SARAR: Shortcut-Aware 3D Brain MRI Question Answering via Retrieval-Augmented Reranking

Mohammad H. Abbasi, Bailey Trang, Shaurnav Ghosh, Favour Nerrise, Yuncong Mao, Mohammad Asadi, Juze Zhang, and Ehsan Adeli*

Stanford University, Stanford, CA 94305, USA
{mabbasi,eadeli}@stanford.edu

Abstract. 3D brain MRI is essential for clinical diagnosis and treatment planning, yet it remains underexplored in medical imaging AI research. A critical challenge in medical visual question answering (VQA) is that models often exploit text shortcuts, answering from linguistic patterns rather than analyzing the input images. We propose SARAR, Shortcut-Aware Retrieval-Augmented Reranking, a multimodal 3D brain MRI question answering framework with two key technical contributions: (1) image dropout training with random label replacement that forces genuine image reliance by removing text-only learning signals, and (2) retrieval-augmented reranking that recovers and surpasses this gap by leveraging visually similar cases. Our architecture combines frozen 3D VAE encoders with a LoRA-tuned language model to process T1-weighted and T2/FLAIR sequences for three question types: region localization (4-way MCQ), diagnosis classification (4-way MCQ), and yes/no clinical findings. On our curated dataset of 520 scans and 2,020 QA pairs, SARAR achieves 73.3% on Diagnosis, 63.7% on Region, and 69.9% on Yes/No, outperforming zero-shot VLMs on image grounding metrics and the 13B-parameter RadFM while using only 0.5B parameters. Unlike other models, when images are replaced with blank input, SARAR's accuracy drops to the chance level, providing empirical evidence that predictions depend on visual input under the blank-image audit. In contrast, state-of-the-art VLMs maintain or even improve accuracy without images, revealing text shortcut exploitation.

Keywords: Medical VQA · 3D Brain MRI · Retrieval-Augmented Reranking · Vision-Language Models · Multimodal Learning · Explainable AI

1 Introduction

Recent medical VQA systems report strong performance, yet it remains unclear whether models truly rely on medical images or exploit linguistic shortcuts. Radiologists routinely examine 3D brain MRI scans to localize abnormalities, identify pathologies, and assess clinical findings. Building AI systems that can answer structured clinical questions about these scans would support clinical

* Corresponding author

workflows, but creating such systems is challenging. Unlike 2D medical images, where vision-language models (VLMs) have shown promise [24, 19], 3D volumetric data requires specialized architectures and carefully curated training data.

A fundamental problem in medical visual question answering (VQA) is that models often learn textual shortcuts, exploiting statistical regularities in questions or answer distributions rather than genuinely understanding image content. Recent analyses show that linguistic priors can be exploitable in medical VQA benchmarks [6, 12], yet existing 3D brain MRI benchmarks [3, 28] have not systematically quantified shortcut reliance using such audits.

In this work, we propose SARAR (**S**hortcut-**A**ware 3D Brain MRI Question Answering via **R**etrieval-**A**ugmented **R**eranking), a framework that addresses shortcut biases through two mechanisms. First, during training, we randomly drop image tokens with probability p=0.15 and replace target labels with random choices, preventing text-only shortcuts; Second, at inference time, we apply retrieval-augmented reranking using visually similar training cases to refine predictions. We further leverage dual-modality input (T1 and T2/FLAIR MRI scans) to capture complementary anatomical and pathological information.

On our curated benchmark of 2,020 QA pairs across 520 brain MRI scans, SARAR achieves 73.3% on Diagnosis, 63.7% on Region, and 69.9% on Yes/No tasks, substantially above chance and outperforming state-of-the-art VLMs. Crucially, when visual inputs are replaced by blank images, accuracy drops to chance level, indicating strong dependence on visual input in contrast to zero-shot VLMs, which show minimal degradation without images.

Our contributions in this study are summarized as follows: (1) a shortcut-audited evaluation framework for 3D brain MRI VQA that reveals systematic differences in image grounding; (2) a training strategy promoting image dependence via image dropout with random label replacement; (3) a case-based clinical reasoning mechanism via retrieval that recovers and surpasses accuracy; and (4) demonstration that dual-modality input strengthens image reliance.

2 Related Work

Medical Visual Question Answering. Medical VQA benchmarks have historically focused on 2D images, including VQA-RAD [14], SLAKE [18], VQA-Med [4], and PMC-VQA [34]. For 3D imaging, M3D [3] released a multimodal dataset including VQA tasks, and mpLLM [28] targets 3D brain mpMRI VQA. In recent years, efforts have been made to develop 3D MRI foundation models [31]. However, none of these explicitly control for text-only shortcuts or provide systematic shortcut-aware evaluation with image-removed auditing, which we address here.

Vision-Language Models in Medicine. Recent progress in medical vision-language models (VLMs) has followed two primary directions: large-scale contrastive pretraining and instruction-tuned multimodal assistants. Contrastive approaches such as MedCLIP [29] and BiomedCLIP [33] align medical images with textual descriptions using paired datasets, enabling zero-shot transfer to down-

stream tasks. Instruction-tuned models, including LLaVA-Med [16] and Med-Flamingo [22], extend this paradigm by adapting large language models to medical reasoning through multimodal instruction tuning. More recently, RadFM [30] expands multimodal foundation modeling to encompass both 2D and 3D data, highlighting the importance of volumetric representations in clinical imaging.

Despite these advances, emerging evidence suggests that reported gains may partially stem from language shortcuts and dataset biases rather than genuine multimodal reasoning. In particular, Butsanets et al. [6] demonstrate that text-only baselines achieve non-trivial performance on widely used medical VQA benchmarks, revealing the presence of strong linguistic priors and shortcut learning. Such shortcutability undermines claims of image-grounded reasoning and raises concerns about model reliability in clinical settings. In contrast, our work explicitly quantifies shortcutability and shows that our model's accuracy drops to near-chance levels when images are removed, providing empirical evidence that predictions are genuinely image-conditioned. This design prioritizes image-conditioned multimodal reasoning over superficial statistical correlations, creating a more clinically trustworthy system.

Retrieval-Augmented Methods. Retrieval augmentation has been explored for medical VQA [32] and chest X-ray tasks [27, 8]. Unlike complex fusion architectures, we adopt a simpler strategy: retrieving visually similar training cases to adjust prediction scores at inference. This case-based refinement is particularly effective for diagnostic tasks, where morphologically similar cases usually share underlying pathology, and complements our shortcut-resistant training by reinforcing image-grounded reasoning.

3 Method

Our framework consists of three main components: 3D visual encoders for visual modalities (T1 and T2/FLAIR), a language model, and a retrieval module for inference-time augmentation. Figure 1 illustrates the overall architecture.

3.1 Architecture

Preprocessing. All MRI volumes undergo standardized preprocessing using the sMRI Processing Pipeline [1]: skull stripping, bias field correction, intensity normalization, and resampling to 128^3 voxels with 1mm isotropic resolution.

3D Visual Encoder. We use two separate frozen 3D AutoencoderKL models [23], one for T1-weighted and one for T2/FLAIR sequences. These encoders were trained in a self-supervised manner on approximately 51,000 brain MRI scans from ADNI [11], PPMI [20], HCP [26], ABCD [7], AIBL [9], BraTS [21], and WMH [13]. Each encoder produces a 32-channel latent representation.

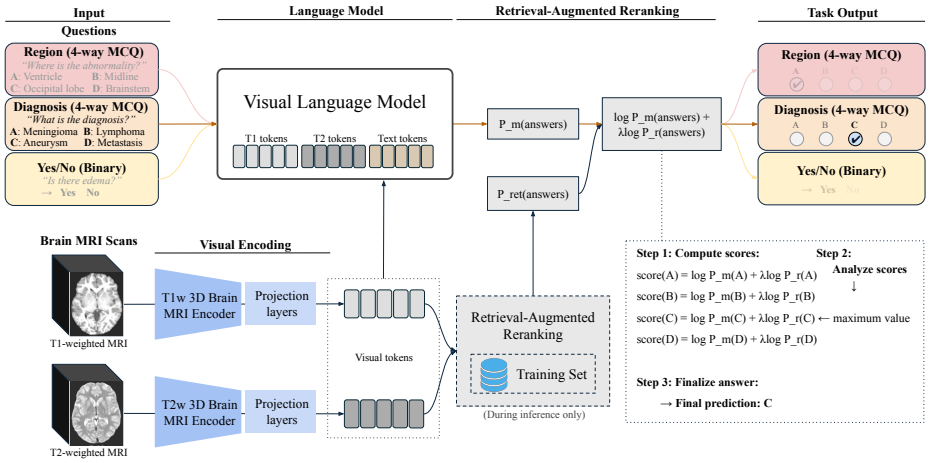

Fig. 1. Overview of SARAR. **Left:** The shortcut problem models may answer from text patterns alone without examining images. **Middle:** Image dropout training with random label replacement breaks text-only shortcuts by forcing image reliance. **Right:** Retrieval-augmented reranking recovers accuracy using visually similar cases, enabling both shortcut-awareness and high performance.

Visual Projection and Fusion. We flatten the VAE latent output and apply adaptive average pooling to obtain K=64 visual tokens per modality. A linear projection layer (29.6K trainable parameters) maps these tokens to the LLM hidden dimension (d=896). For dual-modality models, we concatenate T1 and T2/FLAIR tokens to form 128 visual tokens total.

Language Model. We use Qwen2.5-0.5B-Instruct [25] with LoRA adapters [10] (r=8, α=16), yielding 1.08M trainable parameters (0.22% of backbone). For MCQ, we select the answer token (A/B/C/D) with the highest log-probability; for binary questions, we compare the probability of tokens "True" and "False", selecting the one with higher probability.

3.2 Image Dropout Training

To mitigate text-only shortcuts, we apply image dropout during training (p=0.15), a technique inspired by modality dropout in multimodal learning [2]. When image dropout is activated, all visual tokens are zeroed out, and the target label is replaced with a uniformly random value (1/4 for multiple choice and 0.5 for Yes/No). This label randomization acts as a regularizer: by removing a consistent mapping between question-only inputs and target labels during dropped-image updates, it discourages the model from learning text-only heuristics rather than treating the random targets as meaningful supervision. For optimization, we use focal loss [17] with γ=2.0 for multiple-choice questions and standard cross-

entropy for Yes/No classification, and we apply weighted random sampling to address class imbalance.

3.3 Retrieval-Augmented Reranking

At inference, we retrieve k=10 nearest neighbors using cosine similarity of spatially-averaged VAE latents (d=64 for the concatenated T1+T2 representation). We compute a retrieval prior by aggregating neighbor labels weighted by similarity:

$$P_{\text{ret}}(a \mid x) = \frac{\sum_{(x_i,y_i)\in\mathcal{N}(x)} \text{sim}(x, x_i) \cdot \mathbf{1}[y_i = a]}{\sum_{(x_i,y_i)\in\mathcal{N}(x)} \text{sim}(x, x_i)} \tag{1}$$

The final prediction combines model and retrieval scores:

$$\text{score}(a) = \log P_{\text{model}}(a) + \lambda \cdot \log P_{\text{ret}}(a) \tag{2}$$

where λ is tuned per-task on validation data. Diagnosis benefits most (λ=5), while Yes/No uses λ=0. Note that retrieval aggregates over semantic class labels (e.g., diagnosis category, anatomical region), not answer letters; the retrieved class distribution is then mapped to the query's shuffled answer ordering.

4 Experiments

We train for 5,000 steps with batch size 1 using AdamW (learning rates: 5×10^{-5} projection layer, 5×10^{-6} LoRA) and image dropout p=0.15. Training takes 16 hours on one A100 GPU. At inference, we retrieve k=10 neighbors and tune λ per-task on validation data.

Dataset. Our QA dataset is constructed from BraTS-GLI [21], BraTS-MEN [21], and WMH Challenge [13], using radiologist-authored reports from RadGenome-Brain MRI [15], which is, to our knowledge, the only publicly available source of structured radiology reports for 3D brain MRI. We generate 2,020 QA pairs: **Region** (4-way MCQ, 520), **Diagnosis** (4-way MCQ, 520), and **Yes/No** (980). Following MicroVQA [5], we intentionally use fixed templates and per-sample answer shuffling to eliminate linguistic shortcuts. Template-based evaluation has been shown to reduce text-only exploit ability by removing superficial correlations between question phrasing and answers [5]. Data is split at the subject level with strictly patient-disjoint splits (416/52/52 subjects for the train/validation/test sets, zero overlap), ensuring retrieval benefits arise from pathological similarity rather than data leakage. All reported results use this single fixed split; no repeated cross-validation is performed. We use macro accuracy for MCQ and balanced accuracy for the Yes/No setting. Our goal is controlled shortcut auditing rather than large-scale representation learning. The frozen visual encoders were self-supervised on a broader collection that includes BraTS and WMH, so dataset-source overlap exists between encoder pretraining and the downstream benchmark. No diagnostic or VQA labels were used during encoder pretraining, and the downstream train/validation/test splits remain strictly patient-disjoint.

4.1 Main Results

Table 1 shows results across modalities and blank-image ablation. Because SA

Table 1. Comparative results (accuracy %) across modalities and training settings. The SA model is the base architecture without the Retrieval-Augmented Reranking (RAR), SA + RAR (SARAR) includes both, Blank denotes the blank-image ablation. Methods are evaluated on Region (Reg.), Diagnosis (Diag.), and Yes/No (YN) tasks, highlighting the impact of retrieval and multimodal components. For the blank case, the T1+T2 experiment drops to exactly chance, confirming the shortcut-aware design.

Method→	SA Base			SARAR (Ours)			SARAR (Blank)			RadFM [30] (0-shot)		
Modality↓	Reg.	Diag.	YN	Reg.	Diag.	YN	Reg.	Diag.	YN	Reg.	Diag.	YN
T1	59.1	39.9	69.9	62.9	67.9	69.9	38.4	37.5	50.0	—	—	—
T2	61.0	42.7	69.9	65.1	67.5	69.9	35.5	39.2	50.0	—	—	—
T1+T2	56.1	45.3	69.9	**63.7**	**73.3**	**69.9**	25.0	25.0	50.0	7.0	37.6	—
Random Baseline	25.0	25.0	50.0	25.0	25.0	50.0	25.0	25.0	50.0	25.0	25.0	50.0

Base and SARAR use the same trained architecture and differ only by inference-time retrieval reranking, their performance gap directly isolates the contribution of RAR. Retrieval provides the largest gains on Diagnosis: +28 points for T1+T2 (45.3%→73.3%), with McNemar's test confirming significance ($p<0.01$). While the base model achieves 45.3% on Diagnosis (well above 25% chance), retrieval leverages the medical prior that visually similar scans often share pathology; results are robust to λ within ± 2; large λ for Diagnosis reflects strong pathological similarity among neighbors. The optimal retrieval weight varies by task: λ=5 for Diagnosis, λ=2 for Region, and λ=0 for Yes/No. Results are robust to $k \in \{5, 10, 15\}$. Yes/No accuracy (69.9%) reflects a learned clinical prior rather than per-image reasoning consistent with limited template diversity. We therefore focus evaluation primarily on Region and Diagnosis, which demonstrate clear retrieval benefits and image grounding.

We also compare against RadFM [30], a 13B-parameter 3D medical VLM: on zero-shot evaluation, RadFM achieves only 7.0% on Region and 37.6% on Diagnosis, substantially below our 0.5B model.

Without visual input, T1+T2 accuracy drops to exactly chance level, indicating that the model retains little usable text-only signal under this blank-image audit. Single-modality models retain some residual accuracy (35–39%), suggesting dual-modality dropout provides stronger regularization. Together, the blank-image audit and the SA Base-SARAR comparison show complementary roles: the training strategy reduces reliance on text-only signals, while retrieval substantially improves task accuracy.

Clinical Insights. The performance pattern aligns with clinical workflows: Diagnosis benefits most from retrieval (+28 points), paralleling case-based reasoning in radiology [8], whereas Region relies on per-image anatomy. Dual-modality

Table 2. 2D VLMs evaluated via 10-pass axial slice sampling to maximize volumetric coverage; architectural 3D processing is unavailable to these models by design.

Model	Yes/No			Region			Diagnosis		
	Img	Blk	Δ	Img	Blk	Δ	Img	Blk	Δ
gpt-5	72.5	55.2	-17	30.8	25.8	-5	49.4	35.0	-14
gemini-2.5-pro	63.8	31.9	-32	18.5	9.2	-9	64.8	3.9	-61
llama-4-maverick	65.5	61.9	-4	20.8	25.0	+4	53.3	30.6	-23
llama-4-scout	61.7	50.0	-12	30.0	24.2	-6	50.6	21.7	-29
qwen3-vl-32b	62.6	50.0	-13	19.2	24.2	+5	46.7	36.1	-11
claude-3.7-sonnet	27.6	36.0	+8	25.0	17.5	-8	42.8	13.3	-29
gpt-5-nano	59.8	50.0	-10	30.8	23.3	-8	35.6	1.7	-34
Ours (T1+T2)	**69.9**	50.0	-20	**63.7**	25.0	-39	**73.3**	25.0	-48

eliminates residual shortcuts, reflecting T1/T2 complementarity. Unlike models exploiting linguistic priors [6, 12], our accuracy drops to chance under blank-image auditing a necessary property for clinically reliable AI.

4.2 Ablation Studies

Why Zero-Shot VLM Comparison? Our VLM comparison evaluates shortcut reliance, not architectural superiority. Since 2D VLMs cannot process 3D volumes directly, we evaluate each scan 10 times with different axial slice sets, ensuring comprehensive coverage of the full volume within each model's context limit. Crucially, if our benchmark contained text shortcuts, VLMs should also drop to chance with blank input yet several improve (Table 2), confirming our template design eliminates shortcuts while VLMs exploit pretraining priors.

Comparison with Zero-Shot VLMs. Table 2 and Figure 2 compare with VLMs in a zero-shot setting to evaluate image grounding. Our goal is to demonstrate text shortcut exploitation, observable regardless of fine-tuning. While most VLMs show decreased accuracy with blank input, none drop to exactly chance level. Some VLMs show *improved* performance on certain tasks (e.g., claude-3.7-sonnet on YesNo: +8%, qwen3-vl-32b on Region: +5%), indicating text shortcut exploitation rather than genuine image grounding.

Qualitative Examples. Figure 3 shows correct predictions and failure modes. Common errors include region confusion between adjacent areas and glioblastoma-meningioma misclassification both enhancing lesions with overlapping features.

Clinical Implications. Shortcut-exploiting models risk clinical hallucination by generating plausible answers without examining images. Radiology requires reliable use of image evidence for diagnosis. Our shortcut-aware framework enables measuring whether models truly ground predictions in visual content, a prerequisite for trustworthy clinical translation.

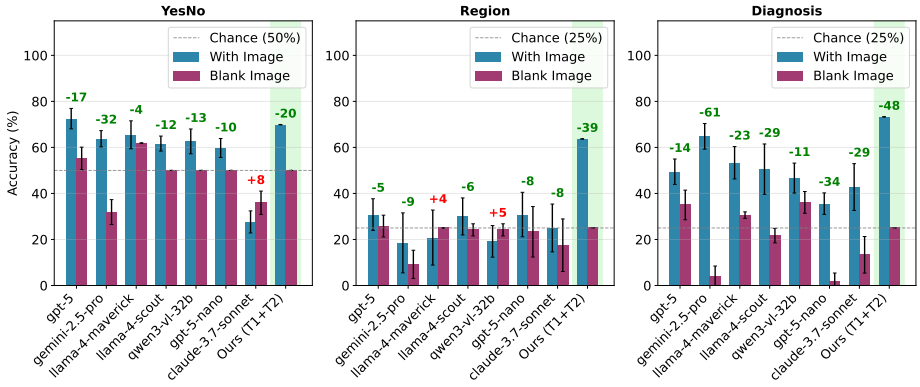

Fig. 2. Image grounding comparison. Blue: with image; Red: blank input. Δ above bars shows accuracy change. Our model (highlighted) achieves highest accuracy while dropping to chance with blank input, unlike VLMs which often *improve* without images.

5 Conclusion

We presented SARAR, a shortcut-aware framework for 3D brain MRI VQA combining image dropout training with retrieval-augmented reranking. Our method achieves 73.3% on Diagnosis and 63.7% on Region while dropping to exactly chance without images, in contrast to VLMs which often *improve* without images, revealing text shortcut exploitation. Beyond performance gains, this work establishes shortcut auditing as an essential evaluation protocol for medical multimodal AI without which claims of image grounding remain unverified.

Limitations and Future Work. This study is positioned as a controlled proof-of-concept for shortcut-aware evaluation in 3D brain MRI VQA. Our benchmark (520 scans, 2,020 QA pairs across three pathology categories) reflects the currently available public structured radiology reports, with RadGenome-Brain MRI being the only such source. While the scale is sufficient to systematically reveal shortcut behavior and grounding differences across models, expanding to larger multi-institutional datasets with broader pathology diversity is an important next step. We adopt template-based questions as a deliberate design choice to eliminate linguistic shortcuts and enable controlled shortcut auditing; extending to open-vocabulary QA under shortcut-controlled settings remains future work. Further clinical validation and prospective studies will be necessary to assess real-world deployment and establish shortcut auditing as standard practice in medical multimodal AI evaluation. Future work should also evaluate mismatched-patient image controls and spatial attribution methods to distinguish general image dependence from anatomically correct visual grounding.

Acknowledgments. This research was supported in part by the National Institutes of Health grant numbers AG089169, AG047366 and AG084471. The content is solely

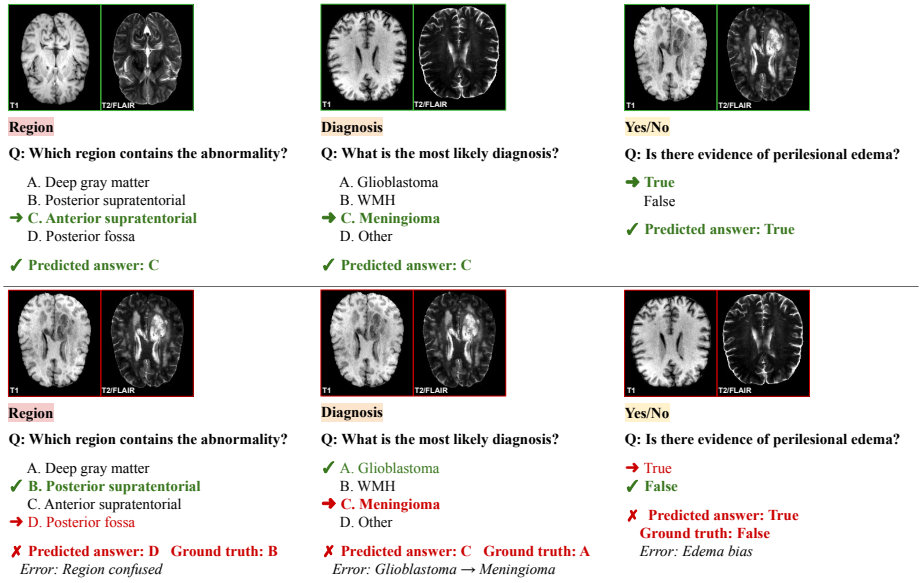

Fig. 3. Qualitative results. Top: correct predictions. Bottom: failure cases.

the responsibility of the authors and does not necessarily represent the official views of the NIH. This study was also supported by the Stanford University Institute for Human-Centered AI (HAI) Hoffman-Yee Award.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abbasi, M.H., Adeli, E.: sMRI processing pipeline: A lightweight, end-to-end workflow for structural brain MRI preprocessing and quality control (2025). https://doi.org/10.5281/zenodo.17503175, https://doi.org/10.5281/zenodo.17503175
2. Bachmann, R., Mizrahi, D., Atanov, A., Zamir, A.: Multimae: Multi-modal multi-task masked autoencoders. In: European Conference on Computer Vision (ECCV). pp. 348–367 (2022)
3. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578 (2024)
4. Ben Abacha, A., Sarrouti, M., Demner-Fushman, D., Hasan, S.A., Müller, H.: Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In: Proceedings of the CLEF 2021 Conference and Labs of the Evaluation Forum-working notes. 21-24 September 2021 (2021)
5. Burgess, J., et al.: Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19552–19564 (June 2025)

6. Butsanets, L., Corbière, C., Khlaut, J., Manceron, P., Dancette, C.: Radimagenet-vqa: A large-scale ct and mri dataset for radiologic visual question answering. arXiv preprint arXiv:2512.17396 (2025)
7. Casey, B.J., et al.: The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites **32**, 43–54 (2018)
8. Choe, J., et al.: Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct. Radiology **302**(1), 187–197 (2022)
9. Ellis, K.A., et al.: The australian imaging, biomarkers and lifestyle (aibl) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of alzheimer's disease. International psychogeriatrics **21**(4), 672–687 (2009)
10. Hu, E.J., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
11. Jack Jr, C.R., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine **27**(4), 685–691 (2008)
12. Jiang, S., Chen, Y., Song, S., Zhang, Y., Jin, Y., Feng, Y., Wu, J., Liu, Z.: Knowing or guessing? robust medical visual question answering via joint consistency and contrastive learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 325–335. Springer (2025)
13. Kuijf, H.J., et al.: Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge **38**(11), 2556–2568 (2019)
14. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific data **5**(1), 180251 (2018)
15. Lei, J., et al.: Autorg-brain: Grounded report generation for brain mri. arXiv preprint arXiv:2407.16684 (2024)
16. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
17. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
18. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In: 2021 IEEE 18th international symposium on biomedical imaging (ISBI). pp. 1650–1654. IEEE (2021)
19. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
20. Marek, K., et al.: The parkinson progression marker initiative (ppmi). Progress in neurobiology **95**(4), 629–635 (2011)
21. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014)
22. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P.: Med-flamingo: a multimodal medical few-shot learner. In: Machine learning for health (ML4H). pp. 353–367. PMLR (2023)
23. Pinaya, W.H., et al.: Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 (2023)

24. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
25. Team, Q., et al.: Qwen2. 5: A party of foundation models (2024)
26. Van Essen, D.C., et al.: The wu-minn human connectome project: an overview. Neuroimage **80**, 62–79 (2013)
27. Van Sonsbeek, T., Worring, M.: X-tra: Improving chest x-ray tasks with cross-modal retrieval augmentation. In: International Conference on Information Processing in Medical Imaging. pp. 471–482. Springer (2023)
28. Vepa, A.M., Yu, Y., Gan, J., Cuturrufo, A., Li, W., Wang, W., Scalzo, F., Sun, Y.: A multimodal llm approach for visual question answering on multiparametric 3d brain mri. arXiv preprint arXiv:2509.25889 (2025)
29. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
30. Wu, C., Zhang, X., Zhang, Y., Hui, H., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Nature Communications **16**(1), 7866 (2025)
31. Yang, H., Wu, Z., Wang, S., Chen, X., Jiang, X., Ma, H., Zhao, L., Zhang, D., Zhu, Y.: Deciphermr: An end-to-end multimodal generalist model for brain mri. arXiv preprint arXiv:2509.21249 (2025)
32. Yuan, Z., Jin, Q., Tan, C., Zhao, Z., Yuan, H., Huang, F., Huang, S.: Ramm: Retrieval-augmented biomedical visual question answering with multi-modal pre-training. In: Proceedings of the 31st ACM international conference on multimedia. pp. 547–556 (2023)
33. Zhang, S., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
34. Zhang, X., Wu, C., Zhao, Z., Lin, W., Zhang, Y., Wang, Y., Xie, W.: Pmc-vqa: Visual instruction tuning for medical visual question answering. arXiv preprint arXiv:2305.10415 (2023)