

# Improving Factuality of 3D Brain MRI Report Generation with Paired Image-domain Retrieval and Text-domain Augmentation

Junhyeok Lee<sup>1#\*</sup>, Yujin Oh<sup>2,3#</sup>, Dahyoun Lee<sup>3</sup>, Hyon Keun Joh<sup>4</sup>, Minchul Kim<sup>5</sup>, Chul-Ho Sohn<sup>6,7</sup>, Sung Hyun Baik<sup>6,8</sup>, Cheolkyu Jung<sup>6,8</sup>, Jung Hyun Park<sup>6,9</sup>, Kyu Sung Choi<sup>6,7†</sup>, Byung-Hoon Kim<sup>3,10,11,12†</sup>, and Jong Chul Ye<sup>13†</sup>

<sup>1</sup> Cancer Biology, Seoul National University College of Medicine, Korea

<sup>2</sup> Radiology, Massachusetts General Hospital and Harvard Medical School, USA

<sup>3</sup> Biomedical Systems Informatics, Yonsei University College of Medicine, Korea

<sup>4</sup> Graduate School, Yonsei University, Korea

<sup>5</sup> Radiology, Kangbuk Samsung Hospital, Korea

<sup>6</sup> Radiology, Seoul National University College of Medicine, Korea

<sup>7</sup> Radiology, Seoul National University Hospital, Korea

<sup>8</sup> Radiology, Seoul National University Bundang Hospital, Korea

<sup>9</sup> Radiology, SMG-SNU Boramae Medical Center, Korea

<sup>10</sup> Psychiatry, Yonsei University College of Medicine, Korea

<sup>11</sup> Behavioral Sciences in Medicine, Yonsei University College of Medicine, Korea

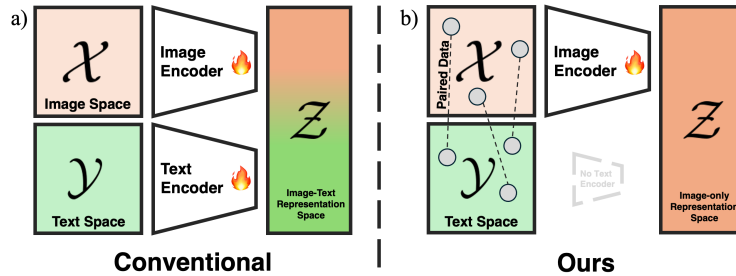
<sup>12</sup> Yonsei Institute for Digital Health, Korea

<sup>13</sup> Kim Jaechul Graduate School of AI, KAIST, Korea

ent1127@snu.ac.kr; egyptdj@yonsei.ac.kr; jong.ye@kaist.ac.kr

**Abstract.** Acute ischemic stroke (AIS) requires time-critical decision-making, where inaccurate interpretation of neuroimaging findings can lead to irreversible disability. Diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) maps from magnetic resonance imaging (MRI) are central to detecting acute infarction, yet generating factually reliable radiology reports directly from 3D MRI remains challenging due to the difficulty of learning robust cross-modal alignments between volumetric images and clinical text. We propose paired image-domain retrieval and text-domain augmentation (PIRTA), a retrieval-augmented generation framework that improves report factuality by avoiding explicit image-text alignment. PIRTA retrieves clinically similar 3D DWI/ADC volumes using a pretrained 3D vision encoder and leverages their paired clinician-authored reports to ground large language model (LLM)-based report generation. Experiments on multi-institutional in-house data, a held-out external privacy-preserving cohort, and the public ISLES benchmark demonstrate that PIRTA achieves strong image-domain retrieval performance and consistently improves ischemic-territory accuracy, a clinically grounded surrogate for report factuality, compared to direct image-to-text baselines. These results indicate that retrieval-grounded generation provides a scalable and reliable paradigm for producing factually consistent radiology reports from complex 3D brain MRI. Source code is available at <https://github.com/jhlee0619/PIRTA>.

\* # Co-first authors. † Co-corresponding authors.



**Fig. 1.** Conceptual comparison of cross-modal alignment strategies. (a) Conventional methods align image and text encoders in a shared representation space, requiring joint modeling across modalities. (b) PIRTA avoids explicit cross-modal alignment by retrieving similar images in the image domain, and using their paired clinician-authored reports for text-domain augmentation.

**Keywords:** Acute ischemic stroke · Diffusion weighted imaging · Radiology report generation · Retrieval-augmented generation · Large language models

## 1 Introduction

Stroke is the second leading cause of death worldwide and a major contributor to long-term disability and dementia [5]. Acute ischemic stroke (AIS), which accounts for approximately 70% of all stroke cases, results from abrupt interruption of cerebral blood flow and requires time-critical intervention, as treatment delays substantially increase the risk of irreversible neurological deficits [2]. Accordingly, current American Heart Association/American Stroke Association (AHA/ASA) guidelines emphasize rapid reperfusion therapy within narrow therapeutic windows to improve functional outcomes [20].

Magnetic resonance imaging (MRI) plays a central role in AIS diagnosis and management by delineating tissue viability and ischemic injury [21]. In particular, diffusion-weighted imaging (DWI) and apparent diffusion coefficient (ADC) maps enable sensitive detection of acute infarction and support assessment of hemorrhagic transformation risk [2]. These findings are summarized in radiology reports, which provide the clinically actionable interpretation guiding thrombolysis or thrombectomy decisions. While recent advances in artificial intelligence (AI) have enabled automated AIS detection and outcome prediction from MRI [19], most prior work has focused on classification or segmentation tasks rather than generating clinically usable radiology reports. Existing report generation approaches typically rely on supervised learning with extensive annotation or segmentation-based pipelines, limiting scalability and cross-institutional generalization [9, 3, 25, 13, 12, 17].

Radiology report generation from MRI is inherently challenging because it requires translating high-dimensional volumetric image evidence into structured clinical language. Although vision-language models (VLMs) and large language

models (LLMs) have demonstrated strong performance on 2D modalities such as chest X-ray and CT [24, 14, 16], reliable report generation from 3D MRI remains limited by the scarcity of pretrained 3D medical image encoders and the difficulty of learning robust cross-modal alignment between volumetric image representations and radiology text (Fig. 1a) [11, 18].

To mitigate these challenges, we propose paired image-domain retrieval and text-domain augmentation (PIRTA), a retrieval-augmented generation (RAG) framework [15] that improves report factuality by avoiding explicit image-text alignment (Fig. 1b). PIRTA retrieves clinically similar 3D DWI/ADC volumes from an in-domain paired image-report database using a pretrained 3D vision encoder, and leverages their paired clinician-authored reports as evidence to guide LLM-based report generation. To enable robust 3D retrieval, we pretrain a 3D Vision Transformer (ViT) [4] with MAE-based self-supervised learning [7] on large-scale unlabeled MRI.

Our contributions are as follows:

- We propose PIRTA, an image-domain retrieval and text-domain augmentation framework that avoids explicit cross-modal alignment, reducing optimization complexity for 3D MRI report generation.
- We develop a robust 3D DWI/ADC encoder through large-scale MAE pre-training, demonstrating significant improvements in cross-institution retrieval generalization.
- We evaluate factuality using clinically actionable ischemic-territory accuracy on multi-institutional internal data and public benchmarks, showing that PIRTA improves factual consistency by grounding generation in retrieved clinical evidence.

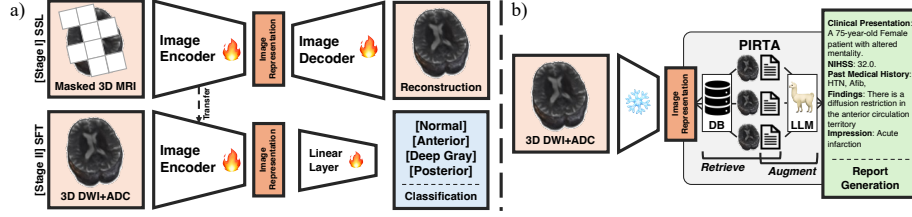
## 2 Methods

### 2.1 Problem Definition and Cross-modal Learning Complexity

Let  $\mathcal{X}$  denote the space of 3D brain MRI volumes and  $\mathcal{Y}$  the space of radiology reports; given a query image  $x_q \in \mathcal{X}$ , the objective is to generate or retrieve a clinically accurate report  $y \in \mathcal{Y}$ . Conventional image-to-text approaches jointly train an image encoder  $f_{\text{image}} : \mathcal{X} \rightarrow \mathcal{Z}$  and a text encoder  $f_{\text{text}} : \mathcal{Y} \rightarrow \mathcal{Z}$  into a shared latent space  $\mathcal{Z}$  by minimizing a distance between paired embeddings,

$$\min d(f_{\text{image}}(x_q), f_{\text{text}}(y)), \quad (1)$$

implicitly learning the joint distribution  $P(x, y)$  over a joint hypothesis space  $\mathcal{H} = \mathcal{H}_{\text{image}} \times \mathcal{H}_{\text{text}}$  with complexity  $C(\mathcal{H}) = C(\mathcal{H}_{\text{image}}) + C(\mathcal{H}_{\text{text}})$  [23]. This is particularly challenging for 3D medical imaging due to limited paired data and the representational gap between volumetric MRI and clinical text.



**Fig. 2.** Overview of PIRTA. (a) Training the 3D MRI image encoder: MAE-based SSL pretraining on unlabeled data followed by supervised fine-tuning for ischemic-territory classification. (b) Cross-modal RAG with PIRTA: using the frozen encoder to retrieve similar images and their paired reports, which augment LLM-based report generation.

## 2.2 PIRTA: Paired Image-domain Retrieval and Text-domain Augmentation

PIRTA reformulates report generation as in-domain image retrieval followed by text-domain augmentation. Given a paired database  $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$  of MRI volumes and clinician-authored reports, we train only an image encoder  $f_{\text{image}} : \mathcal{X} \rightarrow \mathcal{Z}$ . For a query  $x_q$ , PIRTA retrieves the most relevant case by minimizing cosine distance in the embedding space,

$$x^* = \arg \min_{x_i \in \mathcal{D}} d(f_{\text{image}}(x_q), f_{\text{image}}(x_i)), \quad (2)$$

and uses the paired report  $y^*$  as external evidence for report generation. This reduces the hypothesis space to  $\mathcal{H}_{\text{PIRTA}} = \mathcal{H}_{\text{image}}$ , focusing learning on the marginal  $P(x)$  rather than the joint  $P(x, y)$ . Grounding generation in clinician-authored reports retrieved purely in the image space anchors outputs in expert narratives and mitigates ungrounded generation, while reducing data and optimization burden. Fig. 2 overviews PIRTA, including MAE-based pretraining and retrieval-grounded report generation.

## 2.3 3D MRI Image Encoder Training

The image encoder is trained in two stages: self-supervised pretraining followed by supervised fine-tuning. For pretraining, we adopt a masked autoencoder (MAE) framework [7] with a 3D Vision Transformer (ViT) [4] backbone. DWI and ADC volumes are concatenated as a two-channel 3D input and partitioned into non-overlapping volumetric patches, a subset of which is randomly masked. The encoder processes the visible patches, while a lightweight decoder reconstructs the masked patches using mean squared error loss.

Following pretraining, the encoder is fine-tuned for ischemic-territory classification (anterior, deep gray, posterior, normal). The MAE decoder is replaced by a linear classification head, and the model is optimized using cross-entropy loss on labeled internal datasets. This two-stage training yields a 768-dimensional volumetric embedding for retrieval.

## 2.4 Retrieval-based Report Generation

After fine-tuning, latent embeddings of all training images are extracted to construct the retrieval database  $\mathcal{D}$ . Given a query MRI volume, cosine similarity is computed between its embedding and all database embeddings, and the top- $k$  most similar cases are retrieved. The paired radiology reports associated with the retrieved images are then used to augment report generation by a large language model (LLM).

Specifically, we fine-tune a large language model using low-rank adaptation (LoRA) [10] on structured radiological findings. During inference, retrieved reports and their similarity scores are injected into the prompt as external evidence, guiding the LLM to generate a final report. The report is structured into Findings and Impression sections and includes metadata fields (clinical presentation, NIHSS, and PMH) when available.

## 3 Experiments

### 3.1 Datasets

We use multi-institutional in-house MRI data from 1,831 adults (993 acute ischemic stroke cases and 838 controls) for model training and internal evaluation (Internal Cohort). External generalization is assessed using two independent datasets acquired at centers not included in the Internal Cohort: a held-out, privacy-preserving dataset (External Cohort 1,  $n = 580$ ) and the public multi-site ISLES benchmark (External Cohort 2,  $n = 206$ ) [8]. For self-supervised pre-training, we leverage unlabeled DWI/ADC volumes from 38,532 UK Biobank subjects [1]. All reports were transformed into structured findings templates curated by a certified clinical expert. Train/validation/test partitions are enforced strictly at the patient level, with no patient appearing in multiple splits; the retrieval database is built from training-set embeddings only, so validation and test images are never used as retrieval candidates.

### 3.2 Implementation details

The image encoder is pretrained with MAE-based SSL and fine-tuned for four-class territory classification on internal data. Retrieval uses cosine similarity across training-set embeddings to identify top- $k = 5$  neighbors. Report generation is performed using LLaMA3-8B-Instruct [6], instruction-tuned with LoRA on structured findings. LLM fine-tuning is conducted for 5 epochs with a batch size of 8 and LoRA rank  $r = 16$ . For inference, retrieved findings and their similarity scores are provided as textual evidence in the LLM prompt.

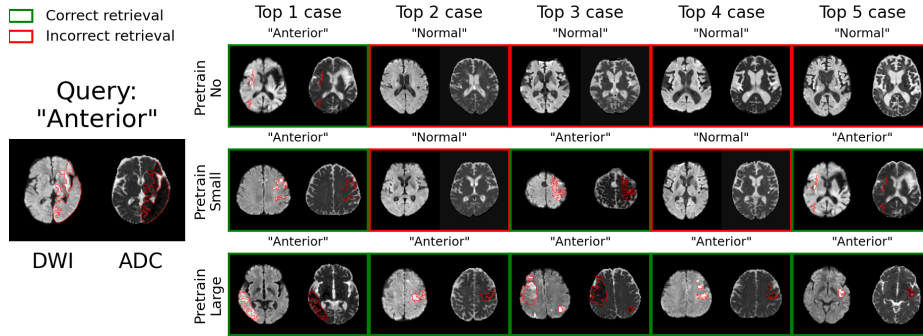
## 4 Results

### 4.1 Image-domain retrieval and representation generalization

We evaluate image-domain retrieval performance using mean average precision (mAP@ $k$ ) and top- $k$  accuracy (Acc@ $k$ ), as summarized in Table 1. While strong

**Table 1.** Image retrieval performance (%) versus pretraining size. *None*: random initialization; *Small*: MAE on the internal cohort only ( $n=1,831$ ); *Large*: MAE on  $n=38,532$  UK Biobank subjects.

| Dataset           | Pretrain | mAP@1        | mAP@5        | mAP@10       | Acc@1        | Acc@5        | Acc@10       |
|-------------------|----------|--------------|--------------|--------------|--------------|--------------|--------------|
| Internal Cohort   | None     | 80.73        | 81.90        | 81.57        | 80.73        | 84.40        | 84.40        |
|                   | Small    | 90.37        | 90.78        | 90.85        | 90.37        | 91.74        | 92.66        |
|                   | Large    | <b>94.04</b> | <b>94.27</b> | <b>94.15</b> | <b>94.04</b> | <b>94.50</b> | <b>94.50</b> |
| External Cohort 1 | None     | 56.72        | 58.22        | 57.94        | 56.72        | 61.55        | 64.48        |
|                   | Small    | 68.97        | 69.19        | 69.11        | 68.97        | 70.17        | 70.69        |
|                   | Large    | <b>71.21</b> | <b>71.43</b> | <b>71.30</b> | <b>71.21</b> | <b>71.72</b> | <b>71.72</b> |
| External Cohort 2 | None     | 38.83        | 40.48        | 40.42        | 38.83        | 44.17        | 48.06        |
|                   | Small    | 65.53        | 66.26        | 66.39        | 65.53        | 67.48        | 69.42        |
|                   | Large    | <b>70.87</b> | <b>71.52</b> | <b>71.24</b> | <b>70.87</b> | <b>72.33</b> | <b>72.82</b> |



**Fig. 3.** Qualitative retrieval examples for a query volume (left) and the top-5 retrieved scans (right). Green indicates correct retrieval; red indicates incorrect retrieval.

retrieval performance is achieved on the Internal Cohort (mAP@1 of 94.0%), the most pronounced gains are observed on the external cohorts, particularly External Cohort 2 (ISLES), indicating improved robustness under substantial domain shift.

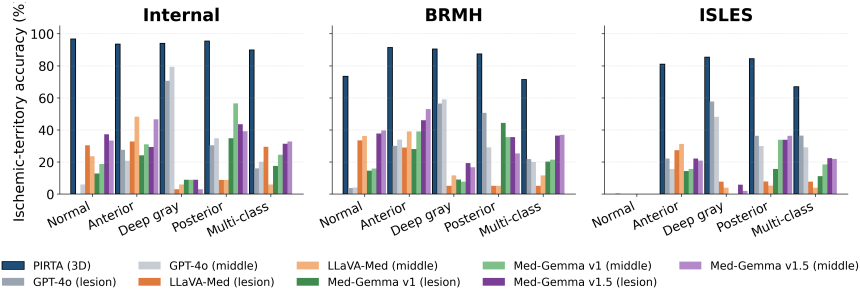
To further assess representation quality, we conduct an auxiliary sanity check using ischemic-territory classification with the same encoder (Table 2). Consistent with the retrieval results, large-scale MAE pretraining substantially improves multi-class classification accuracy across both internal and external cohorts, with especially large improvements on External Cohort 2. These parallel trends in retrieval and classification performance suggest that the encoder learns semantically meaningful and transferable 3D representations, rather than task-specific heuristics.

## 4.2 Clinical factuality evaluation of generated reports

We evaluate report factuality using ischemic-territory accuracy, a clinically actionable attribute that directly influences treatment decisions in acute stroke care.

**Table 2.** Effect of MAE pretraining scale on multi-class territory classification accuracy (Acc@1), shown as a sanity check of representation quality. *None*: random initialization; *Small*: MAE on the internal cohort ( $n=1,831$ ); *Large*: MAE on  $n=38,532$  UK Biobank subjects.

| Dataset           | Pretrain | Class-wise Acc@1 |              |              |              | Multi-class Acc@1 |
|-------------------|----------|------------------|--------------|--------------|--------------|-------------------|
|                   |          | Normal           | Anterior     | Deep gray    | Posterior    |                   |
| Internal Cohort   | None     | 96.3             | 87.96        | 89.81        | 89.81        | 81.94             |
|                   | Small    | 99.54            | 93.06        | 91.67        | 95.37        | 89.81             |
|                   | Large    | <b>100.0</b>     | <b>95.37</b> | <b>93.98</b> | <b>96.76</b> | <b>93.06</b>      |
| External Cohort 1 | None     | 60.87            | 78.26        | 86.29        | 83.95        | 53.18             |
|                   | Small    | 69.23            | 89.8         | 90.47        | 85.95        | 66.22             |
|                   | Large    | <b>72.24</b>     | <b>90.64</b> | <b>90.8</b>  | <b>87.79</b> | <b>69.23</b>      |
| External Cohort 2 | None     | -                | 69.52        | 76.67        | 68.1         | 40.0              |
|                   | Small    | -                | <b>84.29</b> | <b>86.67</b> | 80.95        | 61.9              |
|                   | Large    | -                | 82.38        | 84.76        | <b>83.81</b> | <b>68.1</b>       |



**Fig. 4.** Ischemic-territory accuracy for report generation. PIRTA (3D) is compared with 2D-slice baselines (GPT-4o, LLaVA-Med, and the medical-domain VLM Med-Gemma v1/v1.5). *Multi-class* denotes patients with  $\geq 2$  affected territories.

We compare PIRTA with both general-purpose 2D multimodal baselines (GPT-4o, LLaVA-Med [16]) and a medical-domain 2D vision-language model (Med-Gemma v1 and v1.5 [22]), each receiving an axial slice (mid-brain or lesion-centered when available) and the identical output-format prompt. PIRTA achieves substantially higher ischemic-territory accuracy than every 2D baseline (Fig. 4), leading the strongest 2D baseline by +57.2 / +34.5 / +30.6 multi-class accuracy points on the internal cohort, BRMH, and ISLES, with per-class gaps persisting across territories. This indicates that retrieval-grounded 3D evidence, not merely medical-domain pretraining of a 2D backbone, is the dominant factuality contributor, with failures concentrated at low-similarity, anatomically rare infarct patterns.

Table 3 shows PIRTA generates reports closely aligning with expert references; higher retrieval similarity consistently corresponds to better factual accuracy, directly linking retrieval quality to report factuality.

**Table 3.** Qualitative assessment of text-domain augmentation. Correct ischemic territory information and corresponding similarity scores are highlighted in blue

|                      |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                               |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>DWI &amp; ADC</b> | <p><b>Ground-truth:</b><br/> <b>Clinical Presentation:</b> A 58-year-old male patient.<br/> <b>NIHSS:</b> N/A <b>Past Medical History:</b> N/A<br/> <b>Findings:</b> Diffusion restriction in the <b>anterior circulation</b> territory.<br/> <b>Impression:</b> Acute infarction.</p> <hr/> <p><b>Generated report:</b><br/> <b>Clinical Presentation:</b> A 58-year-old male patient.<br/> <b>NIHSS:</b> N/A <b>Past Medical History:</b> N/A<br/> <b>Findings:</b> Diffusion restriction in the <b>anterior circulation</b> territory.<br/> <b>Impression:</b> Acute infarction.</p> <hr/> <p><b>Retrieved similar cases:</b><br/> [1] <b>Anterior circulation</b>, mild wedge-shaped (Score: <b>0.95</b>)<br/> [2] <b>Anterior circulation</b>, mild small striatocapsular (Score: <b>0.93</b>)<br/> [3] <b>Anterior circulation</b>, mild large vascular (Score: <b>0.91</b>)<br/> [4] <b>Anterior circulation</b>, mild large vascular (Score: <b>0.91</b>)<br/> [5] <b>Anterior circulation</b>, strong small (Score: <b>0.90</b>)</p> |
|----------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## 5 Discussion and Conclusion

By reformulating report generation as an image-domain retrieval and text-domain augmentation task, PIRTA overcomes key limitations of direct image–text alignment for volumetric MRI, reducing optimization complexity in data-constrained settings.

**Clinical implications of retrieval-grounded factuality.** Ischemic-territory information directly informs reperfusion triage in acute stroke care, where misattribution of vascular territory can shift treatment decisions and patient outcomes. PIRTA delivers consistent territory accuracy across the internal cohort and two external cohorts (BRMH, ISLES) without recalibration, with the explicit cosine-similarity score acting as a per-case confidence cue that lets unreliable retrievals be flagged for radiologist review.

**Territory accuracy as a focused factuality surrogate.** N-gram metrics such as BLEU and ROUGE miss clinically important factual errors in radiology reports, motivating ischemic-territory accuracy as a clinically grounded surrogate. We emphasize that territory accuracy is a *focused* factuality measure rather than a complete report-quality metric: it captures whether the most treatment-relevant attribute is preserved through generation but does not directly score lesion size, perfusion mismatch, hemorrhagic transformation, or report fluency. We adopt it as a high-signal, clinically actionable axis while leaving complementary metrics for future evaluation.

**Limitations and planned evaluation axes.** Factual accuracy degrades when rare or atypical infarct patterns are underrepresented in the retrieval database, yielding low cosine similarity and weak grounding. Beyond expanding the retrieval corpus, planned axes include (i) finer-grained ASPECTS-style scoring for per-region attribution, (ii) prospective clinical evaluation with radiologist-rated factuality on held-out reports, and (iii) hybrid multimodal retrieval that augments image embeddings with clinical metadata (NIHSS, presentation, PMH) to recover rare-pattern grounding.

In summary, PIRTA improves the factuality of 3D brain MRI report generation by grounding outputs in clinician-authored reports retrieved from similar cases, offering a scalable paradigm that achieves superior ischemic-territory accuracy without explicit cross-modal alignment.

**Acknowledgments.** This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2026-25479661) (K.S.C.); the SNUH Research Fund (No. 04-2024-0600; No. 04-2025-2060) (K.S.C.); the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI) grant funded by the Ministry of Health&Welfare (No. RS-2024-00439549) (K.S.C.); and the Bio&Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korean government (MSIT; RS-2025-02263045) (B.-H.K.).

**Disclosure of Interests.** Byung-Hoon Kim serves as a director of EverEx, which is unrelated to the matter of this study. The other authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Alfaro-Almagro, F., Jenkinson, M., Bangerter, N.K., Andersson, J.L., Griffanti, L., Douaud, G., Sotiropoulos, S.N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al.: Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *NeuroImage* **166**, 400–424 (2018). <https://doi.org/10.1016/j.neuroimage.2017.10.034>
2. Campbell, B.C., De Silva, D.A., Macleod, M., Coutts, S., Schwamm, L., Davis, S.M., Donnan, G.: Ischaemic stroke. *Nature Reviews Disease Primers* **5**(1), 70 (2019). <https://doi.org/10.1038/s41572-019-0118-8>
3. Çetinoglu, Y., Koska, I.O., Uluç, M., Gelal, M.: Detection and vascular territorial classification of stroke on diffusion-weighted MRI by deep learning. *European Journal of Radiology* **145**, 110050 (2021). <https://doi.org/10.1016/j.ejrad.2021.110050>
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021), <https://arxiv.org/abs/2010.11929>
5. Feigin, V., Brainin, M., Norrving, B., Martins, S., Sacco, R., Hacke, W., Fisher, M., Pandian, J., Lindsay, P.: World stroke organization (WSO): Global stroke fact sheet 2022. *International Journal of Stroke* **17**(1), 18–29 (2022). <https://doi.org/10.1177/17474930211065917>
6. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024), <https://arxiv.org/abs/2407.21783>
7. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.B.: Masked autoencoders are scalable vision learners. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 16000–16009 (2022). <https://doi.org/10.1109/CVPR52688.2022.01553>
8. Hernandez Petzsche, M.R., Rosa, E.L., Hanning, U., Wiest, R., Pinilla, W.E.V., Reyes, M., Meyer, M.I., Liew, S., Kofler, F., Ezhov, I., et al.: ISLES 2022: A multi-center magnetic resonance imaging stroke lesion segmentation dataset. *Scientific Data* **9**(1), 762 (2022). <https://doi.org/10.1038/s41597-022-01875-5>

9. Herzog, L., Murina, E., Dürr, O., Wegener, S., Sick, B.: Integrating uncertainty in deep neural networks for MRI based stroke analysis. *Medical Image Anal.* **65**, 101790 (2020). <https://doi.org/10.1016/j.media.2020.101790>
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Wang, S., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), <https://arxiv.org/abs/2106.09685>
11. Huo, F., Xu, W., Guo, J., Wang, H., Guo, S.: C2kd: Bridging the modality gap for cross-modal knowledge distillation. In: Computer Vision and Pattern Recognition. pp. 16006–16015 (June 2024). <https://doi.org/10.1109/CVPR52733.2024.01515>
12. Koska, I.O., Selver, A., Gelal, F., Uluç, M., Çetinoğlu, Y., Yurttutan, N., Serindere, M., Dicle, O.: Deep learning classification of ischemic stroke territory on diffusion-weighted MRI: Added value of augmenting the input with image transformations. *Journal of Imaging Informatics in Medicine* **38**, 1374–1387 (2025). <https://doi.org/10.1007/s10278-024-01277-6>
13. Lee, K.Y., Liu, C.C., Chen, D., Weng, C.L., Chiu, H.W., Chiang, C.H.: Automatic detection and vascular territory classification of hyperacute staged ischemic stroke on diffusion weighted image (DWI) using convolutional neural networks (CNN). *Scientific Reports* **13**(1), 404 (2023). <https://doi.org/10.1038/s41598-023-27621-4>
14. Lee, S., Kim, W.J., Chang, J., Ye, J.C.: LLM-CXR: Instruction-finetuned llm for CXR image understanding and generation. In: International Conference on Learning Representations (2024), <https://arxiv.org/abs/2305.11490>
15. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.t., Rocktäschel, T., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. In: Advances in Neural Information Processing Systems. vol. 33, pp. 9459–9474 (2020), <https://arxiv.org/abs/2005.11401>
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. In: Advances in Neural Information Processing Systems. vol. 36, pp. 28541–28564. Neural Information Processing Systems Foundation, Inc. (NeurIPS) (2023). <https://doi.org/10.52202/075280-1240>
17. Liu, C.F., Zhao, Y., Yedavalli, V., Leigh, R., Falcao, V., Wintermark, M., Warach, S.J., Albers, G.W., Davis, S.M., Grotta, J.C., et al.: Automatic comprehensive radiological reports for clinical acute stroke mris. *Communications Medicine* **3**(1), 95 (2023). <https://doi.org/10.1038/s43856-023-00327-4>
18. Ma, W., Li, S., Cai, L., Kang, J.: Learning modality knowledge alignment for cross-modality transfer. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 33777–33793 (2024). <https://doi.org/10.48550/arXiv.2406.18864>, <https://arxiv.org/abs/2406.18864>
19. Mouridsen, K., Thurner, P., Zaharchuk, G.: Artificial intelligence applications in stroke. *Stroke* **51**(8), 2573–2579 (2020). <https://doi.org/10.1161/STROKEAHA.119.027479>
20. Powers, W.J., Rabinstein, A.A., Ackerson, T., Adeoye, O.M., Bambakidis, N.C., Becker, K., Biller, J., Brown, M., Demaerschalk, B.M., Hoh, B., et al.: Guidelines for the early management of patients with acute ischemic stroke: 2019 update to the 2018 guidelines for the early management of acute ischemic stroke: A guideline for healthcare professionals from the american heart association/american stroke association. *Stroke* **50**(12), e344–e418 (2019). <https://doi.org/10.1161/str.0000000000000211>

21. Román, L., Menon, B., Blasco, J., Hernández-Pérez, M., Dávalos, A., Majoie, C., Campbell, B., Guillemin, F., Lingsma, H.F., Anxionnat, R., et al.: Imaging features and safety and efficacy of endovascular stroke treatment: a meta-analysis of individual patient-level data. *The Lancet Neurology* **17**(10), 895–904 (2018). [https://doi.org/10.1016/S1474-4422\(18\)30242-4](https://doi.org/10.1016/S1474-4422(18)30242-4)
22. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv (2025), <https://arxiv.org/abs/2507.05201>
23. Shalev-Shwartz, S., Ben-David, S.: *Understanding machine learning: From theory to algorithms*. Cambridge university press (2014)
24. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., Hou, L., Clark, K., Pfohl, S.R., Cole-Lewis, H., et al.: Toward expert-level medical question answering with large language models. *Nature Medicine* **31**(3), 943–950 (2025). <https://doi.org/10.1038/s41591-024-03423-7>
25. Taşci, B., Taşci, I.: Deep feature extraction based brain image classification model using preprocessed images: Pdrnet. *Biomedical Signal Processing and Control* **78**, 103948 (2022). <https://doi.org/10.1016/j.bspc.2022.103948>