Rethinking RAG: Mitigating Text-Driven Hallucinations via Feature-Level Retrieval Integration in Medical LVLMs

Yujoong Kim[1], Soopil Kim[2], Sion An[1], and Sang Hyun Park[3]*

[1] Graduate School of Artificial Intelligence, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea
{kuj1026,sanghyunpark}@postech.ac.kr
[2] AX Research Group for Semiconductors, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Republic of Korea
[3] Department of Computer Science & Engineering, Pohang University of Science and Technology (POSTECH), Pohang, Republic of Korea

Abstract. Medical Large Vision-Language Models (Med-LVLMs) achieve strong capability but remain prone to hallucinations that pose serious diagnostic risks. Retrieval-Augmented Generation (RAG) has been explored to mitigate this issue. However, existing approaches append retrieved evidence as text to the language input, which shifts the model's focus toward textual cues instead of visual reasoning and increases the risk of copying retrieved reports. To address this, we propose a RAG framework that leverages retrieved image–text pairs to enhance the visual representation of a query image through modified cross-attention. We incorporate a novel training objective into this module to prevent over-reliance on a few references and promote semantically faithful alignment. In addition, we observe that commonly used CLIP-based retrieval often assigns high similarity to negated and non-negated sentences with contradictory clinical meanings, resulting in mismatched references and potential hallucinations. To mitigate this, we design a negation-aware retrieval strategy that filters clinically inconsistent cases prior to feature fusion. Experiments on three benchmarks in radiology and ophthalmology demonstrate consistent improvements over the base model and text-prompt RAG baselines, by reducing language shortcuts and reinforcing visual evidence utilization. Source code will be released upon publication.

Keywords: Medical Vision-Language Models · Retrieval-Augmented Generation · Cross-Attention · Visual Feature Enhancement · Hallucination Mitigation.

1 Introduction

Medical Large Vision-Language Models (Med-LVLMs) have attracted growing attention for generating clinical reports and answering diagnostic questions from

* Corresponding author.

medical images [1, 2]. Despite progress, these models frequently produce hallucinations, generating statements that appear plausible yet contradict the visual evidence in the input image [3].

Retrieval-Augmented Generation (RAG) has therefore been adopted to mitigate hallucination by grounding model outputs in external evidence [4], and recent efforts have explored three main directions in this context. The first focuses on improving retrieval quality by training fact-aware multimodal retrievers [5]. The second direction focuses on aligning retrieved contexts with model behavior [6–8]. To mitigate misleading references, these methods apply preference fine-tuning to calibrate the model's reliance on retrieved evidence. More recent approaches further incorporate clinical relevance directly into the preference signal. The third integrates explicit chain-of-thought reasoning into the RAG pipeline to enhance the trustworthiness of model decisions [9]. Despite these advances, existing approaches share a fundamental structural limitation. Retrieved evidence is incorporated solely as additional textual input to the language model. This exacerbates the language bias in VLMs [10], posing a risk of directly copying the reference text rather than grounding in visual evidence [7].

In this paper, we propose a novel RAG framework that utilizes retrieved image–text pairs to enhance visual feature of the query image, preventing the model from copying retrieved text while enabling more effective use of retrieval information. To achieve this, we design a cross-attention module that fuses features from the retrieved image–text pairs with those of the query image, allowing structured and grounded information integration at the feature level. As attention in VLMs tends to focus on only a few retrieved references without explicit regularization, we further introduce a grounding loss to prevent the cross-attention from collapsing onto a small subset of visual tokens. Our grounding loss encourages a more balanced and similarity-aware attention distribution, promoting grounded and reliable evidence integration.

In addition, we identify a key limitation in commonly used CLIP-based retrieval. Although widely adopted for selecting top-k reports, CLIP similarity is insensitive to subtle clinical differences and may assign high similarity to contradictory negated and non-negated sentences. To address this, we propose a negation-aware retrieval strategy that filters clinically inconsistent cases and improves the factual reliability of generated outputs. In summary, the main contributions of this paper are as follows:

1. We introduce a novel RAG-based visual reasoning framework that incorporates clinically similar patient cases directly into visual feature encoding.
2. We introduce an attention grounding loss to prevent attention collapse and promote balanced, clinically consistent evidence integration.
3. We propose a negation-aware retrieval strategy that filters clinically inconsistent cases and improves factual reliability.
4. Our framework achieves state-of-the-art performance across multiple benchmarks, outperforming conventional text-prompt RAG methods.

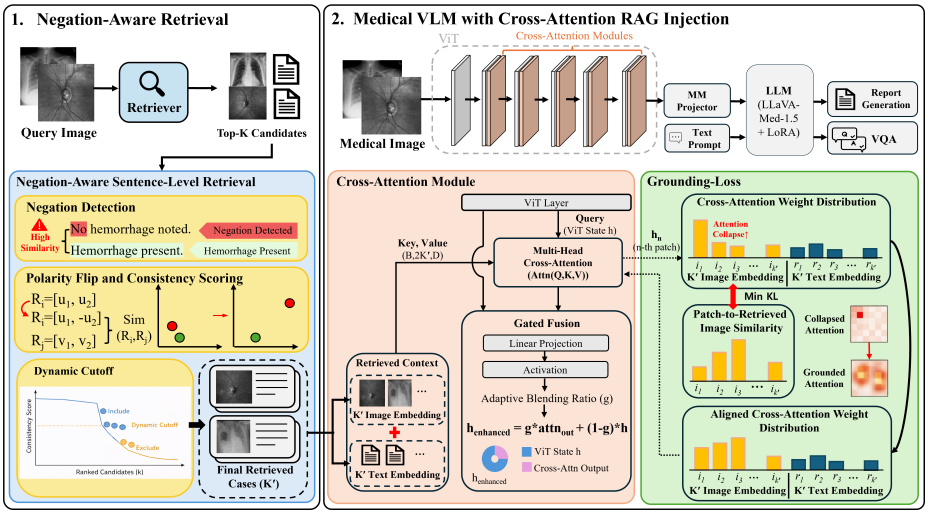

Fig. 1. Overview of the proposed framework. (i) A negation-aware retriever selects clinically consistent cases from the database. (ii) Retrieved image–report pairs are fused into the upper layers of the vision encoder via gated cross-attention, enriching the query's visual representation. (iii) A grounding loss prevents attention from collapsing onto a few tokens, ensuring attention aligns with visual relevance.

2 Method

Overview. We consider the problem of retrieval-augmented medical visual question answering and report generation. Given a query image I_q, a retriever selects K' reference cases $\{(\mathbf{I}_k, \mathbf{R}_k)\}_{k=1}^{K'}$ from a database, where $\mathbf{I}_k$ and $\mathbf{R}_k$ denote the retrieved image and report respectively. The model f_θ takes the query and retrieved references as input to produce an output $\hat{y} = f_\theta\Big(I_q, \{(\mathbf{I}_k, \mathbf{R}_k)\}_{k=1}^{K'}\Big)$. Conventional methods encode the retrieved reports as additional text tokens, i.e., $f_\theta(I_q, [\mathbf{R}_1; \mathbf{R}_2; \ldots; \mathbf{R}_{K'}])$ [7, 10]. Our framework instead fuses $(\mathbf{I}_k, \mathbf{R}_k)$ directly into the visual encoding of I_q, so that retrieved evidence enriches the visual representation before any text is generated. An overview is shown in Fig. 1.

2.1 Negation-Aware Sentence-Level Retrieval

Most existing medical RAG methods rely on global CLIP [11] image–text embeddings for retrieval. However, CLIP similarity is largely driven by writing style and shared keywords, rather than fine-grained clinical semantics. In particular, CLIP struggles to distinguish negated findings from affirmative ones [12]. This limitation can result in mismatched references and potential hallucinations. To address this issue, we propose a negation-aware sentence-level retrieval strategy.

Each report $\mathbf{R}_i$ is decomposed into individual sentences and each sentence is independently encoded with CLIP, to yield sentence embeddings $\mathbf{u}_j^{(i)}$, $j =$

$1, \ldots, n_i$. Each embedding is classified as negated or affirmative using NegEx [13] combined with a biomedical NLP model, scispaCy [14]. For each negated sentence, the sign of its CLIP embedding is inverted: $\mathbf{u}_{\text{neg}} = -\mathbf{u}$. Under cosine similarity, this inversion ensures that contradictory statements are no longer treated as similar.

The inter-report similarity $\hat{S}$ is computed via symmetric bidirectional matching to account for length variability:

$$\hat{S}(\mathbf{R}_i, \mathbf{R}_j) = \frac{1}{2}\left[\frac{1}{n_i}\sum_{\mathbf{u}\in\mathbf{R}_i}\max_{\mathbf{v}\in\mathbf{R}_j}\cos(\mathbf{u},\mathbf{v}) + \frac{1}{n_j}\sum_{\mathbf{v}\in\mathbf{R}_j}\max_{\mathbf{u}\in\mathbf{R}_i}\cos(\mathbf{v},\mathbf{u})\right], \quad (1)$$

where $\cos(\cdot,\cdot)$ denotes cosine similarity and $\mathbf{u}$, $\mathbf{v}$ denote sentence embeddings from $\mathbf{R}_i$ and $\mathbf{R}_j$, respectively. Given an initial set of top-K candidates retrieved by CLIP image similarity, the consistency score $\hat{C}_i$ of each retrieved report is defined as its mean similarity to the remaining candidates:

$$\hat{C}_i = \frac{1}{K-1}\sum_{j\neq i}\hat{S}(\mathbf{R}_i, \mathbf{R}_j). \quad (2)$$

Let $\hat{s}_i$ denote the CLIP cosine similarity between the query image and the i-th retrieved image. The final ranking score is $\phi_i = \alpha_{\text{rep}} \cdot \hat{C}_i + \alpha_{\text{agr}} \cdot \hat{s}_i \cdot \hat{C}_i$, where α_{rep} weights standalone report consistency and α_{agr} weights its interaction with image similarity. This ensures high ranking only when a case is both report-consistent and visually similar to the query. Retrieved cases are sorted by ϕ_i and discarded at the point where the score drops by more than a gap threshold δ, yielding K' clinically coherent cases for the fusion step.

2.2 Cross-Attention RAG Injection in the Vision Encoder

In prior text-prompt RAG methods, retrieved evidence is appended as additional text tokens to the language model input. As discussed in Sec. 1, this design exacerbates the inherent language bias of VLMs and increases the risk of directly copying retrieved text [7, 10]. To avoid this shortcut, we inject retrieved evidence exclusively into the visual encoding process, ensuring that references can only influence the model through enriched visual features.

We insert gated cross-attention modules into the upper layers of the vision encoder [15]. Each retrieved image $\mathbf{I}_k$ and report $\mathbf{R}_k$ are encoded with the CLIP image and text encoders, respectively, yielding image embeddings $\mathbf{i}_k$ and report embeddings $\mathbf{r}_k$. The retrieved context is formed by concatenating all K' pairs: $\mathbf{C} = [\mathbf{i}_1; \mathbf{r}_1; \ldots; \mathbf{i}_{K'}; \mathbf{r}_{K'}] \in \mathbb{R}^{B\times 2K'\times D}$. At each inserted layer l, the ViT hidden state $\mathbf{h}^{(l)} \in \mathbb{R}^{B\times N_p\times D}$ of the query image I_q is used as the attention query, while $\mathbf{C}$ serves as the key and value:

$$\mathbf{A}^{(l)} = \text{softmax}\left(\frac{\mathbf{Q}^{(l)}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}, \quad \mathbf{Q}^{(l)} = \mathbf{h}^{(l)}W_Q,\ \mathbf{K} = \mathbf{C}W_K,\ \mathbf{V} = \mathbf{C}W_V. \quad (3)$$

Since retrieval quality varies across queries, a fixed-weight fusion risks over-incorporating noisy retrievals or under-utilizing relevant ones. Inspired by [15], we adopt a learnable sigmoid gate to adaptively control the contribution of retrieved evidence:

$$\mathbf{g} = \sigma\Big(W_g\big[\mathbf{h}^{(l)}; \mathbf{A}^{(l)}\big] + b_g\Big), \quad \mathbf{h}^{(l)}_{\text{enh}} = \mathbf{g} \odot \mathbf{A}^{(l)} + (1 - \mathbf{g}) \odot \mathbf{h}^{(l)}. \tag{4}$$

The gate bias is initialized negative to preserve pre-trained representations early in training. During training, the cross-attention parameters, MM Projector, and LoRA adapters are optimized to adapt the model to the enriched visual representation.

2.3 Correspondence-Based Grounding Loss

When performing cross-attention over retrieved cases, the model is expected to attend in proportion to their clinical relevance to the query. Without additional supervision, however, attention weights tend to collapse onto a narrow subset of retrieved cases that most rapidly reduce the autoregressive loss [16]. This prevents the model from broadly referencing all retrieved cases, introducing retrieval bias.

To prevent this, we supervise the image-portion of cross-attention with a target derived from patch-level visual similarity. For each patch n, we compute cosine similarity to each retrieved image embedding $\mathbf{i}_k$ and form a soft target:

$$p_{\text{corr}}(k \mid n) = \text{softmax}\left(\frac{\cos(\mathbf{h}^{(l)}_n,\, \mathbf{i}_k)}{\tau}\right), \quad p_{\text{smooth}} = (1 - \varepsilon)\, p_{\text{corr}} + \frac{\varepsilon}{K'}, \tag{5}$$

where τ controls sharpness and ε is a label-smoothing coefficient. Intuitively, this target aligns the attention distribution with actual visual relevance via KL divergence. Since attention is computed over the full $2K'$ context, we extract and renormalize the image-portion weights so that $p^{\text{img}}_{\text{attn}}$ and p_{smooth} share the same K'-dimensional simplex and minimize their KL divergence:

$$\mathcal{L}_{\text{ground}} = \sum_n \text{KL}\Big(p_{\text{smooth}}(\cdot \mid n) \parallel p^{\text{img}}_{\text{attn}}(\cdot \mid n)\Big). \tag{6}$$

The total loss is $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda\, \mathcal{L}_{\text{ground}}$, where λ balances language modeling and attention regularization.

3 Experiments

Experimental Setup. We evaluate on three benchmarks: **IU X-ray** [17], **MIMIC-CXR** [18], and **Harvard-FairVLMed** [19], following the data splits and preprocessing of [6, 7]. VQA benchmarks are constructed by generating yes/no question-answer pairs from medical reports using GPT-4 [20]. VQA is evaluated by Accuracy, F1, and AUC; report generation by ROUGE-L [21], and

Table 1. VQA performance across three benchmarks. **Bold**: best; underline: second best.

	IU X-ray			MIMIC-CXR			Harvard-FairVLMed		
Model	Acc	F1	AUC	Acc	F1	AUC	Acc	F1	AUC
LLaVA-Med-1.5	75.47	64.04	67.46	75.79	80.49	68.84	63.03	74.11	63.05
+ Greedy	76.88	65.59	68.74	78.32	86.75	71.13	82.54	85.98	70.09
+ Beam Search	76.91	66.06	68.77	81.56	86.36	73.79	80.93	88.08	68.94
+ DoLa	78.00	66.75	72.19	81.35	85.73	72.73	76.87	85.53	67.10
+ OPERA	70.59	61.54	63.22	69.34	76.66	62.46	71.41	81.37	65.59
+ VCD	68.99	54.35	61.08	70.89	75.57	64.61	65.88	77.20	64.16
+ MedDr	83.33	67.80	77.15	55.16	56.18	58.47	70.17	80.72	64.15
+ FactMM-RAG	84.51	68.51	77.07	77.58	81.86	70.09	83.67	87.21	72.20
+ RULE	87.84	78.00	85.78	83.92	87.49	83.44	87.12	92.89	77.08
+ MMed-RAG	89.54	80.72	87.13	83.57	**88.49**	85.08	87.94	92.78	80.81
+ **Ours**	**90.83**	**82.72**	**91.71**	**85.36**	87.05	**86.74**	**93.09**	**96.00**	**84.78**

METEOR [22]. The backbone is LLaVA-Med-1.5 [2] (Mistral-7B + CLIP ViT-L/14).

The retriever is based on OpenCLIP [23] ResNet-50 pretrained on CC12M, fine-tuned for 360 epochs with a learning rate of 1×10^{-4} and batch size of 512. Retrieval uses $K = 15$ candidates with gap threshold δ=0.05 and equal weighting (α_{rep}=α_{agr}=0.5). The cross-attention module consists of 6 layers with 8 heads. LoRA is configured with r=32, α=64 [24]. The grounding loss weight is λ=2×10^{-3} with temperature τ=0.1.

We adopt a two-stage training protocol to avoid gradient competition between randomly initialized cross-attention and pre-trained language components. In Stage 1, only the cross-attention parameters are updated, establishing visual retrieval integration before any language-side adaptation. In Stage 2, the cross-attention is frozen and the MM Projector and LoRA adapters are fine-tuned to adapt to the enriched visual representation.

Main Results. We compare against decoding-based methods such as Greedy, Beam Search, DoLa [25], OPERA [26] and VCD [27], and medical RAG systems such as MedDr [28], FactMM-RAG [5], RULE [6], MMed-RAG [7]. As shown in Table 1, our method consistently achieves the highest Accuracy and AUC across all three datasets for VQA. Decoding-based methods yield modest improvements but remain below RAG-based approaches, confirming that retrieval provides complementary evidence. Among RAG baselines, MMed-RAG achieves the strongest results overall, yet our method surpasses it across all benchmarks. The improvement is particularly pronounced on Harvard-FairVLMed, where ophthalmologic findings are highly localized and benefit most from fine-grained visual verification. This suggests that fusing evidence at the vision level captures subtle diagnostic cues that text-prompt injection alone cannot reliably convey to the decoder. For report generation (Table 2), our approach achieves the highest

Table 2. Report generation performance across three benchmarks. **Bold**: best; underline: second best.

	IU X-ray		MIMIC-CXR		Harvard-FairVLMed	
Model	R-L	MTR	R-L	MTR	R-L	MTR
LLaVA-Med-1.5	12.26	8.21	13.05	11.16	11.36	10.75
+ Greedy	15.38	12.69	14.26	14.19	11.49	13.77
+ Beam Search	16.21	13.17	14.74	14.43	12.62	14.50
+ DoLa	15.82	12.72	14.89	14.81	12.51	14.51
+ OPERA	14.70	12.01	12.52	13.72	11.47	13.63
+ VCD	14.14	11.59	12.30	13.38	11.38	13.89
+ MedDr	16.45	13.50	15.72	16.77	13.72	15.40
+ FactMM-RAG	18.05	15.92	15.84	16.82	14.17	15.31
+ RULE	23.16	27.99	15.96	17.42	14.93	17.74
+ MMed-RAG	25.59	32.43	12.34	20.47	16.59	19.85
+ **Ours**	**34.96**	**35.34**	**21.01**	**21.57**	**17.75**	**20.75**

Table 3. Ablation results on MIMIC-CXR. **Bold**: full model.

Model	Acc	F1	AUC
Full model	**85.36**	**87.05**	**86.74**
w/o retrieved report (img only)	83.56	85.12	85.08
w/o gating	83.17	84.99	84.96
w/o negation-aware (fixed top-15)	80.00	81.62	82.08
w/o grounding loss	79.63	81.01	81.35

ROUGE-L and METEOR scores across all three benchmarks, with especially large margins on IU X-ray and MIMIC-CXR. Notably, text-prompt RAG baselines show unstable performance across datasets in this task, whereas our method maintains consistent gains. This indicates that conditioning generation on visually enriched representations produces reports that better reflect actual image content, rather than being biased by injected textual priors.

Ablation Study. As shown in Table 3, we conducted a detailed ablation analysis on each component of the proposed framework. *w/o retrieved text* removes the top-K' retrieved report embeddings from the cross-attention context; *w/o gating* removes the sigmoid gate and applies a fixed blend of retrieved features; *w/o negation-aware* replaces the negation-aware retriever with initial top-K; *w/o grounding loss* sets λ=0, training the cross-attention with the language modeling objective alone.

All four variants degrade performance. Among them, the grounding loss and the negation-aware retriever show the largest impact, with accuracy drops of 5.73pp and 5.36pp respectively. This is consistent with the attention analysis in Fig. 2 (right), where removing the grounding loss leads to attention collapse onto a narrow subset of retrieved cases. Removing retrieved text or the gating

Table 4. Attention ratio (Pre-Vision / Vision / Post-Vision) for VQA across three benchmarks. Values are colored red (increase) or blue (decrease) relative to the base model.

	IU X-ray			MIMIC-CXR			Harvard-FairVLMed		
Model	Pre	Vis	Post	Pre	Vis	Post	Pre	Vis	Post
LLaVA-Med-1.5	0.4994	0.0511	0.1765	0.4879	0.0606	0.1906	0.4923	0.0585	0.1927
+ Text RAG	0.3564	0.0136	0.4165	0.3533	0.0122	0.4101	0.3700	0.0145	0.4200
+ **Ours**	0.3675	0.1965	0.1656	0.3223	0.2708	0.1327	0.4246	0.1826	0.1559

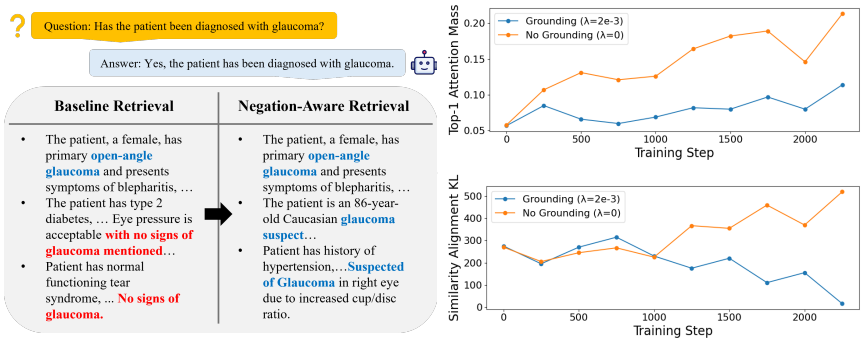

Fig. 2. Left: Negation-aware retrieval excludes diagnostically contradictory cases (e.g., *"no signs of glaucoma"*) that baseline retrieval incorrectly includes. **Right**: Without the grounding loss (λ=0), top-1 attention collapses and KL divergence grows; the grounding loss (λ=2×10^{-3}) maintains visually-aligned attention.

mechanism yields moderate but consistent drops, indicating that both diagnostic context from reports and adaptive fusion contribute to the overall performance.

To further verify that performance gains reflect genuine visual grounding rather than language shortcuts, we decompose the language model's attention into three segments. Pre-Vision corresponds to system prompt tokens, Vision to projected ViT patch embeddings, and Post-Vision to question and retrieved text tokens. As shown in Table 4, text-prompt RAG suppresses the Vision ratio below the base model while inflating Post-Vision, confirming the shortcut behavior. Our method substantially raises the Vision ratio across all three datasets.

4 Conclusion

We presented a novel retrieval-augmented framework for Med-LVLMs that incorporates retrieved image and report pairs directly into the vision encoder. This addresses the fundamental structural limitation of conventional text-prompt RAG. By ensuring that retrieved evidence influences the model only through enriched visual features, our framework prevents bypassing visual reasoning by copying retrieved text. A correspondence-based grounding loss further regularizes cross-

attention to prevent collapse onto a narrow subset of references, promoting diverse and visually grounded evidence utilization. In addition, we introduce a negation-aware retrieval strategy that filters clinically contradictory cases before fusion, improving the factual reliability of retrieved evidence. Experiments across three benchmarks in radiology and ophthalmology show consistent improvements over both the base model and existing RAG baselines, while attention analysis confirms that our framework substantially restores the model's reliance on visual evidence. Extending the evaluation with clinical-efficacy metrics, open-ended VQA, and more recent Med-LVLM backbones remains a direction for future work.

Acknowledgments. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00439264, Development of High-Performance Machine Unlearning Technologies for Privacy Protection), and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-00516124).

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

References

1. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
2. Li, C., et al.: LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In: NeurIPS (2023)
3. Xia, P., et al.: CARES: A comprehensive benchmark of trustworthiness in medical vision language models. In: NeurIPS Datasets and Benchmarks Track (2024)
4. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: NeurIPS, vol. 33, pp. 9459–9474 (2020)
5. Sun, L., Zhao, J., Han, W., Xiong, C.: Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In: NAACL (2025)
6. Xia, P., et al.: RULE: Reliable multimodal RAG for factuality in medical vision language models. In: EMNLP, pp. 1081–1093 (2024)
7. Xia, P., et al.: MMed-RAG: Versatile multimodal RAG system for medical vision language models. In: ICLR (2025)
8. Zhu, K., et al.: MMedPO: Aligning medical vision-language models with clinical-aware multimodal preference optimization. arXiv:2412.06141 (2024)
9. Dai, P., et al.: GoCa: Trustworthy multi-modal RAG with explicit thinking distillation for reliable decision-making in Med-LVLMs. In: MICCAI (2025)
10. Zhao, H., et al.: Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance. In: EMNLP (2025)
11. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)
12. Zhang, S., et al.: Vision-language models do not understand negation. arXiv:2501.09425 (2025)

13. Chapman, W.W., et al.: A simple algorithm for identifying negated findings and diseases in discharge summaries. J. Biomed. Informatics **34**(5), 301–310 (2001)
14. Neumann, M., et al.: ScispaCy: Fast and robust models for biomedical natural language processing. In: BioNLP Workshop, pp. 319–327 (2019)
15. Alayrac, J.B., et al.: Flamingo: A visual language model for few-shot learning. In: NeurIPS, pp. 23716–23736 (2022)
16. Xiao, G., Tian, Y., Chen, B., Han, S., Lewis, M.: Efficient streaming language models with attention sinks. In: ICLR (2024)
17. Demner-Fushman, D., et al.: Preparing a collection of radiology examinations for distribution and retrieval. JAMIA 23(2), 304–310 (2016)
18. Johnson, A.E., et al.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv:1901.07042 (2019)
19. Luo, Y., et al.: FairCLIP: Harnessing fairness in vision-language learning. In: CVPR, pp. 12289–12301 (2024)
20. OpenAI: GPT-4 technical report. arXiv:2303.08774 (2023)
21. Lin, C.Y.: ROUGE: A package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81 (2004)
22. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: ACL Workshop, pp. 65–72 (2005)
23. Cherti, M., et al.: Reproducible scaling laws for contrastive language-image learning. In: CVPR, pp. 22,084–22,094 (2023)
24. Hu, E.J., et al.: LoRA: Low-rank adaptation of large language models. In: ICLR (2022)
25. Chuang, Y.S., et al.: DoLa: Decoding by contrasting layers improves factuality in large language models. arXiv:2309.03883 (2023)
26. Huang, Q., et al.: OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In: CVPR, pp. 13418–13427 (2024)
27. Leng, S., et al.: Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: CVPR, pp. 13,872–13,882 (2024)
28. He, S., et al.: Meddr: Diagnosis-guided bootstrapping for large-scale medical vision-language learning. arXiv preprint arXiv:2404.15127 (2024)