Metadata-Aligned 3D MRI Representations for Contrast and Sequence Understanding

Anonymous Author(s)
Affiliation
Address
email

Abstract

Magnetic Resonance Imaging (MRI) offers diverse contrasts and acquisition protocols, yet the lack of standardized labels across sites and scanners makes automated sequence classification and contrast-aware applications challenging. We propose a metadata-guided CLIP framework for learning 3D MRI contrast representations by aligning images with their DICOM metadata. This alignment enables the model to capture both contrast-specific and acquisition-related variations, yielding embeddings that support diverse downstream tasks such as image–metadata retrieval and sequence classification, and can further serve as a foundation for contrast-invariant representation learning and cross-site harmonization. Evaluated on a large and heterogeneous clinical MRI dataset, our framework yields well-structured latent spaces, achieves strong image metadata retrieval, and forms meaningful unsupervised clusters of MRI sequences. Furthermore, the learned embeddings enable competitive few-shot sequence classification performance compared to fully supervised 3D networks. Code and weights are publicly available at [anonymised].

1 Introduction

Magnetic Resonance Imaging (MRI) is a versatile modality widely used in clinical practice, providing diverse contrasts and acquisition protocols that capture complementary anatomical and functional information. However, clinical MRI datasets are often highly heterogeneous, collected across multiple scanners, sites, and patient populations, and typically lack standardized sequence or contrast labels [1]. This variability poses major challenges for automated sequence classification[2], contrast-aware analysis[3], and downstream tasks such as image retrieval [2] or harmonization [4, 5]. Recent advances in self-supervised and contrastive representation learning, particularly CLIP-style frameworks, have shown that aligning different modalities can yield embeddings with strong generalization and transfer capabilities [6–9]. However, existing approaches often depend on full supervision or do not fully capture the rich acquisition metadata inherent to MRI due to limited data variability, leading to representations that remain sensitive to scanner and protocol specific variations [10–13].

We propose a metadata-guided CLIP framework for 3D MRI that aligns volumetric images with their DICOM [14] metadata to learn contrast-aware embeddings. These embeddings capture acquisition-specific variations, achieve strong image–metadata retrieval performance and form structured latent spaces that naturally cluster MRI sequences. Moreover, the learned representations enable competitive few-shot classification and provide a promising foundation for downstream tasks such as cross-site data harmonization and modality-aware image analysis.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

2 Methods

MR-CLIP learns MRI contrast representations by contrastively aligning volumetric image embeddings with structured DICOM metadata (see Fig. 1). For each acquisition, a 3D image encoder extracts volumetric features, while a metadata encoder projects DICOM tags (via a natural language template) into a shared embedding space. To account for small parameter differences that do not meaningfully affect image contrast, we group metadata by binning numeric fields (e.g., TR, TE) and clustering categorical fields (e.g., Manufacturer), forming semantically similar acquisition groups and reduce 21,660 unique metadata combinations to 1,415 contrast labels, which are then used to guide the contrastive learning process. The list and distribution of used metadata are provided in the Appendix.

MR-CLIP is trained using a Supervised Contrastive (SupCon) Loss [15]. Let z_i denote the anchor embedding for sample i, and let $P(i)$ be the set of positive embeddings for i, including exact matches and other samples from the same metadata group. The loss for anchor i is

$$\mathcal{L}_i = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(z_i^\top z_p/\tau)}{\sum_{a \in A(i)} \exp(z_i^\top z_a/\tau)},$$

where $A(i)$ is the set of all embeddings in the batch excluding i, and τ is a temperature hyperparameter. This loss is calculated separately for image and metadata embeddings and then averaged. Compared to standard InfoNCE [16], which considers only a single positive per anchor, SupCon naturally handles multiple positives, encouraging the model to cluster semantically similar acquisitions. We also train a 2D variant of MR-CLIP that aligns individual slices with their corresponding metadata; in this case, the SupCon objective benefits from including different slices from the same brain, promoting consistent contrast representations and invariance to anatomical variations.

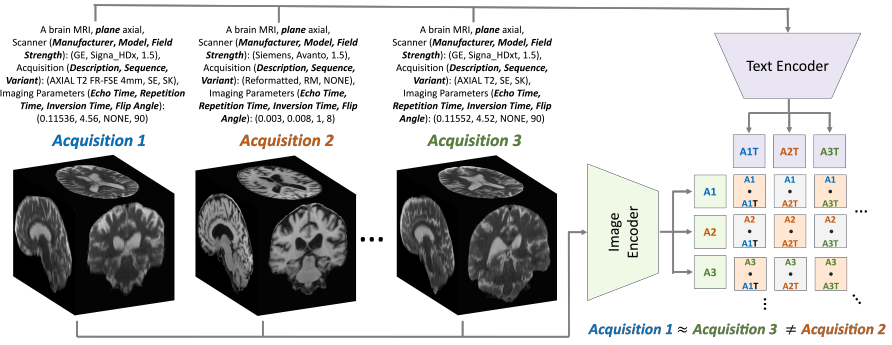

Figure 1: MR-CLIP aligns 3D MRI volumes with their corresponding DICOM metadata, resulting in contrast representations that are robust to anatomical variability and subtle parameter differences.

3 Results and Discussion

We evaluate MR-CLIP through three complementary experiments that assess cross-modal alignment, representation quality, and metadata interpretability. As summarized in Table 1, 2D and 3D MR-CLIP

Table 1: Cross-modal retrieval performance (%). Showing Recall@K (R@1/5/10) for image-to-text, 3D scan-to-text, and text-to-image retrieval. Linear classification accuracy (%) is shown in the rightmost column. Highest values in each column are bolded.

Model	**Image→Text**			**3D Scan→Text**			**Text→Image**			**Linear Acc.**
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	
BiomedCLIP	1.4	5.0	8.4	2.5	9.8	15.0	3.6	9.5	13.1	39.0
BiomedCLIP (Fine-tuned)	50.0	78.5	82.6	67.4	89.1	92.1	38.5	65.8	71.8	75.5
ViT-B/16 (InfoNCE Loss)	65.6	**85.2**	**90.4**	68.8	92.2	94.4	49.3	69.3	76.6	71.3
ViT-S/16	46.7	79.1	84.4	69.0	92.2	95.2	64.6	77.8	80.9	73.6
2D MR-CLIP (ViT-B/16)	**66.0**	77.3	78.3	**78.7**	**94.2**	**95.3**	**90.9**	**93.6**	**94.4**	82.6
3D MR-CLIP	-	-	-	60.2	79.0	82.0	79.3	91.6	94.0	**86.9**

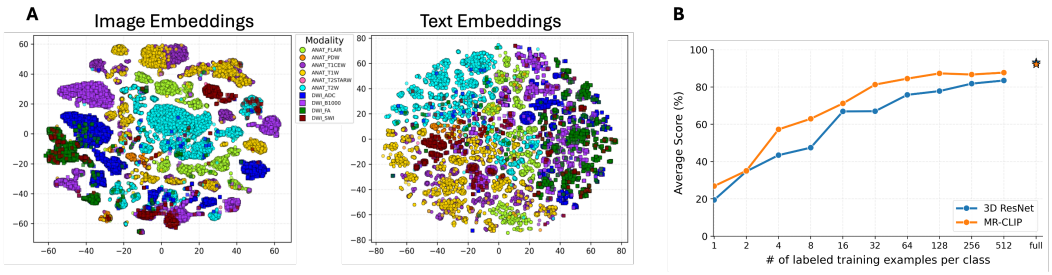

Figure 2: A: t-SNE visualizations of image and text embeddings, color coded by sequence. B: Few-shot learning performance of MR-CLIP, compared to supervied 3D ResNet baseline.

variants outperform baselines in image-to-metadata, metadata-to-image, and 3D scan-to-metadata retrieval (for 2D models, retrieval results are aggregated across slices to produce comparable 2.5D scan-level scores) across R@1, R@5, R@10, and in linear metadata classification. The 2D MR-CLIP achieves the highest overall retrieval scores, reflecting precise slice-level alignment and efficient feature utilization, while the 3D variant closely follows, demonstrating volumetric representations that generalize effectively across diverse imaging protocols.

To visualize the learned representation structure, we project image and metadata embeddings using t-SNE (Fig. 2A). MR-CLIP embeddings form distinct clusters across MRI sequence types, clearly separating anatomical and diffusion-weighted images. In few-shot sequence classification (Fig. 2B), MR-CLIP consistently outperforms a 3D ResNet baseline, particularly under low-shot settings (1–64 samples per class), while performing comparably when trained on the full dataset, highlighting its stronger inductive bias under limited supervision.

Finally, we analyze per-tag prediction accuracy under linear probing across 2D, 2.5D, and 3D MR-CLIP variants in Fig. 3. The 2.5D model performs best overall, suggesting that aggregating local slice context provides effective balance between efficiency and representational capacity. Discrete fields such as Acquisition Plane and Field Strength are classified with near-zero error, while numerical parameters (e.g., TE, TR) exhibit higher bin misclassifications but small average deviations, indicating predictions close to the true values.

Overall, MR-CLIP effectively disentangles image contrast from anatomical content, producing robust, contrast-aware embeddings that generalize across scanners and support downstream tasks such as retrieval, sequence recognition, and metadata analysis. Although grouping quality and metadata incompleteness and inconsistencies may introduce noise in training process, the framework establishes a scalable foundation for metadata-guided MRI representation learning, bridging image features with acquisition semantics for improved analysis and harmonization.

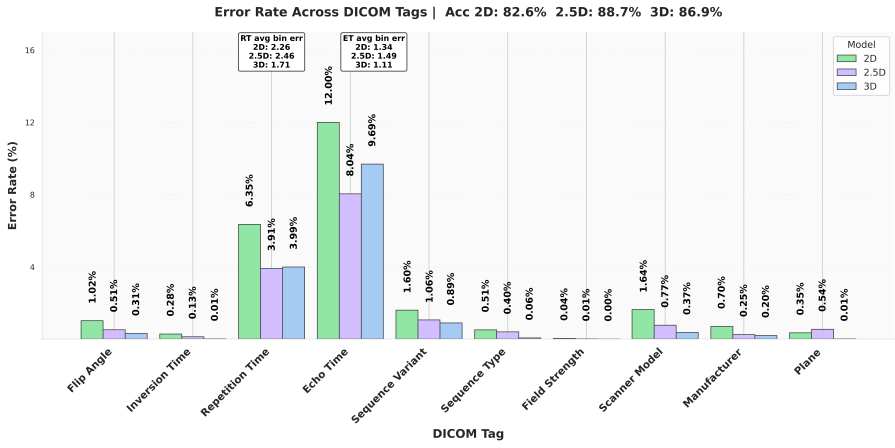

Figure 3: Error rates across DICOM tags based on linear probe classification results.

References

[1] Himanshu Sinha and Pradeep Reddy Raamana. Solving the pervasive problem of protocol non-compliance in MRI using an open-source tool mrQA. *Neuroinformatics*, 22:297–315, 2024.

[2] Romain Gauriau, Catriona Bridge, Lu Chen, Ben Glocker, Joseph Hajnal, Daniel Rueckert, and Wenjia Bai. Using DICOM metadata for radiological image series categorization: A feasibility study on large clinical brain MRI datasets. *Journal of Digital Imaging*, 33(3):747–762, 2020.

[3] Jonas Denck, Jens Guehring, Andreas Maier, and Eva Rothgang. MR-contrast-aware image-to-image translations with generative adversarial networks. *International Journal of Computer Assisted Radiology and Surgery*, 16(12):2069–2078, June 2021.

[4] Lianrui Zuo, Yihao Liu, Yuan Xue, Blake E. Dewey, Samuel W. Remedios, Savannah P. Hays, Murat Bilgel, Ellen M. Mowry, Scott D. Newsome, Peter A. Calabresi, Susan M. Resnick, Jerry L. Prince, and Aaron Carass. HACA3: A unified approach for multi-site MR image harmonization. *Computerized Medical Imaging and Graphics*, 109:102285, 2023.

[5] Jintang Ouyang, Ehsan Adeli, Kilian M Pohl, Qingyu Zhao, and Gregory Zaharchuk. Representation disentanglement for multi-modal brain MRI analysis. In *Information Processing in Medical Imaging (IPMI)*, volume 12729 of *Lecture Notes in Computer Science*, pages 321–333. Springer, 2021.

[6] Alec Radford, Jong Wook Kim, Christopher Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021.

[7] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024.

[8] Haoran Wang, Ke Liu, Nathan Ng, et al. MedCLIP: Contrastive learning from unpaired medical images and text. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022.

[9] Junde Wu, Yusheng Zhang, Yuanpu Xie, Tao Ma, and et al. PMC-CLIP: Contrastive vision-language pretraining on biomedical literature. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[10] Yulin Wang, Honglin Xiong, Kaicong Sun, Shuwei Bai, Ling Dai, Zhongxiang Ding, Jiameng Liu, Qian Wang, Qian Liu, and Dinggang Shen. Toward general text-guided multimodal brain mri synthesis for diagnosis and medical image analysis. *Cell Reports Medicine*, 6(6):102182, 2025.

[11] Ruixuan Du and Vin Vardhanabhuti. 3D-RADNet: Extracting labels from DICOM metadata for training general medical domain deep 3D convolution neural networks. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning*, volume 121, pages 174–192. PMLR, 2020.

[12] Jeff W McDaniel, Anna Z Moore, Thomas H Mareci, Stephen L Price, Diane J Conklin, Neal Wagle, Marcelo C Pinho, and Bennett A Landman. Improving the automatic classification of brain MRI acquisition contrast with machine learning. *Academic Radiology*, 2022.

[13] Shuai Liang, Derek Beaton, Stephen R. Arnott, Tom Gee, Mojdeh Zamyadi, Robert Bartha, Sean Symons, Glenda M. MacQueen, Stefanie Hassel, Jason P. Lerch, Evdokia Anagnostou, Raymond W. Lam, Benicio N. Frey, Roumen Milev, Daniel J. Müller, Sidney H. Kennedy, Christopher J. M. Scott, ONDRI Investigators, and Stephen C. Strother. Magnetic Resonance Imaging sequence identification using a metadata learning approach. *Frontiers in Neuroinformatics*, 15:622951, 2021.

[14] National Electrical Manufacturers Association. Digital imaging and communications in medicine (DICOM) standard. `https://www.dicomstandard.org/`, May 2025. NEMA PS3 / ISO 12052.

[15] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[16] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2019.

[17] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

A Appendix

Data Data usage approved under [anonymised]. Full list of used DICOM tags are given in Table 2 and distribution of tags are given in Fig. 4. All 3D MRI volumes are rigidly registered to the MNI template space and skull-stripped. From each registered volume, we extract a representative subset of slices by selecting every second slice from the central 100 slices, capturing the most diagnostically relevant anatomy while controlling dataset size. Acquisition plane (axial, coronal, sagittal) is determined from voxel resolution, with the highest-resolution dimension chosen as the slicing axis; for isotropic volumes, the axial plane is selected by default. Full pipeline is given in code repository.

Implementation Details MR-CLIP is implemented in PyTorch and trained on three NVIDIA A100 GPUs (40 GB each) with a batch size of 3000 for 2D and 150 for 3D per GPU, using sharded loss as in the CLIP implementation [17]. Optimization uses Adam (lr $= 1e{-}4$, $\beta_1 = 0.9$, $\beta_2 = 0.98$) with weight decay 0.2, over 100 epochs with 2000 warm-up steps. Gradient checkpointing reduces memory usage, while patch dropout (0.5) and text dropout (0.2) are applied alongside standard image augmentations, including random affine transforms, resized crops, Gaussian blur, and horizontal flips. The codes are built upon OpenCLIP repository (`https://github.com/mlfoundations/open_clip`) (License provided in repository)

Table 2: DICOM metadata fields used in MR-CLIP for contrast and sequence representation learning.

DICOM Tag
Magnetic Field Strength
Manufacturer
Manufacturer's Model Name
Series Description
Scanning Sequence
Sequence Variant
Acquisition Plane (extracted from voxel size)
Echo Time (TE)
Repetition Time (TR)
Inversion Time (IR)
Flip Angle

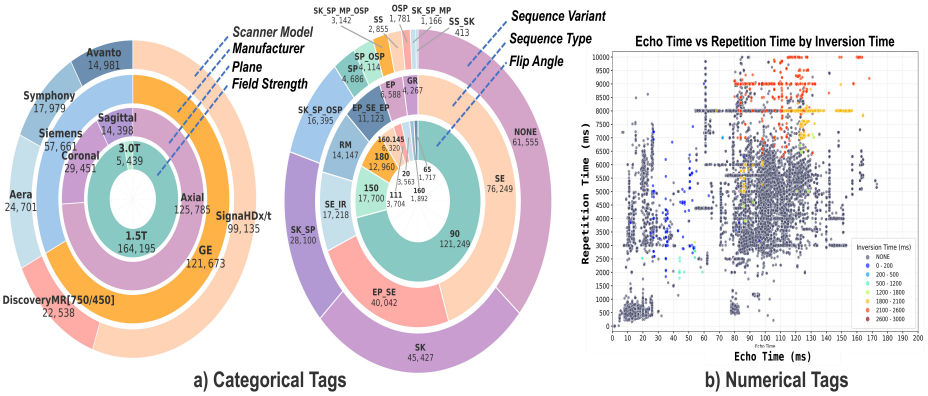

Figure 4: Overview of metadata distribution in our dataset. (a) Categorical tags including scanner, plane, field strength, and sequence information and flip angle. (b) Numerical distribution of echo and repetition times, color-coded by inversion time.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that the paper introduces a metadata-guided CLIP framework for 3D MRI, which aligns volumetric images with their DICOM metadata to learn contrast-aware embeddings. These claims are consistent with the described methodology and supported by experiments demonstrating contrast-aware representation learning, retrieval performance, and few-shot classification capabilities.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations related to the metadata aspect, emphasizing that the model's performance depends on the completeness, accuracy, and consistency of DICOM metadata. It acknowledges that missing or noisy metadata fields may affect the reliability of image–metadata alignment and downstream performance, reflecting an awareness of the scope and assumptions tied to metadata quality.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper focuses on the empirical and methodological aspects of metadata-driven representation learning for MRI rather than formal theoretical development. It does not include theorems or proofs, as the work is primarily experimental and applied in nature.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide detailed explanations of our dataset processing, model implementation, and training procedures, and our code is publicly available. While the primary clinical dataset is private, the code can be applied to public datasets to reproduce the main results. In the paper, we include the key details necessary to support our claims, and readers can refer to the released code for full reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide public access to the code, including detailed instructions for training and evaluation. Although the dataset is private due to clinical restrictions, the code repository clearly describes the data preprocessing steps, model training setup, and evaluation protocols. This level of detail allows other researchers to reproduce the methodology on similar datasets, and the main experimental claims can be verified using the provided code and described procedures.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The Methods section and the Appendix provide the details about the experimental setup, and code repository is referred where necessary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: While we do not report traditional error bars or confidence intervals, our experiments evaluate multiple model variants (2D, 2.5D, 3D MR-CLIP), few-shot settings, and different retrieval/classification metrics to demonstrate consistent performance trends. For key experiments such as linear-probe metadata classification and cross-modal retrieval, results are stable across slices, volumes, and acquisition types, which provides practical evidence of robustness and reproducibility of our main claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The necessary information is provided in Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper conforms with NeurIPS Code of Ethics.

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our work aims to improve the robustness, and scalability of medical imaging AI by aligning MRI scans with their DICOM metadata, enabling contrast-aware and data-efficient learning without requiring manual annotations. This has clear positive societal impacts in healthcare, such as accelerating clinical AI development, facilitating harmonization across scanners and sites, and reducing biases introduced by inconsistent metadata or manual labeling. However, potential negative impacts include the risk of model misuse in clinical decision-making without proper validation, or unintended bias propagation if the training metadata reflect institutional or demographic imbalances. To mitigate these risks, our work focuses on foundation-level representation learning rather than diagnostic automation, and we advocate for responsible use under clinical supervision and open benchmarking on diverse datasets.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There is no high risk for misuse of the model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.

- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We build on OpenCLIP repository which is credited and referred to their code repository for the license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release our codes and documentation is provided in the code repository.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Data usage ethics approval is stated in Appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The usage of LLMs is not a component of this paper

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.