

Rapid patient-specific neural networks for X-ray to volume registration

Corresponding Author: Mr Vivek Gopalakrishnan

Any redactions in this file are there to maintain patient confidentiality, the confidentiality of unpublished data, or to remove third-party material.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

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)

This manuscript describes a method for single view, single object rigid 3D CT to 2D X-ray registration using patient-specific learning. Experiments include 2 public yet small benchmark datasets (pelvis and head/DSA) and an internal dataset closely related to the head/DSA dataset, which is larger in terms of subjects (50 compared to 6 and 10, respectively) but still small in terms of X-rays (122 total vs 366 and 20, respectively) and further without a fundamentally reliable source of pose ground truth (expert review).

Having worked on 3D/2D registration extensively in the past and still today, I empathize with some of the challenges that the method presented in this manuscript is seeking to overcome. It is true that, thinking back only about 10-ish years, the scientific progress of 3D/2D registration was severely hampered by the sheer unavailability of (usable?) open-source code packages. If your research group was not working on this and had built their own framework, then working on 3D/2D registration would deal a metaphorical killing blow to your progress simply because of the infrastructure that would need to be created to even consider solving the simplest problem setting of single-view, single-object rigid 3D/2D registration. With the open-source culture rejuvenated through deep learning, this situation has improved. Many frameworks that enable users to study 3D/2D registration now exist, and many more ideas on how to improve its notorious challenges (such as heavy parameter dependence, small capture range of optimizers especially in traditional non-ML formulations of the problem, etc.) have been proposed and evaluated, often showing drastic improvements over some of the more basic baselines. Indeed, the patient-specific modeling paradigm at the center of this manuscript is squarely among these innovations, presented in a much less sophisticated form by Esteban et al. [47] already in 2019 and then extended to patient-specific learning quite similar to the work here in the associated journal extension by Grimm et al. [R1] in 2021. Similarly, the authors themselves have presented the same idea several times before, e.g., in [38] and [53] (w.r.t. which the authors argue that this paper is a complete overhaul, but the core idea remains the core idea).

There is of course quite some merit to this work; the experiments are well designed and reported (though I would argue that given the wealth of prior work on machine learning-enhanced 3D/2D registration especially for this most simple task of single-view, single-object rigid the chosen baselines are decisively weak), the reduction in pre-training time is a reasonable engineering effort using transfer learning based on quite specific pretraining, and the demonstration that this paradigm can work on real data of varied anatomy is good (though the chosen anatomy is some of the easiest ones).

These strengths, however, are to be contrasted with some of what I wrote before and some other considerations. First, single-object, single view rigid 3D/2D registration is by far the easiest possible setting. It is commendable that with this method, we may now be able to move more reliably to adequate registration without the (sometimes cumbersome) finetuning of registration parameters; yet, many solutions to this problem (including patient specific ones) exist. While perhaps indeed this method may be better (the evidence is not provided to demonstrate clear superiority w.r.t. more recent ML-based methods, and doing so will be tough due to the specificity of this type of research) it is more of a logical continuation of prior work than a paradigmatic shift. It is also not clear whether the method is "future proof", in the sense that it may offer equally convincing opportunities for the more general 3D/2D registration problems, including those with multiple objects, multiple views, and compound or non-rigid motion. There are more and less obvious limitations towards making this work, and considerably less

(yet not none) literature has focused on how to approach these problems, including with machine learning. The practicality of the approach is further hampered by the fact that current pre-training is also anatomy specific, which qualifies the 5-minute claim since sufficient data must be available for pre-training the models. Recent foundation model approaches, e.g. [R3] and associated datasets, also leveraging data synthesis, may help overcome this challenge and should be considered. There are additional shortcomings that overall rank lower for me but should be noted, such as for example, the limitation that X-ray synthesis is being described as realistic but seems to be considerably incomplete, as the model considers single material, single energy absorption (there are several sentences in the manuscript, such as the one stating that appearance differences are primarily due to geometric distortions, which are certainly incorrect). The benefits of this renderer is differentiability, but more realistic open-source rendering frameworks such as DeepDRR exist, and should or at least could be considered for data synthesis of the target 2D x-rays to further enhance model robustness. In addition, while I commend the user-based study to establish a semblance of registration accuracy in the in-house dataset, it is known that such techniques fail to capture deviations in registration quality when the registration errors are low [R2]. The conclusion may still hold that the presented framework performs well on this data, however, the comparisons etc. should likely be treated with care.

In conclusion, the method and this manuscript are well developed, and perhaps indeed useful in some way; yet, from my perspective they do not offer a fundamentally different paradigm, clear departure from prior work, or universally useful solution for 3D/2D registration, which limit my enthusiasm for this specific venue. I can see opportunities in a method/code framework focused venue where the utility offered here (namely a potentially quite useable framework) outweighs the more modest scientific contribution.

[R1] Grimm, M., et al. (2021). Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration. IEEE transactions on medical imaging, 40(9), 2221-2232.

[R2] Cho, Sue Min, et al. "Visualization in 2D/3D registration matters for assuring technology-assisted image-guided surgery." International journal of computer assisted radiology and surgery 18.6 (2023): 1017-1024.

[R3] Killeen, Benjamin D., et al. "Fluorosam: A language-aligned foundation model for x-ray image segmentation." arXiv preprint arXiv:2403.08059 (2024).

(Remarks on code availability)

I confirmed the repository is available but have not attempted running the code.

Referee #2

(Remarks to the Author)

This interesting manuscript proposes the use of patient-specific neural networks for intraoperative x-ray to CT volume registration. More specifically, it proposes the use of these networks for patient-specific PRE-registration, followed by refinement using a more conventional 2D/3D registration method to achieve a final registration result. The patient-specific registration network is trained via transfer learning from a previously trained patient-agnostic network derived from the CTPelvic1K dataset, and training times on the order of 5 minutes are reported. Pre-registration errors on pelvis images of $\sim 67.4 \pm 20$ mm are reported and 3D target registration errors of ~ 1 mm and are reported after the refinement step for manually designated "fiducials" at designated anatomic landmarks from what seems to be single x-ray projections. Sub-mm 3D registration errors are achieved in 41-44% of cases in their test data set. Impressive reprojection errors ($mPE = \sim 18$ mm, $mRPE = \sim 6$ mm) are also reported. The authors also report the application of their method on DSA volumes for brain angiography, with roughly comparable results.

Overall, I found this to be an interesting and original approach, but I do have some questions that ought to be addressed before final publication.

1. Single-image 2D-3D registration results can be extremely sensitive to the accuracy of the C-arm intrinsic parameters. There should be some discussion of this issue and (ideally) some sort of simulation experiment to assess this sensitivity.
2. Although the use of manually-identified "fiducials" only appear EXPLICITLY in the evaluation of the results, these points are chosen to correspond to readily-identified anatomic features, which are places where the 2D and 3D images have significant geometric information and local image contrast. Thus, there may be some IMPLICIT bias in an evaluation based on these points. It would be useful to see some discussion of this point. Perhaps added discussion of the dGEO results would help.
3. It would also be very useful to see the standard deviations for each of the results reported in table S1.
4. One thing that I noticed is that the post-refinement results in Table S1 for the authors' three xvr methods (patient-agnostic, patient-specific, finetuned) are essentially the same, one suspects within the limits of statistical significance. Thus, what are the actual advantages of using any form of patient-specific training vs. patient-agnostic training only, followed by refinement. It may be that a more powerful statistical analysis (paired t-test, perhaps) could demonstrate this. In any case, this point should be addressed.
5. The statement "...achieving submillimeter-accurate registration at intraoperative speeds and improving upon existing methods by an order of magnitude" seems like an over-claim and perhaps should be modified, partly because of the reasons mentioned in my point 4. All the sub-mm results also rely on a further refinement step (M.6) and only achieve sub-mm results

in less than 50% of cases. These limitations need to be discussed, and (again) the relative advantages of the patient-specific training need to be demonstrated more convincingly.

6. It would be useful to see the times for the pose-refinement step, as well as the patient-specific training step in the reported results. How do these compare with the different methods?

7. Relative to times, "5 minutes" is only meaningful if one can say more about the computational hardware involved. How large are the networks involved? How many GPUs? Which ones? Etc.

8. Although the "fiducials" are manually-selected, it wasn't entirely clear whether this was done on the 3D volumes directly or on 2D projections. I think these may be the same landmarks that the authors used for their "landmark initialization" comparison, but it would be useful to see a comparison.

9. One general concern, common to most medical image analysis papers, is that reporting "success" rates based on thresholds or even the distributions based on experimental results on sample data somewhat begs the question of what the accuracy is for a particular, previously unseen, patient. It is probably unfair to expect the authors to address this issue as a condition for publication of this particular paper, but any insights that they can contribute would be welcome, even as a possibility for future research.

To summarize, I think that this is an interesting paper that reports a good idea. The authors have done a good job describing their method in sufficient detail to allow others to replicate and adapt their results. I look forward to seeing the revised version.

(Remarks on code availability)

I did not have time to review the code, but I think that the paper's description of the method

Version 1:

Reviewer comments:

Referee #2

(Remarks to the Author)

The authors have answered my questions and have significantly improved their interesting manuscript. I think that it is fine to publish.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

Review of "Rapid patient-specific neural networks for intraoperative X-ray to volume registration" by Gopalakrishnan et al.

1. Overall, the revisions are responsive and the manuscript is considerably stronger than the original submission. I commend the authors for conducting new analyses interrogating the generalizability of xvr, releasing model weights on HuggingFace, running additional sensitivity analyses, and increasing the clinical rigor of the evaluations. There are nonetheless still some concerns I have about generalizability that I list below.

2. The baseline comparisons to other methods are improved but still quite limited for a paper in a top journal. In particular, the authors write that "Comparisons of the reported results on the same anatomical region show that xvr achieves the most accurate and reliable registrations across all previous methods on a common anatomical structure." yet this claim rests on comparisons of xvr to only two comparators, Gao et al. and Grimm et al., on just one dataset based on only 6 cadaveric CTs (DeepFluoro), and that too on non-identical data (as the authors acknowledge). No formal test of statistical significance is reported, and given that the 362 X-rays in DeepFluoro are not independent (many are from the same subject), the effective sample size is much smaller than it appears. The superiority claims in the results and discussion should be softened to reflect these limitations.

3. The core contribution of patient-specific finetuning remains inadequately demonstrated. In particular, the results for the three xvr variants (finetuned, de novo, foundation) presented in Table S1 reveal that the finetuned and de novo models achieve nearly identical performance across DeepFluoro and Ljubljana. Both attain a mTRE of 0.9(1.1) mm and 41.4%/40.9% submillimeter success rate % on DeepFluoro, and the two methods achieve identical mTRE and submillimeter success rate % on Ljubljana. Is the advantage of patient-specific finetuning over de novo training exclusively one of speed (5 min vs. 12 h) and not accuracy? If so, this should be made more explicit. Additionally, the comparisons between these three variants in these two datasets rest on only 6 and 10 unique subjects, making it difficult to draw firm conclusions about when and by how much finetuning helps.

4. Related to point 3, while the authors argue that "...rapid patient-specific finetuning eliminates the out-of-distribution failures

that plague previous methods when encountering patients not well-represented in their training sets, ensuring reliable performance for every individual undergoing an intervention.” the paper does not actually test this hypothesis rigorously. The authors do not test the finetuned model on anatomical structures or patient populations (e.g., pediatric, severely pathological) that are underrepresented in the pretraining corpus, which may be precisely where the foundation-then-finetune paradigm would need to succeed to support the generalizability claims being made.

5. The framing in the introduction and discussion are improved, but the presentation should still be toned down given the issues above.

(Remarks on code availability)

Well-documented README included with installation and use instructions, I was easily able to install the package (xvr) with the standard Python package manager (pip) from the command line.

Version 2:

Reviewer comments:

Referee #2

(Remarks to the Author)

The authors have further improved their paper in response to Reviewer 3's extensive comments. I think that this paper is fine to publish.

(Remarks on code availability)

Referee #3

(Remarks to the Author)

The authors have done a solid job responding to my comments; the results are more robust and the presentation has improved considerably. I commend the authors for their efforts in these revisions and have no further comments at this point.

(Remarks on code availability)

Please see my comments from the previous round of revisions -- the authors made it easy to install & run their tool.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license,

unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Dear Editors and Reviewers:

We deeply appreciate the reviewers for their thorough and insightful feedback on our work. We are glad to hear that they found our method, manuscript, and associated software package to be “*well developed and perhaps indeed useful*” with “*a potentially quite usable framework*” (R1), and our approach to be “*interesting and original*” with “*impressive reprojection errors*” and strong overall performance (R2). Both reviewers appreciated the extensive evaluation showing that our 2D/3D registration method, **xvr**, generalizes to real clinical data from diverse anatomical regions and patients.

The reviewers raised important concerns that we have addressed throughout the revision. R1 questioned whether our approach to 2D/3D registration offered clear advantages over existing solutions, and whether **xvr** would be able to generalize to multiple anatomical structures. R2 raised technical questions about the sensitivity of **xvr** to the misestimation of intrinsic parameters, along with requests for additional lower-level details.

We have addressed and incorporated all of the reviewers’ feedback, substantially strengthening the manuscript with the following new contributions, where **(R1)** and **(R2)** indicate points addressed to Reviewers 1 and 2, respectively:

(R1) A foundation model for 2D/3D registration (Fig. 2): We have developed a pretrained pose regression model for the entire human body, making rapid initialization accessible for any anatomical region without application-specific training. The model was trained on synthetic X-rays generated in 360° viewing directions from a multi-dataset aggregation of over 2,000 rigidly co-registered CT volumes covering the whole body. This foundation model is adaptable to any rigid structure imaged at any angle. **We show that, with only 5 minutes of finetuning, this single model generalizes to new subjects and anatomical regions, achieving the highest and most consistent registration accuracy of all tested methods.** We have also released the weights alongside our open-source package.

(R1) New dataset evaluation (Fig. 3): We have further validated the ability of our foundation model to generalize to novel anatomical regions by conducting a comprehensive analysis of a new dataset containing paired X-ray and CT images of the femur, which was released after our initial submission.¹

(R1) Expanded baseline comparisons (Fig. 4): We have included comparisons to two additional previous state-of-the-art 2D/3D registration methods, including one suggested by R1, further supporting the improved registration accuracy and generalizability of **xvr**.^{2,3}

(R1) Robust ground truth validation (Fig. 5): To improve the analysis of our internal clinical dataset, expert clinicians spent over 40 hours manually annotating 2,000+ pairs of 2D/3D keypoints, which we used to establish reliable ground truth C-arm poses. In this revision, we have replaced our previous expert user study with a quantitative evaluation.

(R2) Sensitivity analysis (Fig. S4): We conducted a new simulation study to systematically evaluate the robustness of our method to errors in recorded C-arm intrinsic parameters.

(R2) Additional reporting: We have included specific details as requested, including the runtime of all pose-refinement methods (Fig. 4A-C), IQRs for all registration error metrics (Tab. S1), and all computational resources used in the experiments (Methods M.9).

¹ Flepp, R., Nissen, L., Sigrist, B., Nieuwland, A., Cavalcanti, N., Fürnstahl, P., Dreher, T. and Calvet, L., 2025. Automatic multi-view X-ray/CT registration using bone substructure contours. *International Journal of Computer Assisted Radiology and Surgery*, pp.1-8.

² Grimm, M., Esteban, J., Unberath, M. and Navab, N., 2021. Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration. *IEEE Transactions on Medical Imaging*, 40(9), pp.2221-2232.

³ Gao, C., Feng, A., Liu, X., Taylor, R.H., Armand, M. and Unberath, M., 2023. A fully differentiable framework for 2D/3D registration and the projective spatial transformers. *IEEE Transactions on Medical Imaging*, 43(1), pp.275-285.

These contributions address all questions raised by both reviewers and demonstrate that our method provides a superior solution for automatic 2D/3D rigid registration. Our expanded experiments show that the revised **xvr** framework reliably and accurately generalizes out-of-the-box to any subject and anatomical region without requiring manual labels or hyperparameter tuning—achieving a broadly applicable solution to this challenging open clinical problem. We are grateful to the reviewers and editors for helping us to create a more robust framework for 2D/3D registration and improve the manuscript.

All revisions made to the manuscript have been annotated with blue font.

Response to the Editor

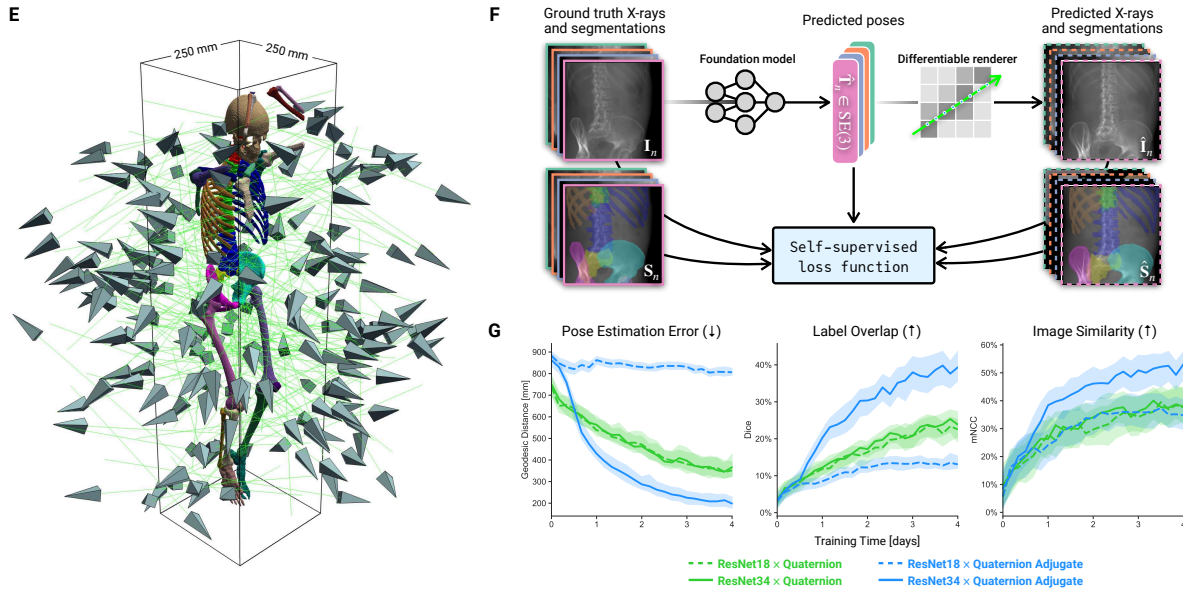
We would ask that the revised study fully address the issues raised by the reviewers, including the concerns raised by reviewer #1 regarding the question as to why this solution to registration without finetuning is superior to other proposed solutions, and about the generalizability of the model for use with multiple anatomies.

In the revised manuscript, we have fully addressed all issues raised by the reviewers with new experiments, expanded analyses, and clarifications to the text. With regards to the specific concerns raised by Reviewer #1, we have now shown that our solution for registration (without finetuning) is superior to other previously proposed methods **precisely because it can automatically and accurately generalize to any anatomy**.

Previous state-of-the-art registration methods, including baselines already included in our experiments [53] and those suggested by Reviewer #1,⁴ are keypoint-based approaches that require trained experts to perform time-consuming manual annotation of anatomical landmarks on training CTs for each new structure. In contrast, **xvr** removes this annotation bottleneck by simulating its own training data, enabling it to efficiently scale and generalize to new anatomical structures and clinical settings. In the revised manuscript and response letter, we include additional comparisons to these and other previous methods,⁵ demonstrating that both the foundation and finetuned models in **xvr** achieve more accurate and consistent results (Fig. 4).

To further address the reviewer's concerns regarding generalization to multiple anatomical regions, we have strengthened our contributions by developing a region-agnostic foundation model for 2D/3D registration. Trained on thousands of CT scans with whole-body coverage, this model automatically adapts to diverse anatomical structures across the human body (see Fig. below). Additional experiments in this revision show that the proposed foundation model can be successfully finetuned to new datasets and anatomical regions from unseen patients and medical centers with only 5 minutes of self-supervised training (Fig. 3).

To summarize, **xvr** provides a fundamentally different solution to 2D/3D registration: it requires no manual annotations, achieves state-of-the-art accuracy, and generalizes automatically to novel anatomical structures through a single open-source foundation model accessible via an easy-to-use pip-installable Python package.



⁴ Grimm, M., Esteban, J., Unberath, M. and Navab, N., 2021. Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration. *IEEE Transactions on Medical Imaging*, 40(9), pp.2221-2232.

⁵ Gao, C., Feng, A., Liu, X., Taylor, R.H., Armand, M. and Unberath, M., 2023. A fully differentiable framework for 2D/3D registration and the projective spatial transformers. *IEEE Transactions on Medical Imaging*, 43(1), pp.275-285.

Response to Reviewer 1

1. This manuscript describes a method for single view, single object rigid 3D CT to 2D X-ray registration using patient-specific learning. Experiments include 2 public yet small benchmark datasets (pelvis and head/DSA) and an internal dataset closely related to the head/DSA dataset, which is larger in terms of subjects (50 compared to 6 and 10, respectively) but still small in terms of X-rays (122 total vs 366 and 20, respectively) and further without a fundamentally reliable source of pose ground truth (expert review).

We agree that the sample sizes of these datasets are limited relative to other biomedical imaging problem spaces (e.g., segmentation). However, we emphasize that the presented experiments represent the largest and most anatomically diverse evaluation of a 2D/3D registration algorithm on real X-ray and CT images to date. At the time of submission, the two public datasets we analyzed (DeepFluoro and Ljubljana) were the only openly available sources of real X-ray/CT image pairs from human subjects, making these evaluations as large and diverse as possible on public datasets for comparison to prior work.

After submission, an additional public dataset comprising paired and calibrated X-ray and CT images of human femurs was released.⁶ In this revision, we have now expanded our evaluations to include these 49 distal femur X-rays from 5 subjects (see Figs. 3 and 4 in the revised manuscript). Consistent with our results on the pelvis, neurovasculature, and skull, **xvr** generalizes out-of-the-box to subjects in this novel dataset. Finetuning the new whole-body 2D/3D registration foundation model we present in this revision for only five minutes per subject enabled **xvr** to achieve a median mTRE of 8.0 mm (IQR: 5.7–11.4). **xvr** is robust and generalizes to novel anatomical regions and subjects **without requiring any manual annotations**, demonstrating its utility as a generic solution for the rigid 2D/3D registration problem.

Regarding the internal dataset (Brigham), we agree that the lack of reliable ground truth limited our previous analysis of **xvr**'s performance in this setting. To address this concern, our clinical collaborators manually annotated corresponding 2D and 3D keypoints on each of the 122 X-ray and CT image pairs in our dataset, and we estimated ground truth C-arm poses using the PnP algorithm [76]. Quantitative comparisons with ground truth keypoints show that **xvr** achieves a median mTRE of 1.2 mm (IQR: 0.5–1.9), further demonstrating **xvr**'s robustness in real-world clinical settings (Fig. 5).

Importantly, manually annotating these keypoints required over 40 hours of work by clinicians. In contrast, **xvr** automatically registered the entire dataset of 122 X-rays in just 10 minutes on a relatively weak consumer GPU (NVIDIA GeForce RTX 3060; 12 GB VRAM).

2. Having worked on 3D/2D registration extensively in the past and still today, I empathize with some of the challenges that the method presented in this manuscript is seeking to overcome. It is true that, thinking back only about 10-ish years, the scientific progress of 3D/2D registration was severely hampered by the sheer unavailability of (usable?) open-source code packages. If your research group was not working on this and had built their own framework, then working on 3D/2D registration would deal a metaphorical killing blow to your progress simply because of the infrastructure that would need to be created to even consider solving the simplest problem setting of single-view, single-object rigid 3D/2D registration. With the open-source culture rejuvenated through deep learning, this situation has improved. Many frameworks that enable users to study 3D/2D registration now exists, and many more ideas on how to improve its notorious challenges (such as heavy parameter dependence, small capture range of optimizers especially in traditional non-ML formulations of the problem, etc.) have been proposed and evaluated, often showing drastic improvements over some of the more basic baselines.

We deeply appreciate the reviewer's perspective and experience on this point. However, we disagree with the reviewer's position that the rejuvenation of open-source science has solved this problem.

⁶ Flepp, R., Nissen, L., Sigrist, B., Nieuwland, A., Cavalcanti, N., Fürnstahl, P., Dreher, T. and Calvet, L., 2025. Automatic multi-view X-ray/CT registration using bone substructure contours. *International Journal of Computer Assisted Radiology and Surgery*, pp.1-8.

Many papers have been written on 2D/3D registration in the past 10 years, with some releasing at least partial codebases (e.g., [ProST](#), [CMAES-reg](#), [ITK](#)). However, our experience is that most of these repositories are research prototypes or unmaintained software, not modern practitioner-oriented tools: they often assume project-specific data formats, lack pre-trained models, or require non-trivial C++/CUDA build chains. As a result, we were not able to apply them to real clinical X-ray images without considerable engineering effort and complete re-implementation in some cases. From the standpoint of a practitioner who wishes to use 2D/3D registration on their own data without rebuilding an entire pipeline from scratch, this remains a substantial barrier.

Among the available frameworks, **xReg** (github.com/rg2/xReg) is, to the best of our knowledge, the only general-purpose 2D/3D registration package that we could install and run on real clinical data with reasonable effort (although its C++ build is very complicated and requires substantial system-specific modification). However, **xReg** is a traditional optimization-based framework and is therefore subject to the well-known challenges the reviewer highlights (sensitivity to initialization and parameter tuning, limited capture range, etc.); it even describes itself as “*an initial baseline for evaluating the neural network approaches du jour*.” Our extensive evaluations of **xReg** and the keypoint-based neural networks built upon it (Grupp *et al.*, *IJCARS* 2020) [53] clearly demonstrate the improved performance of our image-based patient-specific framework, both in terms of registration accuracy and generalization to new anatomical regions (Fig. 4).

To further address the reviewer’s question about how our method compares to prior 2D/3D registration approaches, we have compiled performance metrics reported in the previous publications and used them to benchmark **xvr** and **xReg** on pelvic data (see [Comment R1.6](#) for the full results). For these methods, the code and/or trained models were not publicly available or could not be run successfully, so we were unable to replicate their results on our own datasets and instead had to rely on the performance statistics reported in the original papers. This analysis confirms that **xvr** is more accurate than existing methods for 2D/3D registration on pelvic data, to say nothing of the fact that these previous methods cannot be used on real-world data without excessive effort on the part of downstream users.

For these reasons, we argue that the need the reviewer identifies for a universal practitioner-friendly solution for automatic 2D/3D registration—especially one with machine learning capabilities that enable out-of-the-box generalization to novel patients and procedures—remains unmet. **xvr** and its accompanying easy-to-use pip-installable software package directly addresses this open problem with high accuracy.

3. Indeed, the patient-specific modeling paradigm at the center of this manuscript is squarely among these innovations, presented in a much less sophisticated form by Esteban et al. [47] already in 2019 and then extended to patient-specific learning quite similar to the work here in the associated journal extension by Grimm et al. [R1] in 2021.

Thank you for pointing us towards the journal extension [R1]. We have included a reference to it in the Introduction and a comparison of its performance relative to **xvr** on pelvic X-rays in the Results (Fig. 4E). This comparison, fully reported in the response regarding baselines (see [Comment R1.6](#)), shows that **xvr** outperforms [R1]. To clarify, our formulation of the patient-specific learning paradigm is **self-supervised**, while [47] and [R1] are **semi-supervised**, a fundamental difference that enables **xvr** to generalize to new anatomical structures and interventional settings with far more ease.

Specifically, [47] and [R1] are keypoint-based registration methods. They are innovative in that they remove the requirement that keypoints be manually annotated for every new CT. However, to achieve this, these methods still require that a substantial number of keypoints be manually annotated on a large training set (e.g., more than 10 CTs for the experiments presented in [47, R1]). Unlike the pretrained 3D segmentation networks that we freely leverage (e.g., TotalSegmentator), there are no pretrained whole-body keypoint detection models. **Therefore, contrary to the papers’ claims, these methods are not truly automatic.** Indeed, they are cumbersome to adapt to new anatomies as a domain expert must manually annotate 3D keypoints on a new set of training CTs every time one wants to train a model for a novel structure.

In contrast, **xvr** does not use keypoints. Rather, it regresses the C-arm’s pose directly from X-ray images. This removes the need for manual annotations as the training data is generated automatically on-the-fly from routinely acquired preoperative volumetric imaging **without the need for any manual annotations**. Given any CT scan, a user can use our framework to automatically finetune a patient-specific pose regression model in just 5 minutes with a single command. In fact, our evaluation of the Ljubljana dataset (which does not include any 3D segmentations) demonstrates that **xvr** excels even without segmentation supervision. We clarified this point in the Introduction:

...supervised keypoint estimation models require expert knowledge of anatomical landmarks that are reliably visible in procedure-specific fluoroscopy images, as well as manual annotation of these 3D structures for training and inference in every new preoperative scan [53]. Even semi-supervised approaches that propose to transfer 3D landmarks from previous patients to new ones require a large training corpus of manually annotated scans [47, R1], meaning training such methods is not truly automatic. Moreover, these approaches remain constrained to the specific anatomy and procedure for which they were developed, restricting their clinical utility. This dependence on extensive manual annotation by clinical experts fundamentally limits the scalability and generalization of existing (semi-)supervised approaches, precluding their deployment as universal solutions for intraoperative 2D/3D registration.

4. Similarly, the authors themselves have presented the same idea several times before, e.g., in [38] and [55] (w.r.t. which the authors argue that this paper is a complete overhaul, but the core idea remains the core idea).

We respectfully note that Nature explicitly allows submissions that build on prior conference proceedings.⁷ While our short workshop [38] and conference [55] papers introduced individual components (differentiable X-ray rendering and patient-specific synthetic pretraining, respectively), they are only a subset of the core ideas of the present work and are not sufficient for practical use on their own. This manuscript introduces several scientific insights and methodological advances necessary to make **xvr** a generic, off-the-shelf solution that integrates easily into clinical workflows and enables accurate registration of hundreds of X-rays in minutes. Below, we highlight the specific contributions that make this possible:

1. **Reducing training time from hours to minutes.** We develop an automatic and robust pretraining strategy that dramatically accelerates patient-specific training times from 12 hours in [55] to just 5 minutes. Removing this bottleneck makes patient-specific pose regression networks usable in emergency settings and on large multi-subject datasets.
2. **Introducing a foundation model for 2D/3D registration.** Thanks to this reviewer’s suggestion to generalize our model to multiple anatomical structures (**Comment R1.7**), we trained a pose regression model on synthetic X-rays from over 2,000 CTs with whole-body coverage. Future users can now finetune or directly apply this foundation model to any scan, instead of collecting and training on large task-specific CT datasets themselves, greatly improving the practicality of **xvr**.
3. **Allowing users to specify clinically interpretable C-arm pose distributions.** In the Methods, we highlighted a change from [55] in our parameterization of camera poses from the conventional $[R|t]$ to $[R|Rt]$ (see Eq. 4). This subtle change has a dramatic effect on usability. Specifically, switching from rotate-then-translate to translate-then-rotate enables users to specify clinically interpretable distributions over C-arm poses using radiologic conventions (RAO/LAO, CRA/CAU; Fig. 1C). This makes it significantly easier for clinical users to encode routine acquisition protocols through our command-line interface.
4. **Adapting to variable intrinsic parameters.** A major technical challenge precluding clinical deployment that we solved in this work is accommodating changing intrinsic parameters. For example, clinicians

⁷ <https://www.nature.com/nature/editorial-policies/preprints-conference-proceedings#conference-proceedings>

frequently change the intrinsic parameters of the C-arm (e.g., zoom, panning, collimating, etc.), but neural networks for 2D/3D alignment are trained using synthetic images generated with fixed intrinsic parameters (including our own [55] and previous baselines). Without modification, these models would fail at test time under practical, clinical usage. To overcome this limitation, we developed a simple image processing pipeline composed purely of morphological operations to resample acquired images with a fixed set of intrinsics (Fig. S3). This modification makes the resampled images in-distribution for any neural network, including both our pose regression networks and keypoint identifiers used in the baselines, demonstrating applicability beyond our work.

5. **Extended evaluation.** Compared to [55], we substantially extend the results, showing that our finetuned foundation model achieves the most accurate and consistent registration across four datasets, including two not evaluated in the conference paper.

5. There is of course quite some merit to this work; the experiments are well designed and reported (though I would argue that given the wealth of prior work on machine learning-enhanced 3D/2D registration especially for this most simple task of single-view, single-object rigid the chosen baselines are decisively weak), the reduction in pre-training time is a reasonable engineering effort using transfer learning based on quite specific pretraining, and the demonstration that this paradigm can work on real data of varied anatomy is good (though the chosen anatomy is some of the easiest ones).

We are glad to hear that the reviewer found the design and reporting of our experiments to be meritorious. We also agree that coverage across anatomies is important and wish to emphasize that we evaluated **xvr** on every publicly available source of real X-ray and CT images from human subjects that we could identify. This includes new evaluations of the Femur dataset (released after our original submission), in which we find that **xvr** generalizes off-the-shelf to this data without modification. Additionally, we have organized these datasets into a single repository in standardized and documented DICOM format for future researchers.

From a methodological standpoint, our original baselines (**xReg** and the keypoint-based networks built upon it) were chosen because they are, to our knowledge, the only general-purpose 2D/3D registration frameworks that can be reliably installed and run on real clinical data for the single-view rigid setting. While we fully acknowledge the wealth of prior work on ML-enhanced 2D/3D registration, very few of these methods provide code and trained models that can be directly applied to clinical data without substantial re-implementation and engineering effort. Thus, while we concur that there is substantial prior work, we respectfully disagree that our initial baselines were “*decisively weak*”; rather, they reflected the strongest reproducible and clinically applicable methods available for this problem at the time of submission. If there are other public datasets or reproducible baselines the reviewer would like us to include, we would be very happy to evaluate them as well.

Finally, inspired by a later suggestion from this reviewer (**Comment R1.7**), we now demonstrate that our pretraining strategy is not anatomy-specific but can be scaled to a whole-body foundation model for pose regression, enabling out-of-the-box generalization to many rigid structures across the human body. This addresses the concern that our reductions in pre-training time rely on “*quite specific pretraining*” and further supports the practicality of the proposed patient-specific paradigm on diverse real-world data.

6. These strengths, however, are to be contrasted with some of what I wrote before and some other considerations. First, single-object, single view rigid 3D/2D registration is by far the easiest possible setting. It is commendable that with this method, we may now be able to move more reliably to adequate registration without the (sometimes cumbersome) finetuning of registration parameters; yet, many solutions to this problem (including patient specific ones) exist. While perhaps indeed this method may be better (the evidence is not provided to demonstrate clear superiority w.r.t. more recent ML-based methods, and doing so will be tough due to the specificity of this type of research) it is more of a logical continuation of prior work than a paradigmatic shift.

It is also not clear whether the method is “future proof”, in the sense that it may offer equally convincing opportunities for the more general 3D/2D registration problems, including those with multiple objects, multiple views, and compound or non-rigid motion. There are more and less obvious limitations towards making this work, and considerably less (yet not none) literature has focused on how to approach these problems, including with machine learning.

We are primarily focused on one task: reliably recovering the pose of an X-ray image relative to a preoperative 3D volume for any patient or anatomical structure. As we have shown throughout our experiments (Fig. 4A-E, Tab. S1), this is far from a solved problem, both in terms of the accuracy of recovered C-arm poses, generalizability to diverse clinical scenarios, and ease-of-use. Furthermore, rapidly and reliably identifying the C-arm’s pose is a prerequisite step for most of the tasks that the reviewer listed, such as non-rigid 2D/3D registration or 3D reconstruction. Therefore, the lack of a robust solution to this problem is still a substantial barrier to widespread adoption and incorporation of 2D/3D registration across medicine and surgery.

In addition, we now explicitly address the reviewer’s concern about evidence of superiority over more recent ML-based methods by adding comparisons to the following registration baselines: Grimm *et al.*, 2021 [R1] and Gao *et al.*, 2023 [43]. The reviewer is correct that comparing these methods to **xvr** is challenging, as Grimm *et al.* did not release code for their method and Gao *et al.*’s [43] GitHub repository only provides a script for a toy example with synthetic data and hardcoded initial pose estimates. Nevertheless, we were able to establish meaningful comparisons based on the published evaluations.

Both papers evaluate their methods on real pelvic X-rays, with Gao *et al.* [43] using the same pelvic dataset as our study. We evaluated **xvr** and the **xReg**-based baselines using the same metrics as the original publication, and compared our performance against the previously reported results. These empirical comparisons demonstrate that **xvr** achieves more accurate and reliable registration than previous methods on a common anatomical structure (see the table below), while also eliminating the need for manual annotations and improving generalization across datasets. We acknowledge that these are not one-to-one comparisons, as we do not have access to the original code, models, or data/splits. Accordingly, these reproduced results are indicated by asterisks (*) and are accompanied by appropriate caveats in the text:

Gao et al. [43] also evaluate their method using the DeepFluoro dataset, but calculate mTRE using points on the surface on 3D bone segmentations. Grimm et al. [R1] evaluate a different pelvic dataset that is not publicly available.

	Transl. error [mm]	Rot. error [degrees]	mTRE [mm]	Transl. error $\leq$ 30 mm	mTRE $\leq$ 10 mm
Fixed init. (xReg) [45]	102.2 $\pm$ 78.1	32.7 $\pm$ 22.8	352.2 $\pm$ 239.3	22.9%	20.7%
Landmark reg. (xReg) [53]	38.7 $\pm$ 79.3	5.5 $\pm$ 11.1	82.1 $\pm$ 162.0	75.1%	70.2%
Landmark init. (xReg) [53]	28.6 $\pm$ 127.1	7.3 $\pm$ 22.6	79.6 $\pm$ 238.7	86.7%	80.1%
Fixed init. (xvr) [38]	101.1 $\pm$ 66.5	26.3 $\pm$ 13.5	340.5 $\pm$ 170.1	15.3%	10.8%
Gao <i>et al.</i> , 2023 [43]*	33.9 $\pm$ 70.1	13.9 $\pm$ 30.5	45.0 $\pm$ 88.1	Not reported	73.2%
Grimm <i>et al.</i> , 2021 [R1]*	4.2 $\pm$ 4.0	1.1 $\pm$ 1.3	4.2 $\pm$ 3.9	86.8%	Not reported
Finetuned model (xvr)	1.2 $\pm$ 2.5	0.2 $\pm$ 0.4	2.6 $\pm$ 5.1	100%	95.0%

Regarding novelty, we have clarified in the manuscript that the central conceptual contribution of **xvr** is to provide accurate single-view rigid 2D/3D registration without any manual annotations, in contrast to previous methods, including prior patient-specific approaches that rely on annotated correspondences or

extensive parameter tuning [47, R1]. We fully agree that more general 3D/2D registration problems—multiple objects, multiple views, and compound or non-rigid motion—are both highly relevant and less well explored. Therefore, we have emphasized that our registration framework is very flexible and can easily be extended to more general use cases in the revised Discussion:

*Our framework is readily extensible to several important 2D/3D registration scenarios. First, instead of regressing the pose of a single structure, the pose regression network could be adapted to predict poses of multiple rigid bodies, enabling the estimation of compound, polyrigid transformations [91]. Second, **xvr** is directly extensible to multiview rigid registration by incorporating a differentiable cone-beam CT reconstruction operation [92] and a volumetric loss term during the iterative optimization step. [...] These extensions would broaden the applicability of our method to diverse surgical and interventional scenarios.*

7. The practicality of the approach is further hampered by the fact that current pre-training is also anatomy specific, which qualifies the 5-minute claim since sufficient data must be available for pre-training the models. Recent foundation model approaches, e.g. [R3] and associated datasets, also leveraging data synthesis, may help overcome this challenge and should be considered.

Thank you very much for this interesting suggestion. We agree that requiring sufficient anatomy-specific CT data to be easily accessible for pretraining may be prohibitive for some users working on more niche applications.

Inspired by the reviewer’s comment, our revised manuscript now presents a single, pan-anatomic pose regression model trained on synthetic X-rays generated from a multi-dataset collection of over 2,000 rigidly co-registered volumetric CT scans covering the entire body. Dataset aggregation details are provided in the updated Results section (Page 3). The synthetic X-ray images were sampled in a 360° range in the LAO/RAO axis and a $\pm 60^\circ$ range in the CRA/CAU axis, along with large ranges for the in-plane rotation and three translations, in order to train on a broader set of C-arm poses than just the commonly acquired clinician views. Successfully training a model that fit this large pose distribution also required innovation to our original training setup (Fig. 2). Specifically, we had to change the rotation regression target from Euler angles to the quaternion adjugate [67], a 10-dimensional rotation parameterization that resolves discontinuities in mappings to **SE**(3).

Throughout the revised manuscript, we have shown that with 5 minutes of finetuning, this foundation model adapts to subjects in all the public datasets we evaluate, including the newly released femur dataset that we have added in this revision. This directly removes the need for users to identify and pretrain on large corpora of application-specific data, as in the original submission. While the original version of our software made it easy for a user to pretrain a patient-agnostic model with a public dataset and a single command, we agree that this foundation model for 2D/3D registration with whole-body coverage dramatically enhances the usability of **xvr**, and we thank the reviewer for their suggestion.

8. There are additional shortcomings that overall rank lower for me but should be noted, such as for example, the limitation that X-ray synthesis is being described as realistic but seems to be considerably incomplete, as the model considers single material, single energy absorption (there are several sentences in the manuscript, such as the one stating that appearance differences are primarily due to geometric distortions, which are certainly incorrect). The benefits of this renderer is differentiability, but more realistic open-source rendering frameworks such as DeepDRR exist, and should or at least could be considered for data synthesis of the target 2D x-rays to further enhance model robustness.

We believe that there is a miscommunication here. To clarify, we only claim that our model of X-ray synthesis is sufficiently accurate for *2D/3D registration*. We do not claim to apply to all clinical tasks one might wish to learn to perform with synthetic X-rays, such as segmentation or classification, where a more physically

accurate renderer like DeepDRR might perform better. The corresponding text in the manuscript has been edited to clarify this point:

Our model of X-ray synthesis does not consider various higher-order effects such as scattering, the polychromatic nature of X-ray beams, or the energy-dependent response of different tissue types. While our model is sufficiently accurate for 2D/3D registration, other machine learning tasks (e.g., segmentation or classification) trained using synthetic X-ray images may benefit from a more physically accurate, albeit non-differentiable, renderer such as DeepDRR [96].

Additionally, by “geometric distortions”, we meant that the left femur appears in different positions in the real and synthetic X-rays because the patient’s pose changes between the pre- and intra-operative acquisitions (Fig. 2A). We have updated the Results to emphasize this distinction:

*Synthetic X-rays rendered by **xvr** are **geometrically** consistent with real X-ray images (Fig. 2A-D). Differences in appearance are **primarily driven by patient motion** between the pre- and intra-operative imaging (e.g., the subject’s left femur moves between the two acquisitions in the second columns of Fig. 2A).*

9. In addition, while I commend the user-based study to establish a semblance of registration accuracy in the in-house dataset, it is known that such techniques fail to capture deviations in registration quality when the registration errors are low [R2]. The conclusion may still hold that the presented framework performs well on this data, however, the comparisons etc. should likely be treated with care.

Thank you for this interesting reference. After reading the article, we agree that our user study was insufficient to assess registration accuracy. Therefore, our expert clinical collaborators manually annotated corresponding keypoints on each X-ray and CT image pair, and we computed ground truth C-arm poses with the PnP algorithm. We have added this new analysis to the Results:

***xvr** achieved a median mTRE of 1.2 mm (IQR: 0.5–1.9), further demonstrating generalization to novel anatomical structures in real-world clinical data. Manual annotation of these keypoints required over 40 hours of clinicians’ time, whereas pose estimation and refinement with **xvr** ran completely automatically in just 10 minutes on a consumer-grade GPU (NVIDIA GeForce RTX 3060). Thus, **xvr** attains superior performance without prohibitively expensive keypoint annotations, highlighting a key advantage of our pose regression network over keypoint-based 2D/3D registration methods. [...] For comparison, we also evaluated iterative optimization initialized from the C-arm pose encoded in the DICOM header (Methods M.8). Because these parameters do not account for the positioning of the patient relative to the C-arm, this approach yielded highly inaccurate registrations in most cases (180.3 mm (IQR: 106.2–245.0); Fig. 5C-E).*

10. In conclusion, the method and this manuscript are well developed, and perhaps indeed useful in some way; yet, from my perspective they do not offer a fundamentally different paradigm, clear departure from prior work, or universally useful solution for 3D/2D registration, which limit my enthusiasm for this specific venue. I can see opportunities in a method/code framework focused venue where the utility offered here (namely a potentially quite usable framework) outweighs the more modest scientific contribution.

We sincerely thank the reviewer for their questions, constructive suggestions, and time spent reviewing our work. The expanded methodological contributions and experiments conducted based on the reviewer’s comments have helped clarify the thesis of our manuscript: **xvr** is the first fully automatic and generalizable pan-anatomic solution to 2D/3D rigid registration.

This manuscript presents many key innovations, which broadly increase the applicability and scalability of the proposed method over prior work. Thanks to the reviewer’s suggestion, we have better highlighted these contributions, including developing the whole-body foundation model for pan-application 2D/3D registration,

the reduction in training time to 5 minutes, and the ability to adapt to changing intrinsic parameters, amongst others. Most importantly, **xvr** achieves more accurate and consistent results compared to previous methods without requiring prohibitively expensive manual annotations, highlighting one of the critical advantages of our image-based pose regression neural network.

The fundamental methodological advancements and streamlined interface for general use on real data in **xvr** are valuable contributions. Indeed, registration with our package has already proved useful for downstream tasks, such as intraoperative 3D reconstruction for improved interventional guidance.^{8,9,10,11,12} Ultimately, the technical innovations in **xvr** make it a broadly applicable tool that solves the single-view rigid 2D/3D registration problem, which, despite decades of investigation, has lacked a reliable, accurate, and easy-to-use solution to this day.

[R1] Grimm, M., et al. (2021). Pose-dependent weights and domain randomization for fully automatic X-ray to CT registration. *IEEE transactions on medical imaging*, 40(9), 2221-2x232.

[R2] Cho, Sue Min, et al. "Visualization in 2D/3D registration matters for assuring technology-assisted image-guided surgery." *International journal of computer assisted radiology and surgery* 18.6 (2023): 1017-1024.

[R3] Killeen, Benjamin D., et al. "Fluorosam: A language-aligned foundation model for x-ray image segmentation." *arXiv preprint arXiv:2403.08059* (2024).

⁸ Frisken, S., Gopalakrishnan, V., Chlorogiannis, D.D., Haouchine, N., Cafaro, A., Golby, A.J., Wells III, W.M. and Du, R., 2025. Spatiotemporally constrained 3D reconstruction from biplanar digital subtraction angiography. *International Journal of Computer Assisted Radiology and Surgery*, pp.1-13.

⁹ Chouksey, P., Kashyap, M. and Subbarao, A., 2025. Deep learning-based optimization for accurate multimodal medical image registration. *Scientific Reports*.

¹⁰ Yu, W., Guo, X., Liu, X., Liu, Y., Zheng, H., Huang, Y. and Yuan, Y., 2025. GaussianReg: Rapid 2D/3D Registration for Emergency Surgery via Explicit 3D Modeling with Gaussian Primitives. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 21482-21491).

¹¹ Zhang, W., Zhao, L., Gou, H., Gong, Y., Zhou, Y. and Feng, Q., 2024. PRSCS-Net: Progressive 3D/2D rigid Registration network with the guidance of Single-view Cycle Synthesis. *Medical Image Analysis*, 97, p.103283.

¹² Lecomte-Denis, F., Verde, J., Dillenseger, J.L. and Cotin, S., 2025. Domain agnostic 2D-3D deformable registration Application to fluoroscopic guidance without contrast agent. *Medical Image Analysis*, 105, p.103688.

Response to Reviewer 2

This interesting manuscript proposes the use of patient-specific neural networks for intraoperative x-ray to CT volume registration. More specifically, it proposes the use of these networks for patient-specific PRE-registration, followed by refinement using a more conventional 2D/3D registration method to achieve a final registration result. The patient-specific registration network is trained via transfer learning from a previously trained patient-agnostic network derived from the CTPelvic1K dataset, and training times on the order of 5 minutes are reported. Pre-registration errors on pelvis images of 67.4 +/- 20 mm are reported and 3D target registration errors of 1mm and are reported after the refinement step for manually designated "fiducials" at designated anatomic landmarks from what seems to be single x-ray projections. Sub-mm 3D registration errors are achieved in 41-44% of cases in their test data set. Impressive reprojection errors ($mPE = .18mm$, $mRPE = .6mm$) are also reported. The authors also report the application of their method on DSA volumes for brain angiography, with roughly comparable results.

Overall, I found this to be an interesting and original approach, but I do have some questions that ought to be addressed before final publication.

1. Single-image 2D-3D registration results can be extremely sensitive to the accuracy of the C-arm intrinsic parameters. There should be some discussion of this issue and (ideally) some sort of simulation experiment to assess this sensitivity.

We agree, and thank the reviewer for this interesting suggestion. To quantify the sensitivity of our method to inaccuracies in the C-arm intrinsic parameters, we performed a simulation study in which we systematically corrupted either the pixel spacing or the source-to-detector distance (SDD) provided to the pose regression neural network. Specifically, we evaluated our finetuned models on the DeepFluoro dataset while perturbing each parameter independently by up to $\pm 25\%$ of its nominal value. The nominal pixel spacing was 0.194 mm, and we sampled corrupted values in the range [0.146, 0.242] mm. The nominal SDD was 1020 mm, and we sampled corrupted values in the range [765, 1275] mm.

This simulation shows that misestimation of the C-arm intrinsic parameters indeed induces large errors in the pose regression neural network (Fig. S4). The model is particularly sensitive to misestimation of the pixel spacing, as even a hundredth of a millimeter's worth of error increased the mTRE by tens of millimeters. However, the model was much less sensitive to misestimation of the SDD. This observation is important as it is typically very easy to estimate the pixel spacing (e.g., the size and number of pixels of the X-ray detector plane are known from the manufacturer's specification).

We have included this simulation study in the Methods (M.5):

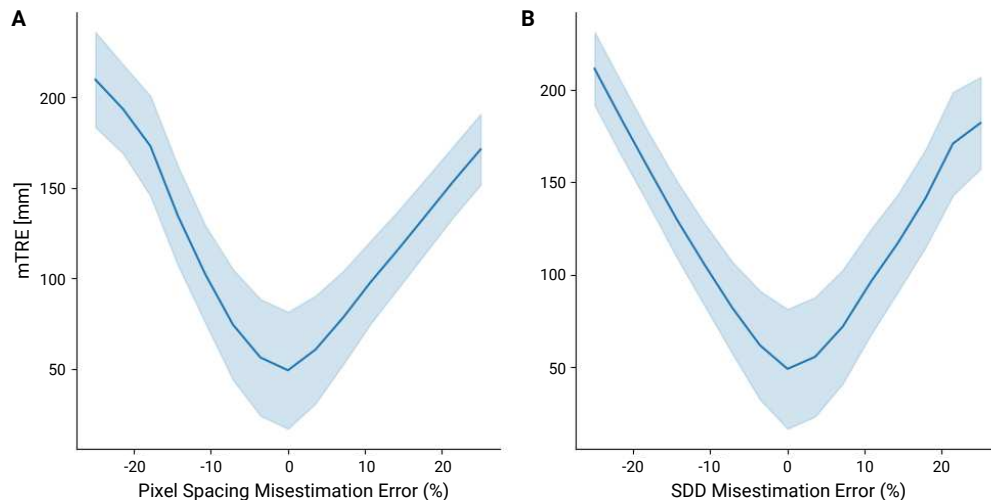


Figure S4. Effect of misestimated intrinsic parameters on initial pose estimation error. Mean target registration error (mTRE) as a function of systematic error in C-arm intrinsic parameters for patient-specific models evaluated on the DeepFluoro dataset. (A) Pixel spacing misestimation across 0.194 ± 0.0485 mm ($\pm 25\%$ of nominal value). (B) Source-to-detector distance (SDD) misestimation across 1020 ± 255 mm ($\pm 25\%$ of nominal value). Shaded regions represent 95% confidence intervals.

2. Although the use of manually-identified “fiducials” only appear EXPLICITLY in the evaluation of the results, these points are chosen to correspond to readily-identified anatomic features, which are places where the 2D and 3D images have significant geometric information and local image contrast. Thus, there may be some IMPLICIT bias in an evaluation based on these points. It would be useful to see some discussion of this point. Perhaps added discussion of the dGEO results would help.

Thank you for this nuanced point. We have clarified the text in the Methods to emphasize this potential source of implicit bias:

Manually selected fiducials chosen to evaluate mPD, mRPE, and mTRE metrics typically correspond to important anatomical landmarks. The geometry and local image contrast that makes these fiducials easy to identify for humans in 2D and 3D imaging may also induce implicit bias towards these structures for registration error estimation at the expense of other anatomical regions. This makes dGeo an important complementary error metric, as it does not rely on manually selected fiducials and instead assesses misregistration across the full anatomical region.

3. It would also be very useful to see the standard deviations for each of the results reported in table S1.

Thank you for this helpful suggestion. To be consistent with the main text, which reported the interquartile range (IQR), we have included the IQR for each method in Table S1.

	DeepFluoro				Ljubljana			
	mPE (2D) [mm]	mRPE (2.5D) [mm]	mTRE (3D) [mm]	dGeo (3D) [mm]	mPE (2D) [mm]	mRPE (2.5D) [mm]	mTRE (3D) [mm]	dGeo (3D) [mm]
Fixed init. (xReg) [45]	86.3(80.2) 18.2%	151.3(133.0) 8.8%	335.4(249.1) 0.0%	365.9(329.5) 0.0%			N/A	
Landmark reg. (xReg) [53]	1.0(3.2) 53.6%	2.2(29.6) 16.9%	3.1(39.9) 1.4%	4.3(52.3) 3.3%			N/A	
Landmark init. (xReg) [53]	0.9(0.4) 59.9%	1.9(2.2) 20.2%	2.8(2.7) 2.7%	3.8(4.1) 6.0%			N/A	
Fixed init. (xvr) [38]	74.5(60.4) 10.8%	137.6(83.9) 9.7%	277.4(137.1) 7.6%	352.8(203.3) 8.0%	10.9(16.5) 30.0%	129.6(146.2) 25.0%	149.9(175.2) 10.0%	149.8(160.2) 10.0%
Foundation (xvr)	0.2(3.0) 71.0%	0.9(16.5) 52.5%	1.3(19.0) 31.2%	1.6(23.4) 41.7%	394.7(660.5) 0.0%	826.7(576.2) 0.0%	1315.5(392.1) 0.0%	1384.1(719.1) 0.0%
De novo (xvr)	0.2(0.2) 93.4%	0.6(0.9) 70.2%	0.9(1.1) 41.4%	1.2(1.3) 55.5%	0.7(0.1) 100.0%	1.0(1.0) 55.0%	1.4(1.5) 35.0%	1.2(1.1) 20.0%
Finetuned (xvr)	0.2(0.2) 93.9%	0.6(0.8) 69.1%	0.9(1.1) 40.9%	1.1(1.3) 53.3%	0.7(0.1) 95.0%	1.0(0.9) 45.0%	1.4(1.5) 35.0%	1.2(1.3) 20.0%

4. One thing that I noticed is that the post-refinement results in Table S1 for the authors’ three xvr methods (patient-agnostic, patient-specific, finetuned) are essentially the same, one suspects within the limits of statistical significance. Thus, what are the actual advantages of using any form of patient-specific training vs. patient-agnostic training only, followed by refinement. It may be that a more powerful statistical analysis (paired t-test, perhaps) could demonstrate this. In any case, this point should be addressed.

The reviewer is absolutely correct that the post-refinement results of the three methods (patient-agnostic, patient-specific, and finetuned) are very similar **on the DeepFluoro dataset (pelvis)**. However, this is not

the case for the Ljubljana dataset (neurovasculature). On this dataset, the patient-agnostic model achieved a median mTRE of 45.1 mm (IQR: 1.5–212.3), whereas the patient-specific and finetuned models achieved mTREs of 1.4 mm (IQR: 0.8–2.2) and 1.5 mm (IQR: 1.0–2.2), respectively. These differences arise from the fact that predictions made by the patient-agnostic model need to be mapped to the reference frame of the patient’s CT with a rigid transform. As we mentioned in the Results, this can be done more accurately for the pelvis (DeepFluoro) than the neurovasculature (Ljubljana):

Finally, the relative simplicity of pelvic anatomy makes it easy to correct the pose estimates from a foundation model with 3D rigid registration. In contrast, when working with the Ljubljana dataset, we were forced to rely on center-alignment, which is less accurate.

We have also clarified the advantages of from-scratch patient-specific training, as opposed to patient-agnostic modeling (with or without finetuning):

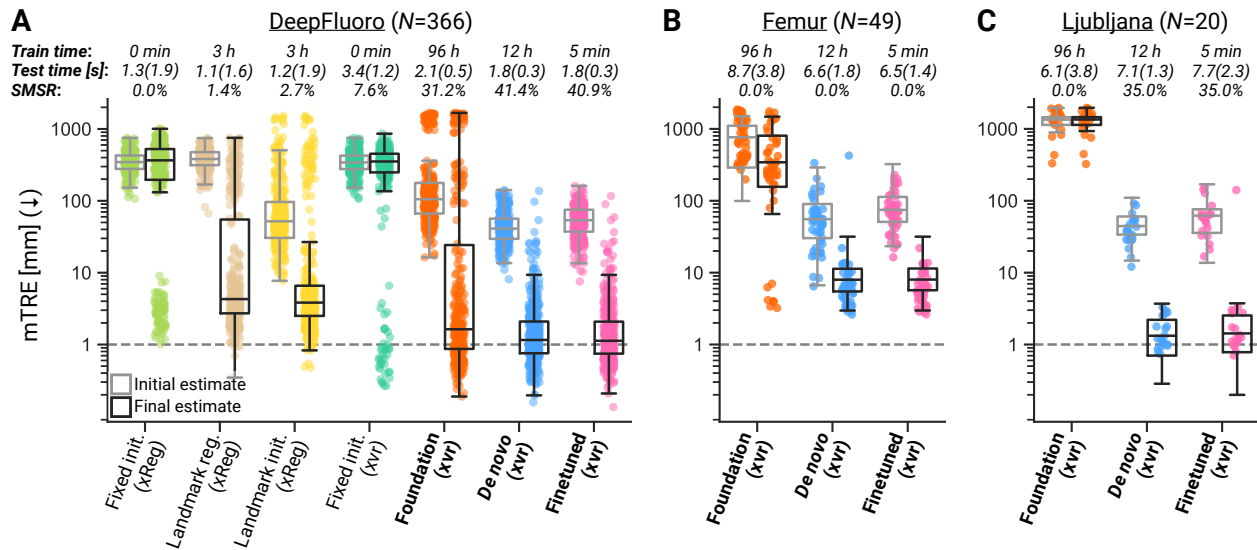
*We emphasize that the ability to train a de novo subject-specific network with **xvr** remains important for specialized scenarios, such as patients with highly distinct anatomies (e.g., situs inversus) or applications in veterinary surgery.*

5. The statement “... achieving submillimeter-accurate registration at intraoperative speeds and improving upon existing methods by an order of magnitude” seems like an over-claim and perhaps should be modified, partly because of the reasons mentioned in my point 4. All the sub-mm results also rely on a further refinement step (M.6) and only achieve sub-mm results in less than 50% of cases. These limitations need to be discussed, and (again) the relative advantages of the patient-specific training need to be demonstrated more convincingly.

We agree with the reviewer and have accordingly removed the sub-millimeter claim across the manuscript.

6. It would be useful to see the times for the pose-refinement step, as well as the patient-specific training step in the reported results. How do these compare with the different methods?

We have updated Fig. 4A-C to report the mean and standard deviation pose-refinement time for each evaluated method (see the updated figure below). Compared with methods that use the **xReg** backend, our pose-refinement step is approximately 0.5 s slower on average, but exhibits a smaller standard deviation and yields substantially more accurate pose estimates.



7. Relative to times, “5 minutes” is only meaningful if one can say more about the computational hardware involved. How large are the networks involved? How many GPUs? Which ones? Etc.

This is an important point and we thank the reviewer for raising it. We have now included these details throughout the Results when the various models are introduced. For completeness, we have included a thorough summary in the Methods (M.9):

All our neural networks use a ResNet34 backbone (21,278,400 parameters) and two linear layers for rotation and translation regression (6,669 parameters). Our foundation model was trained on a single NVIDIA H200 GPU (141 GB of VRAM). All other models were trained on a single NVIDIA RTX 6000 Ada GPU (48 GB VRAM) in a 48-core (96-thread) Linux system with dual-socket Intel(R) Xeon(R) Gold 5418Y CPUs (24 cores per socket, up to 3.80 GHz). With the exception of the Brigham dataset, pose refinement for all tested methods was performed on the same computational platform as used for network training (NVIDIA RTX 6000 Ada GPU). Due to privacy constraints, the Brigham dataset was registered on a secure workstation equipped with a single NVIDIA GeForce RTX 3060 GPU (12 GB VRAM).

8. Although the “fiducials” are manually-selected, it wasn’t entirely clear whether this was done on the 3D volumes directly or on 2D projections. I think these may be the same landmarks that the authors used for their “landmark initialization” comparison, but it would be useful to see a comparison.

We agree that this was unclear originally and have accordingly edited the revision. To clarify, the authors of the DeepFluoro and Ljubljana datasets annotated landmarks in 3D and then projected them into 2D. The authors of the newly added Femur dataset do not provide landmarks, so we extract them from the femur surface on CT. These 2D landmarks are indeed the same ones used for the landmark initialization baseline for the DeepFluoro dataset. However, we use the manually annotated 3D landmarks to compute registration accuracy via all metrics. We clarified this point in the updated caption for Fig. 3:

In the DeepFluoro and Ljubljana datasets, fiducial markers were manually annotated in 3D, while fiducials in the Femur dataset were extracted from the femur surface on CT. In all datasets, 3D fiducials were projected into 2D using ground truth C-arm poses.

9. One general concern, common to most medical image analysis papers, is that reporting “success” rates based on thresholds or even the distributions based on experimental results on sample data somewhat begs the question of what the accuracy is for a particular, previously unseen, patient. It is probably unfair to expect the authors to address this issue as a condition for publication of this particular paper, but any insights that they can contribute would be welcome, even as a possibility for future research.

We thank the reviewer for this question. We would argue that every subject we present evaluation metrics on is previously unseen. Specifically, the corpus of whole-body CTs we use to train our foundation model for 2D/3D registration does not contain any of the subjects from the DeepFluoro, Ljubljana, Brigham, or newly included Femur datasets. Furthermore, we use the same finetuning and pose refinement setup for each subject in the DeepFluoro, Ljubljana, and Femur datasets (e.g., the same neural network architecture, learning rate, training schedule, image similarity functions, etc.), demonstrating the robustness of our package across multiple datasets. Finally, our evaluation of the Brigham dataset shows the accuracy of **xvr** on a previously unseen patient *without any patient-specific adaptation*. Directly applying a population-level network to unseen patients yielded accurate pose registration, achieving a median mTRE of 1.2 mm (IQR: 0.5–1.9). Taken together, these results show that our framework is general and maintains high accuracy on previously unseen patients and anatomy.

To summarize, I think that this is an interesting paper that reports a good idea. The authors have done a good job describing their method in sufficient detail to allow others to replicate and adapt their results. I look forward to seeing the revised version.

We thank you very much for your expert evaluation. Your comments have substantially improved and clarified our manuscript.

Response to the Editor

Dear Dr. Caputa,

Thank you for coordinating our reviews. In the revised manuscript and response letter, we have fully addressed all remaining issues raised by the reviewers.

In this revision, we ask that you provide an example of an out of distribution patient with fine tuning optimizes model performance, vs not fine tuning, as asked for by reviewer #3 (point 4).

We have now included a new experiment in the manuscript that finds that our method successfully generalizes to entirely out-of-distribution patients with novel pathologies. Specifically, we tested our foundation model's ability to align 2D DSAs to 3D MRAs of neonates with an exceptionally rare congenital arteriovenous malformation. As our model was trained exclusively on adults without neurovascular lesions, this represents an extreme test of out-of-distribution robustness.

This experiment uses preoperative 3D MRAs and intraoperative 2D DSA imaging from four neonatal patients at Boston Children's Hospital. After patient-specific finetuning (5 minutes per subject), our method produced successful registrations even for these extremely out-of-distribution cases (see the new Figure 5).

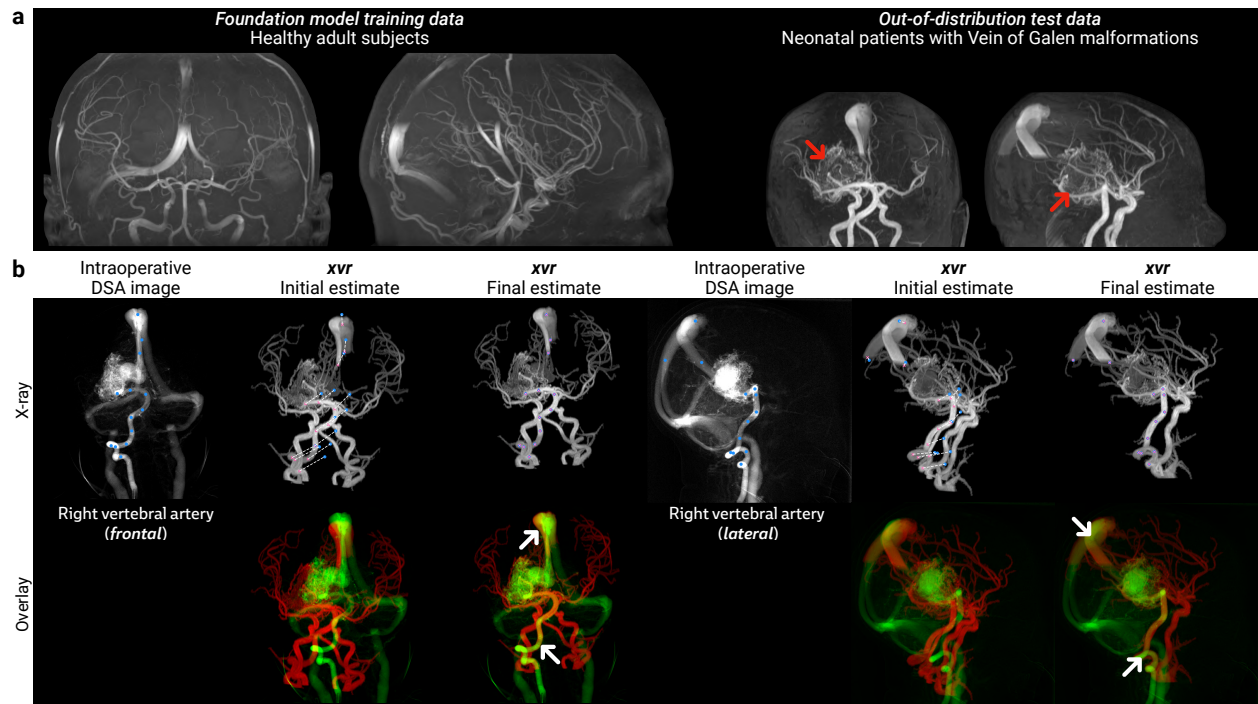


Figure 5: Extreme out-of-distribution generalization. **a**, Maximum intensity projections of 3D MRAs from a healthy adult subject in our pretraining corpus (*left*) versus a neonatal patient with a large Vein of Galen malformation (*right*, indicated by red arrows). **b**, Our foundation model finetuned on neonatal MRAs for 5 min per patient ($N = 4$) successfully registered intraoperative DSAs, achieving median mTRE 2.3 mm (IQR: 1.5–3.0), despite being pretrained exclusively on adult anatomy. Note that MRA captures only large arterial flow, while selective DSA injection reveals microvasculature and venous drainage absent from the preoperative scan, challenging registration, as corresponding structures may be invisible between modalities. White arrows indicate overlapping anatomical landmarks post registration.

You will also need to make some editorial changes to your paper so that it is as brief as possible and complies with our Guide to Authors.

We have accordingly also made the following changes to make our manuscript more direct and comply with the journal's guidelines:

- ✓ Main text shortened from roughly 4,800 words to 3,800 words
- ✓ Abstract reduced to ≤ 200 words
- ✓ Introduction reduced to ≤ 500 words
- ✓ Title revised to ≤ 75 characters
- ✓ All section headers revised to ≤ 40 characters
- ✓ All figure captions revised to ≤ 300 words
- ✓ Main text reference counts lowered to ≤ 50 references
- ✓ Added statistical details to all figure captions

All revisions made to the manuscript have been annotated and tracked with *blue font*.

We thank the editor and the reviewers again for their feedback, as it has made the manuscript considerably stronger. We believe the revised manuscript has now resolved all reviewer comments and look forward to your decision.

Response to Reviewer 2

The authors have answered my questions and have significantly improved their interesting manuscript. I think that it is fine to publish.

Thank you very much for your time and expert feedback.

Response to Reviewer 3

1. Overall, the revisions are responsive and the manuscript is considerably stronger than the original submission. I commend the authors for conducting new analyses interrogating the generalizability of **xvr**, releasing model weights on HuggingFace, running additional sensitivity analyses, and increasing the clinical rigor of the evaluations. There are nonetheless still some concerns I have about generalizability that I list below.

2. The baseline comparisons to other methods are improved but still quite limited for a paper in a top journal. In particular, the authors write that “Comparisons of the reported results on the same anatomical region show that **xvr** achieves the most accurate and reliable registrations across all previous methods on a common anatomical structure,” yet this claim rests on comparisons of **xvr** to only two comparators, Gao *et al.* and Grimm *et al.*, on just one dataset based on only 6 cadaveric CTs (DeepFluoro), and that too on non-identical data (as the authors acknowledge). No formal test of statistical significance is reported, and given that the 362 X-rays in DeepFluoro are not independent (many are from the same subject), the effective sample size is much smaller than it appears. The superiority claims in the results and discussion should be softened to reflect these limitations.

We thank the reviewer for their time and thoughtful feedback. We now report formal statistical tests for all tested methods and have softened our claims in the manuscript. The specific critiques in Point 2 are addressed below.

“The authors’ claims rest on comparisons against only two methods on just one dataset (DeepFluoro).”

We acknowledge the limited number of comparator methods, which reflects the scarcity of open-source 2D/3D registration methods with both public code and/or pretrained models available for rigorous evaluation on new data. To clarify, in addition to comparisons with Gao *et al.* [19] and Grimm *et al.* [6], our paper does also evaluate four other baselines with different initialization and refinement strategies: Gopalakrishnan and Golland [18] with fixed initialization and Grupp *et al.* [5, 20] with fixed initialization, landmark-based initialization, and landmark-based regularization.

Regarding datasets, our quantitative evaluations also include two other independent datasets (Femur and Ljubljana) in addition to DeepFluoro, each representing distinct anatomical regions, imaging protocols, and clinical contexts from DeepFluoro. We also include two other internal clinical datasets (Brigham and Boston Children’s) that demonstrate **xvr**’s robustness in real-world neurovascular settings. Altogether, our evaluation suite spans multiple datasets, initialization schemes, and model variants, strengthening the evidence for our method’s generalizability.

“Effective sample size is small.” We agree with the reviewer that the effective sample size in our quantitative experiments is relatively small due to the correlation between X-rays from individual subjects and that this should be acknowledged as a limitation. We have added a paragraph to the Discussion to incorporate this:

*While our study is the largest and most anatomically diverse evaluation of 2D/3D registration algorithms on real X-ray images and MR/CT volumes to date, our overall sample size remains relatively small. We evaluate **xvr** on 21 subjects from public datasets and 54 from private clinical sources, which reflects the paucity of openly available paired X-ray/CT data and the high cost of acquiring such images for evaluation. Although **xvr** is statistically significantly better than prior work across heterogeneous datasets, these claims of out-performance should still be interpreted with caution. Stronger evidence of robustness will require larger cohorts spanning additional anatomical regions, disease presentations, and imaging protocols, alongside pilot studies prospectively integrating **xvr** into intraoperative settings.*

“No statistical testing + data not independent.” We agree that our statistical presentation would improve with greater rigor and now present statistical tests with explicit modeling of inter-sample dependence. For each method, we fit a linear mixed-effects model on log-transformed mTRE with subject ID as a random intercept, directly accounting for the correlation between X-rays taken from a single subject. We report two complementary analyses: (1) all pairwise comparisons to establish method rankings and (2) two-sided comparisons of each baseline against our finetuned patient-specific model using Wald tests with Bonferroni correction for multiple comparisons (Table S3).

On DeepFluoro, our finetuned model achieves a median mTRE of 1.1 mm (IQR 1.4 mm). Statistically, all baseline methods are significantly worse (all Bonferroni-corrected $p < 0.0001$). Further, there is no significant difference in accuracy between our finetuned and *de novo* models (which achieves median 1.2 mm, IQR 1.3 mm; Bonferroni-corrected $p = 1.0$), trained for 5 mins and 12 hrs each, respectively. These results support our claim that patient-specific finetuning of our foundation model reaches the same accuracy as a *de novo* model in $144\times$ less training time. The full pairwise analysis yields a clear ranking: patient-specific methods (rank #1) > foundation model (rank #3) > landmark initialization [5] (rank #4) > landmark regularization [5] (rank #5) > fixed initialization [18, 20] (rank #6), with all differences remaining significant after correcting for multiple comparisons. On Femur and Ljubljana, our patient-specific models similarly achieve the lowest mTRE with statistical significance across all comparisons.

Following the reviewer’s suggestion, we have softened the broad superiority claims and instead revised the Results and Discussion to state that our method achieves significantly lower mTRE than all tested comparators on the evaluated datasets, as determined by significance tests incorporating inter-sample correlation with additional Bonferroni correction.

	Rank	Method	Renderer	mTRE [mm]	z	p_{adj}	Significance
DeepFluoro	#1	Finetuned	xvr	1.1(1.3)	—	—	—
	#1	<i>De novo</i>	xvr	1.2(1.3)	0.08	1.0000	n.s.
	#3	Foundation	xvr	1.3(3.0)	7.34	<0.0001	***
	#4	Landmark init. [5]	xReg	3.8(4.1)	16.29	<0.0001	***
	#5	Landmark reg. [5]	xReg	4.3(52.3)	18.56	<0.0001	***
	#6	Fixed init. [20]	xReg	365.9(329.5)	42.87	<0.0001	***
	#6	Fixed init. [18]	xvr	352.8(203.3)	47.30	<0.0001	***
Femur	#1	Finetuned	xvr	8.0(5.7)	—	—	—
	#1	<i>De novo</i>	xvr	8.0(5.8)	0.82	1.0000	n.s.
	#3	Foundation	xvr	343.1(647.4)	15.09	<0.0001	***
Ljubljana	#1	Finetuned	xvr	1.4(1.8)	—	—	—
	#1	<i>De novo</i>	xvr	1.3(1.5)	−1.09	0.8262	n.s.
	#3	Foundation	xvr	1351.1(320.1)	21.78	<0.0001	***

Table S3: Ranking of each method’s final pose estimation error per dataset, reported as median(IQR) in mm. z and p_{adj} are from pairwise linear mixed-effects models comparing each method to Finetuned (reference) on log(mTRE) with subject as random intercept and X-ray as a variance component, accounting for the paired design where every X-ray is evaluated by every method. p -values are Bonferroni-corrected within each dataset. Ranks are derived from all pairwise comparisons; tied ranks indicate no significant difference ($p_{\text{adj}} > 0.05$).

3. The core contribution of patient-specific finetuning remains inadequately demonstrated. In particular, the results for the three **xvr** variants (finetuned, de novo, foundation) presented in Table S1 reveal that the finetuned and de novo models achieve nearly identical performance across DeepFluoro and Ljubljana. Both attain a mTRE of 0.9(1.1) mm and 41.4%/40.9% submillimeter success rate % on DeepFluoro, and the two methods achieve identical mTRE and submillimeter success rate % on Ljubljana. Is the advantage of patient-specific finetuning over de novo training exclusively one of speed (5 min vs. 12 h) and not accuracy? If so, this should be made more explicit. Additionally, the comparisons between these three variants in these two datasets rest on only 6 and 10 unique subjects, making it difficult to draw firm conclusions about when and by how much finetuning helps.

We thank the reviewer for this careful reading of Table S1. The reviewer’s interpretation is correct: in our context, the major advantage of patient-specific finetuning of our foundation model over *de novo* training (i.e., training on that subject from random initialization) is in its training time and not necessarily its accuracy. **xvr** focuses on training speed, as its target applications (i.e., surgery, image-guided interventions, radiation oncology, and surgical robotics) all demand tools that can be deployed off-the-shelf with minimal setup time. When given a clinically realistic time budget of 5 mins, our finetuned foundation model achieves statistically equivalent registration accuracy to a *de novo* model trained for 144× longer (12 hrs), making the proposed model much more practical for clinical and surgical use cases.

We have revised the manuscript to state this explicitly:

Our patient-specific finetuned (trained for 5 mins) and de novo models (trained for 12 hrs) were not significantly different in accuracy from each other on any dataset (Bonferroni-corrected $p_{\text{adj}} \geq 0.05$), confirming that rapid patient-specific adaptation of our foundation model achieves comparable accuracy to training from scratch on any specific patient, while requiring 144× less training time (Table S3; Methods M.10).

Moreover, we acknowledge the limited sample sizes (6, 10, and 5 subjects across DeepFluoro, Ljubljana, and Femur, respectively) and have included it as a limitation in the Discussion as described above. However, we emphasize that these statistically significant gains hold consistently across all three datasets even when small effective sample sizes limit statistical power. On Femur, finetuned and *de novo* models again achieve statistically indistinguishable mTRE ($p_{\text{adj}} = 1.0$, $z = 0.59$) while maintaining the same 144× reduction in training time. We have added these results to Table S1 for completeness.

4. Related to point 3, while the authors argue that “...rapid patient-specific finetuning eliminates the out-of-distribution failures that plague previous methods when encountering patients not well-represented in their training sets, ensuring reliable performance for every individual undergoing an intervention,” the paper does not actually test this hypothesis rigorously. The authors do not test the finetuned model on anatomical structures or patient populations (e.g., pediatric, severely pathological) that are underrepresented in the pretraining corpus, which may be precisely where the foundation-then-finetune paradigm would need to succeed to support the generalizability claims being made.

We thank the reviewer for this excellent point. To rigorously test performance on under-represented and pathological subjects, we collected new data for this revision from four neonatal patients with rare congenital Vein of Galen malformations at Boston Children’s Hospital.

Our foundation model was trained exclusively on adult patients without neurovascular lesions. Testing on neonates with this rare neurovascular malformation represents a compounded distribution shift in patient population (pediatric vs. adult), anatomy (age-related head size and shape changes), and pathology (severe vascular abnormality vs. normal anatomy). After 5 minutes of patient-specific finetuning per subject, our method produced successful registrations with accurate anatomical alignment with an average mTRE of 2.3 mm (IQR: 1.5–3.0) on expert-annotated keypoints across 4 subjects, shown in the newly added figure below.

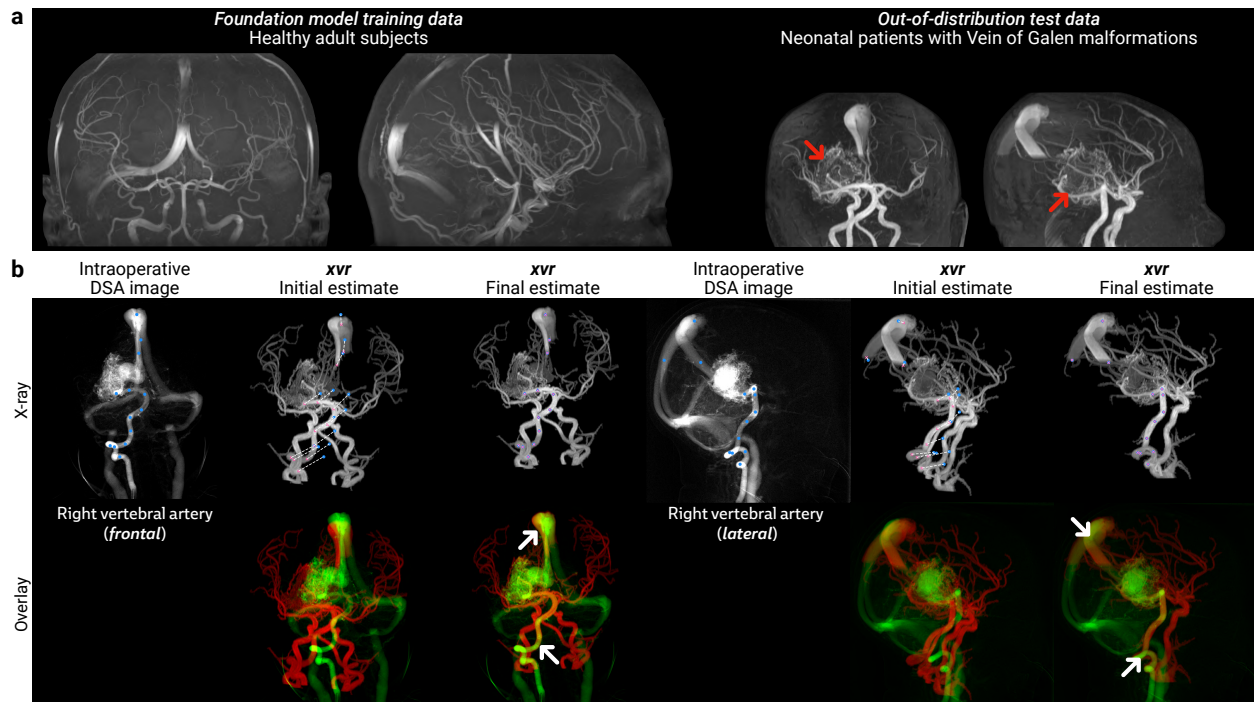


Figure 5: Extreme out-of-distribution generalization. **a**, Maximum intensity projections of 3D MRAs from a healthy adult subject in our pretraining corpus (*left*) versus a neonatal patient with a large Vein of Galen malformation (*right*, indicated by red arrows). **b**, Our foundation model finetuned on neonatal MRAs for 5 min per patient ($N = 4$) successfully registered intraoperative DSAs, achieving median mTRE 2.3 mm (IQR: 1.5–3.0), despite being pretrained exclusively on adult anatomy. Note that MRA captures only large arterial flow, while selective DSA injection reveals microvasculature and venous drainage absent from the preoperative scan, challenging registration, as corresponding structures may be invisible between modalities. White arrows indicate overlapping anatomical landmarks post registration.

We have added these results to the manuscript:

Extreme out-of-distribution generalization. To evaluate performance on populations and conditions completely unrepresented in our pretraining data, we patient-specific finetuned our foundation model on time-of-flight MRAs from four neonatal patients at Boston Children’s Hospital presenting with Vein of Galen malformations. This cohort represents an extreme distribution shift from our pretraining data: pediatric patients are entirely absent from our adult-only corpus, and this vascular pathology is exceptionally rare (less than 1 in 50,000 live births [41]). These patients also differ dramatically from healthy adults in size and exhibit pathological lesions (Fig. 5a). Furthermore, the 2D DSAs also visualizes microvasculature and venous drainage entirely invisible on the 3D MRAs. Despite this multi-faceted domain shift, 5 min of patient-specific finetuning on average intensity projections simulated from preoperative 3D MRAs enabled accurate registration to intraoperative 2D DSA. The models achieved mTREs of 2.3 mm (IQR: 1.5–3.0), demonstrating that our approach generalizes reliably to patient populations and clinical contexts entirely absent from pretraining data (Fig. 5b).

5. The framing in the introduction and discussion are improved, but the presentation should still be toned down given the issues above.

We appreciate the reviewer’s feedback and have accordingly revised the language throughout the introduction and discussion to reflect the limitations raised above. More specifically, we have removed claims of global

superiority and now state that our method achieves statistically significantly lower registration error than all tested comparators on the evaluated datasets. As described above, we have now also added a paragraph discussing the limitations of our sample size to the Discussion, repeated here for convenience:

*While our study is the largest and most anatomically diverse evaluation of 2D/3D registration algorithms on real X-ray images and MR/CT volumes to date, our overall sample size remains relatively small. We evaluate **xvr** on 21 subjects from public datasets and 54 from private clinical sources, which reflects the paucity of openly available paired X-ray/CT data and the high cost of acquiring such images for evaluation. Although **xvr** is statistically significantly better than prior work across heterogeneous datasets, these claims of out-performance should still be interpreted with caution. Stronger evidence of robustness will require larger cohorts spanning additional anatomical regions, disease presentations, and imaging protocols, alongside pilot studies prospectively integrating **xvr** into intraoperative settings.*

| *Well-documented README included with installation and use instructions, I was easily able to install the package (**xvr**) with the standard Python package manager (pip) from the command line.*

Thank you very much for trying our package! We are committed to its long-term maintenance.