkable performance in real-world clinical tasks, such as demonstrating high disease-classification performance, achieving accuracies of 87.8% for non-ischemic cardiomyopathy (NICM), 88.8% for ischemic cardiomyopathy (ICM), 96.9% for cardiac amyloidosis (CA), and 99.0% for hypertrophic cardiomyopathy (HCM).”

R.2.4. The manuscript could benefit from editing, specifically the following:
R.2.4.1. The quality of the figures is too low and should be improved to ensure proper visualization and interpretation of results.

Thank you for highlighting this issue. We have regenerated and replaced all figures with high-resolution versions (≥300 dpi).

R.2.4.2. Ensure that all acronyms (e.g., ACDC) are spelled out the first time they are mentioned in the abstract or introduction.

We thank the reviewer for pointing out this issue. We have carefully reviewed the abstract and introduction and now spell out every acronym at first use.

R.2.5. A clearer and more detailed description of inclusion and exclusion criteria would improve reproducibility and transparency. More comprehensive information on the dataset should be provided, including sequences used, and disease types represented.

We agree with the reviewer for this suggestion. In the revised manuscript we have added a new Dataset subsection “Data and Data Preprocessing” that now includes:

1. **inclusion and exclusion criteria**

   Studies were excluded if they failed to include all views, Cine LAX and SAX or LGE LAX, SAX, 2ch, and 3ch. We also excluded any study in which the mid-slice Cine LAX and SAX series comprised fewer than 30 frames. Studies with LGE SAX stacks containing fewer than 10 slices or more than 15 slices were removed. We excluded studies which we were not able to retrieve a report for (e.g. report was saved in another system we could not access).

2. Indications for the CMR which are presented in the new Table 1.
3. CMR sequences & preprocessing

   In our image preprocessing pipeline, the mid-slice Cine LAX and Cine SAX sequences were treated as temporal series, with each frame representing time. By contrast, in the LGE SAX sequence each frame corresponded to a depth. For LGE LAX, 2ch, and 3ch views, we replicated each view so that its total frame count matched that of the LGE SAX stack. Finally, we concatenated all frames into a single continuous sequence and applied uniform sampling to extract 64 frames, thereby producing a fixed-length input (Fig. 5).

R.2.6. The description of the CMR imaging protocol currently mixes technical and clinical aspects, which may confuse readers. I suggest separating technical details (e.g., scanner specifications, sequences) from clinical indications for better clarity.

We thank the reviewer for this suggestion although the clinical and technical aspects of imaging are closely correlated with respect to CMR. However, we have rewritten the results paragraph to improve clarity for readers by grouping clinical information and technical information in a more intuitive manner. This is copied below for the reviewer's convenience.

> *"*CMR-CLIP is a vision embedding trained in self-supervised manner using 13,786 paired CMR studies and reports from 12,500 patients collected between 2008 and 2023. The cohort consists of 6,133 females and 7,653 males, and an age range is 54.1±16.6. Additional details of general indications for CMRs, findings in the studies, and scanners used for imaging are included in Table 1. In total, the dataset included over 260,000 Cine videos of standard cardiac views including 4 chamber long axis (LAX), 2 chamber long axis (2ch), short axis (SAX), and 3 chamber long axis (3ch), and late gadolinium enhancement (LGE) images. Clips of each individual cardiac view and image type from a single study were undersampled and concatenated together. 95% of CMR studies contained between 400-500 Cine and LGE images, and for two standard deviations from the mean, and the number of words within the impression section ranges from between 95% of the impression sentences contained 30-150 words.*"*

R.2.7. Did the authors observe differences in model performance based on the scanner used (e.g., 1.5T vs. 3.0T)? This could be an important factor influencing generalizability.

We thank the reviewer for raising this question and have added a supplemental table to provide subgroup analysis based on vendor and field strength. There was no trends in any individual subgroups which would suggest the model does not generalize to unseen clinical situations. The results are copied below for the reviewer's convience.

> *"*CMR-CLIP demonstrated robust zero-shot generalizability across scanner vendors and field strengths (Supplemental Table 12-15). On internal Philips data (N=2,726), AUCs of the findings ranged from 0.681-0.850, while a small internal Siemens subset (N=32) yielded similar AUC (0.617-0.926) with greater variance. In an external Siemens cohort (N=428), AUC (0.594-0.836) remained consistent, indicating resilience to domain shift. Stratification by field strength on the internal cohort showed comparable performance at 1.5T (N=1,844; AUC 0.694-0.851) and 3.0T (N=914; AUC 0.652-0.854), and on the external cohort at 1.5 T (N=421; AUC 0.605–0.830), with a very small external 3.0T subgroup (N=7) showing wider metric variability."

R.2.8. More details are needed regarding the radiology report data used. For instance, did the authors focus only on the "impression" section? How did they handle variability in language across reports?

We agree with the reviewer that the details around text processing is not clear. We have expanded the *CMR Text Reports* subsection to describe our report-processing steps in detail:

> “We focus on the impression section, which summarizes each CMR study’s key findings, differential diagnosis, and provides recommendations for plan of care. We used a rule-based parser which identified subsection heading to extract the impression section from the radiology report. The impressions section was then cleaned to remove formatting such as bullet points, numbered lists, and communication boilerplates. We purposefully preserved synonyms and abbreviations to increase text variation and embedding generalizability. “

R.2.9. There is no mention of ethical approval, data privacy protections, or patient consent procedures. These should be explicitly addressed.

We thank the reviewer for pointing out this oversight. We have added an ethics statement to the Methods section.

> *“*The Cleveland Clinic Institutional Review Board waived the need to obtain informed consent for purposes of this study (IRB #22-1213).*”*

Response to Reviewers: NCOMMS-25-09811-T

CMR-CLIP: Contrastive Language Image Pretraining for a Cardiac Magnetic Resonance Image Embedding with Zero-shot Capabilities

We thank the editors and reviewers for their careful and insightful review and the opportunity to further address them. We have further clarified the test data and evaluation settings to provide readers with easier time with reproduction of the work.

Note: Reviewers' specific critiques are written in bold and the response provided below.

Specific response to reviewer's 2 critiques:

Reviewer Summary

This manuscript introduces CMR-CLIP, a vision–language foundation model that treats a complete CMR study as a video-like sequence and aligns it with the impression section of the clinical report via contrastive self-supervision. The authors pretrain on a substantial single-system cohort (13.8k studies), evaluate on an internal hold-out and an external site, compare against a video baseline and a task-specific amyloidosis model, test zero-shot/few-shot classification, cross-modal retrieval, transfer learning, robustness to perturbations, and report generation (retrieval-based and autoregressive).
I appreciate the replies to prior reviews: the added external site, the stability experiments, the generative evaluation, and the baseline comparisons materially strengthen the paper.
In my view the manuscript is already mature and potentially impactful, the modeling choice fits CMR's multi-view temporal nature, and the scale is noteworthy. To reach its full potential (generalizability, clinical utility, and reproducibility), I recommend the following focused additions/clarifications.

Conclusion
The manuscript is solid and already mature, and the authors' responses have substantially improved scope and rigor. With targeted additions the work would convincingly demonstrate generalizability and clinical utility.

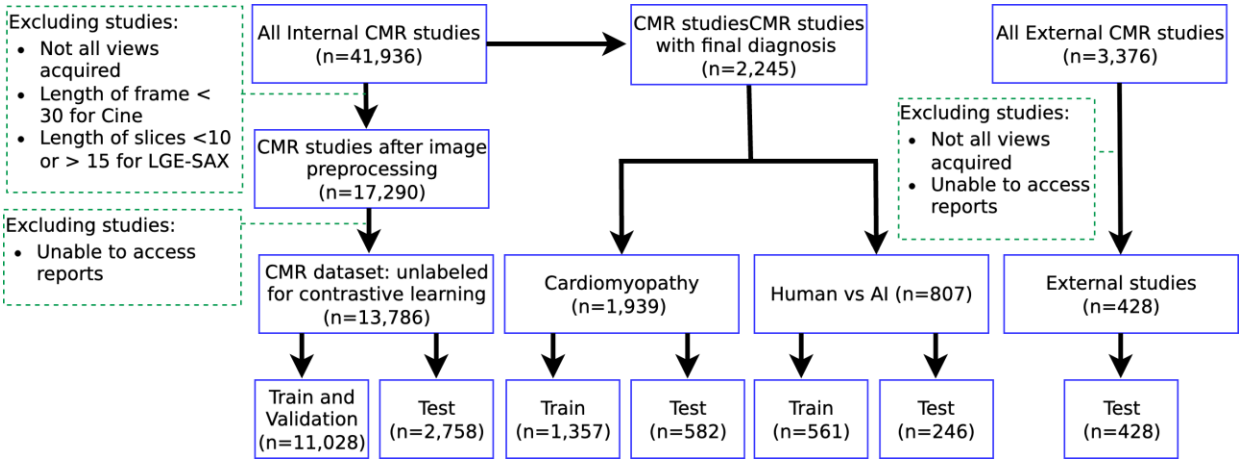
R.2.1. Explicitly confirm patient-level splits for pretraining/val/test and all downstream tasks.

We have added "we split the remaining data using a 70/10/20 split by patient" in the Methods. We further updated the cohort figure (now split to Figure 2 attached below) to describe the split.

R.2.2. For comparison cohorts (amyloidosis/HCM), confirm that no images or reports from those patients entered pretraining to avoid implicit text leakage. A simple data-flow diagram would help.

As the reviewer as suggested, we have added additional data flow in new Figure 2 to visualize how data was used for validation. To be clear, all test data was separate from

pretraining. However, 501 studies in Cardiomyopathy and Human vs AI cohorts have overlap. These were not used for validation. The data flow diagram is attached below:



R.2.3. The revised prompting improved LGE AUC, but LGE remains label-sensitive. Please consider adding subanalysis with curated LGE labels:

We strongly agree that validation of CMR-CLIP using more thoroughly curated LGE labels would be valuable. However, such analysis is not feasible at the current time given variability of documentation of LGE pattern in the report. It would require expert readers to re-interpret the studies with a standardized interpretation schema. The most consistent label is still based on final disease analysis and LGE positivity which we have provided in the initial analysis. However, we have added this limitation to the updated manuscript.

> "the LGE validation is based on documented patterns of LGE in the report. Such patterns are inconsistently documented[38] and may not accurately reflect the pattern in the actual images."

R.2.4. Specify the provenance of disease labels (final diagnosis vs report impression) and how potential mislabels were handled.

We agree with the reviewer that our validation datasets are not clearly described. We rewritten the methods to more clearly describe the different sources of labels. Specifically, the zero-shot and few shot data was done with data extracted from report impressions where as the supervised learning was done with data from final diagnosis. The details are recreated below:

> "Disease and clinical findings mentions in the impressions were extracted using a finetuned Large Language model Meta Artificial intelligence (LLaMA)[38].
>
> The data was split into training and test sets based on patient to prevent data leakage (splits illustrated in Fig. 2). We further leveraged previously acquired manually annotated data for testing purposes. The first such dataset is a cohort of patients with known cardiomyopathy diagnosis (CM dataset). The CM dataset includes 1,939 studies acquired on adult patients between 2008 and 2023. All patients were found to have either hypertrophic cardiomyopathy (HCM), cardiac amyloidosis (CA), undifferentiated non-ischemic cardiomyopathy (NICM), or ischemic cardiomyopathy (ICM). Final diagnosis was determined through manual review of a patient's medical records following relevant guidelines. HCM was identified based on the CMR imaging

biomarkers of left ventricular wall thickness >15mm, absence of abnormal loading conditions, and absence of infiltrative cardiomyopathies[39]. CA was determined by positive cardiac pyrophosphate scan (PYP) or biopsy confirmed amyloidosis. NICM defined as any case of heart failure with preserved ejection fraction, no evidence of ischemic process, and no specific heart failure etiology[40]. Definitive diagnosis of ICM was determined by ejection fraction <40% and either positive LGE or confirmed occlusion of a coronary artery[41]. Of this cohort, 1,357 studies were placed in the training data, while 582 were held out for testing. For human vs AI evaluation cohort, there are 807 studies, comprises of final diagnosis of CA, HCM, and other conditions. This dataset was identified as cases referred for evaluation of infiltrative disease from the documented indication for the exam. A subset of patients (N=501) overlapped with the CM dataset. The dataset was randomly partitioned into training (70%) and test (30%) subsets. For the test set (N=264), clinical reader confidence were manually reviewed[4]"

R.2.5. I would sugget to publish the full prompt list. Provide a prompt-sensitivity analysis (synonyms, negations, minor wording changes) and consider prompt ensembling/logit averaging.

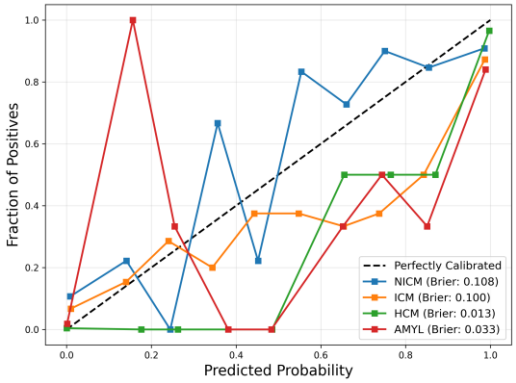
We have added a full prompt list in the supplemental materials.

Left ventricle dysfunction	Left ventricle with systolic dysfunction
	Depressed left ventricular function
Right ventricle dysfunction	Right ventricle with systolic dysfunction
	Depressed right ventricular function
Left ventricle dilation	Left ventricle is dilated
	Left ventricle is enlarged in size
Right ventricle dilation	Right ventricle is dilated
	Right ventricle is enlarged in size
Left ventricle hypertrophy	There is left ventricular hypertrophy
	There is septal hypertrophy
	LV is thickened
LGE	No enhancement and delayed imaging to suggest ischemic insult or infiltrate pathology
	There are no findings to suggest prior ischemic damage or an infiltrative process
Aortic dilation	Aortic root is ecstatic
	Ascending aortic is dilated
	Descending aorta is dilated

With respect to prompt-sensitivity, we currently do not have the necessary bandwidth to accomplish this in a comprehensive and scientifically driven manner as we cannot guarantee that LLM generated synonyms and negations are clinically correct. We also do not currently have access to a clinical fellow/clinical collaborator with the necessary bandwidth to do this manually.

R.2.6. I suggest adding in appendix Brier scores and calibration curves. If possible, include Decision Curve Analysis to quantify net benefit for key tasks.

We have added calibration curves for the supervised learning tasks. We have added this data (reproduced below) as figures in the supplemental data (Supplemental Figure 1).



We further agree with the reviewer that decision curve analysis would add significant value to this work. However, it is difficult to define given that decisions based on imaging are often multi-factorial and total value is not easily defined. Therefore, we believe such analysis is outside the scope of this work.

R.2.7. Complement ROUGE/BLEU with clinician-blinded factuality on key fields (e.g., EF category, LVH presence, LGE pattern) and quantify hallucination rates (are there hallucination in the experiment?).

We agree with the reviewer that factual validation of the generated reports would increase the rigor of the experiments. We leverage the same NER model used to extract entities from clinical generated impressions to identify [disease entities/findings] in our generated report. We compare the overall accuracy of the generated reports against the clinical reports. We find CMR-TARGET achieved an overall accuracy of 0.870. The results are added as a new row in Supplemental Table 11.

Label	Accuracy	F1-score
LVDys	0.892	0.799
RVDys	0.931	0.403
LVD	0.896	0.763
RVD	0.931	0.258
LVH	0.896	0.668
LGE	0.715	0.801
AoD	0.832	0.253

Although we agree with the reviewer that quantifying hallucination rates in generated reports would be useful, we believe this is outside the scope of the work given the difficulty to define hallucinations. The clinical reports themselves are not entirely accurate and hallucinations are not well defined. For instance, if the original report does not mention infiltrative cardiomyopathy, but does include HCM, would it be wrong for the generated report to mention both entities?

R.2.8. Include a brief privacy audit for text: de-identification steps and a membership-inference/nearest-neighbor check to rule out verbatim reuse of training impressions.

We have added description of de-identification steps in our data processing steps. Specifically, it includes extracting all impressions from the full report to remove either patient or physician identifiers, and running the impressions to identify any potential names or dates. Any impression identified with a potential name and/or date was manually reviewed and the identifying information was removed prior to training. The updated description is included below:

> ”We focus on the impression section, which summarizes each CMR study’s key findings, differential diagnosis, and provides recommendations for plan of care. We used a rules-based parser which identified subsection heading to extract the impression section from the radiology report. This has a secondary impact of de-identifying the text data given patient identifiers (name, dates, ordering, and interpreting physicians, etc.) are contained in sections outside of the impression. The impressions section was then cleaned to remove formatting such as bullet points, numbered lists, and communication boilerplates.”

We did a nearest neighbor check as suggested by the reviewer. We encoded all report impressions with a SentenceTransformer model. We then identified the highest similarity report in the training set for each test impression. We then computed standard automated text similarity metrics (BLEU-n, METEOR, ROUGE-L, CIDERr, BERTscore) between the retrieved and test impression. The low metrics reflect the large semantic and syntactic differences in reports between the test and training sets. This is expected given the impression section of the reports are not standardized, there are large numbers of different readers, and the wide variation in patients represented in our data. The results are shown below:

BLEU-1	0.569
BLEU-2	0.442
BLEU-3	0.363
BLEU-4	0.308
METEOR	0.278
ROUGE-L	0.463
CIDERr	0.692
BERTscore	0.781

R.2.9. Excellent that weights are released. Could be useful to add a model card with full training configs (hyperparameters, seeds, optimizer, batch size, temperature),

preprocessing (Impression parser), sampling of the 64 frames, and 3+ reruns with different seeds reporting mean ± SD?

As the reviewer as requested, we have updated the model card with training configs. It is attached below for convenience:

2. Configuration

Modify the configuration file `configs/cmr.json` with your training parameters:

Training Configuration

Parameter	Value
n gpu	4
optimizer	AdamW
learning rate (lr)	3e-5
loss	NormSoftmaxLoss
epochs	100
batch size	16
extraction resolution	256
input resolution	224
projection dimension	512
stride	1

Video Parameters

Parameter	Value
model	SpaceTimeTransformer
architecture config	base patch16 224
number of frames	64
pretrained	true
time initialization	Zeros

Text Parameters

Parameter	Value
model	Bio_ClinicalBERT
pretrained	true

We further provide some estimates of variability of our results by bootstrapping results. Unfortunately, re-running experiments is very costly and not feasible given our compute environment is shared with various different groups.

R.2.10. Editorial consistency: correct minor slips (“CMR-CONVIRT”, “aggravating” → “aggregating”, “compared of” → “comprised of”).

We have gone through further editorialization to correct such slips. Thank you for helping to identify these!

R.2.11. Figures/Tables: Good that images are now ≥300 dpi. Please standardize font sizes/scales, axis labels, and units. Expand abbreviations at first occurrence in Table 1 and captions (example T = Tesla, Y = years…=

We have updated figures to standardized font sizes, axis labels, and units as much as possible. We have also carefully editorialized the manuscript to expand first abbreviations.

R.2.12. Ethics/Privacy: The IRB statement is now present; briefly restate de-identification, and exactly what is released (weights, code; no data).

We have made the following updates to the methods: “The Cleveland Clinic Institutional Review Board waived the need to obtain informed consent for purposes of this study (IRB #22-1213). All identifying data was stripped from image data and associated reports. Code to train the model and weights of the model are publicly available. Cleveland Clinic images and reports are not available for public use.”