Efficient and Faithful Retrieval-Augmented Medical VQA via Intention-Guided On-Policy Self-Distillation

Linhao Li, Yiwen Ye, Yong Xia*

National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China
lilinhao@mail.nwpu.edu.cn, ywye@mail.nwpu.edu.cn, yxia@nwpu.edu.cn

Abstract. Retrieval-augmented generation (RAG) is increasingly used to ground medical vision-language models (VLMs) in clinical visual question answering (VQA). Yet real clinical corpora such as chest X-ray (CXR) report collections are highly templated, so near-identical reports differ only in a few decisive phrases and retrieval becomes noisy. We argue that optimizing the knowledge base matters as much as optimizing the model, and we address both with *Intent-RAG*, an intention-guided on-policy self-distillation framework for efficient and faithful retrieval-augmented medical VQA. Its fine-grained corpus curation decomposes reports into sentence-level chunks, applies semantic deduplication to merge templated near-duplicates, selects cluster-centroid representatives to build a compact high-coverage core corpus, and aligns retrieval to the statement form of the question rather than whole-image similarity. Its intention-guided on-policy self-distillation (OPSD) then uses the same model as student and teacher and controls the visibility of the retrieved reports and the ground-truth answer across differently-privileged teachers, applying dense token-level supervision to distinct perception, reasoning, and answer segments without any architectural change. On IU-Xray and MIMIC-CXR we introduce the Retrieval Answer Leakage Rate as a diagnostic for visual grounding and report Retrieval Robustness under corrupted retrieval. Overall, Intent-RAG surpasses the baseline while cutting the knowledge base by more than 80% at no cost to accuracy.

Keywords: Retrieval-augmented generation · Medical visual question answering · Efficient corpus curation · On-policy self-distillation.

1 Introduction

Large vision-language models (VLMs) show great promise in medical applications [18, 8, 13, 9], and retrieval-augmented generation [7] (RAG) has become an effective evidence-grounding mechanism that is widely adopted in high-stakes

* Corresponding author

clinical settings, where a model retrieves similar historical cases from a clinical corpus to support its reasoning [23, 24].

Using real clinical case data such as MIMIC-CXR reports as the RAG corpus is often ineffective. Radiologists write imaging reports from templates, so reports for different images are highly similar and differ only in a few key phrases, and this redundancy introduces noise that drowns out the useful signal and makes it hard to retrieve genuinely informative reports. To attack this redundancy directly, we propose *fine-grained corpus curation*: we decompose each report into sentence-level chunks, apply semantic deduplication to merge templated near-duplicates and drop short fragments, and select cluster-centroid representatives to build a compact yet high-coverage core corpus. On MIMIC-CXR a core corpus of under 20% of the full one matches or even exceeds full-corpus retrieval. This is our first contribution, treating the corpus as a first-class optimization target rather than a static given.

Optimizing the corpus alone does not cure a second problem: when using retrieved evidence the model takes shortcuts, over-trusting the retrieved text so that attention drifts to text tokens, the image is ignored, and an otherwise-correct answer can be flipped [20, 11, 21, 10]. To counter this, existing work mainly relies on preference fine-tuning with DPO [16] to align the model on *whether* to trust retrieval, where RULE [23] and MMed-RAG [24] calibrate cross-modal reliance and ProtoMedAgent [11] shows that clinical RAG can trigger *retrieval sycophancy*. Yet this sparse sequence-level signal governs only *whether* to use retrieval and weakly teaches *how* to look at the image. More recently, on-policy self-distillation [4, 17, 1] has emerged as a powerful way to steer *how* a model learns, injecting a targeted prior through a privileged teacher and outperforming sparse RL or SFT signals [17, 1]: Vision-OPD [25] injects a regional-to-global perception prior, ViGOS [20] a perception-reasoning decoupling prior, ViCuR [19] recoverable visual cues that close the train-test privilege gap, and SD-Search [12] a hindsight prior for search-augmented reasoning.

These priors all elicit a capability the model already has but does not spontaneously use, and the shortcut above is exactly such a case: the model can already read the image, yet ordinary fine-tuning lets it lean on retrieved text instead. We therefore propose *Intent-RAG*, **intent**ion-guided OPSD for efficient and faithful **RAG**. On the corpus side, we distill the templated report base into a compact core corpus and retrieve with the statement form of the question rather than whole-image similarity. On the model side, intention-guided OPSD injects the prior through three differently-privileged teachers that supervise the perception, reasoning, and answer segments with the same model as student and teacher, withholding retrieval from the perception teacher to keep perception image-grounded and exposing it only to the reasoning teacher so the student weighs retrieval critically, without any architectural change. On MIMIC-CXR and IU-Xray closed-form VQA, Intent-RAG surpasses the baseline while retrieving from under 20% of the full corpus, cutting the knowledge base by more than 80% at no cost to accuracy. We summarize our main contributions as follows:

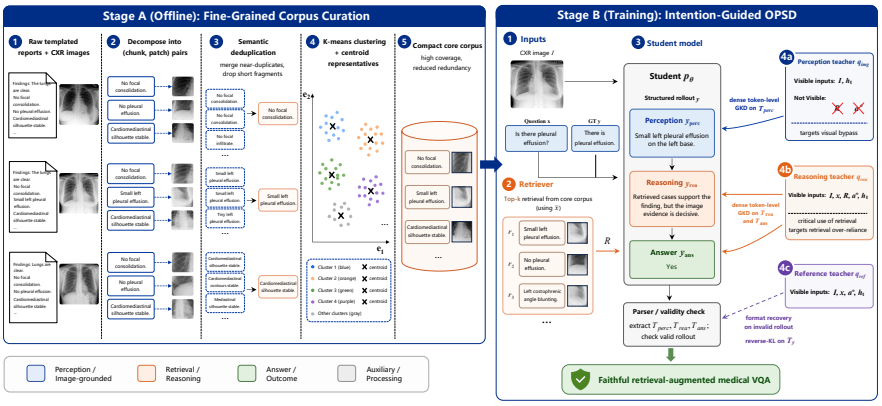

Fig. 1. Overview of Intent-RAG. **Stage A (offline):** raw templated reports are decomposed into (chunk, patch) pairs, semantically deduplicated, and clustered into cluster-centroid representatives that form a compact core corpus. **Stage B (training):** the statement-form question retrieves from the core corpus, the student samples a three-segment rollout of perception, reasoning, and answer, and three differently-privileged teachers supervise distinct spans, targeting visual bypass and retrieval over-reliance.

- **Fine-grained corpus curation.** We treat the corpus as a first-class optimization target, distilling templated reports into a compact core corpus of chunk-patch pairs through decomposition, deduplication, and clustering-based coreset selection.
- **Intention-guided OPSD.** Three differently-privileged teachers supervise the perception, reasoning, and answer segments, targeting visual bypass and retrieval over-reliance without any architectural change.
- **Evaluation and diagnostics.** On IU-Xray and MIMIC-CXR closed-form VQA, we introduce the Retrieval Answer Leakage Rate for visual grounding and report Retrieval Robustness under corrupted retrieval.

2 Method

2.1 Overview

Given a chest X-ray (CXR) image I, a question x, and a knowledge base $\mathcal{C}$ of historical reports, a retriever returns the top-k evidence $R = \{r_1, \ldots, r_k\}$, and the student p_θ generates an answer y from (I, x, R). We organize the output as three segments $y = (y_{\text{perc}}, y_{\text{rea}}, y_{\text{ans}})$: perception describes image-grounded findings from the image alone, reasoning combines the retrieved evidence with that perception, and answer gives the closed-form Yes/No conclusion. The objective is to predict y_{ans} correctly while mitigating two shortcuts, ignoring the image, which we term *visual bypass*, and blindly following retrieval, which we term *retrieval over-reliance*.

Intent-RAG has two stages. Offline, fine-grained corpus curation distills the templated report base into a compact core corpus. At training, intention-guided OPSD supervises the student with segment-wise privileged teachers so it grounds perception in the image and uses retrieval critically. At inference, the statement-form question retrieves from the core corpus and the student produces a three-segment answer. Figure 1 shows the two stages and their data flow.

2.2 Fine-Grained Corpus Curation

We construct a compact, high-information *core corpus* through fine-grained decomposition, semantic deduplication, and clustering-based coreset selection.

Fine-grained decomposition, deduplication, and alignment. Each report is split into chunks. To counter the redundancy induced by template-based reporting, we apply *semantic deduplication* [6], merging near-duplicate chunks and discarding overly short fragments so that only semantically unique chunks remain. For each retained chunk we localize an evidence patch on its image, giving a "chunk-patch" aligned unit. Retrieval combines a text channel and an optional patch channel, following the contrastive alignment paradigm of CLIP [15],

$$\mathrm{score}(q, s) = \mathrm{sim}(q, s) + \lambda\, \mathrm{sim}\big(I_q, \mathrm{patch}(s)\big), \tag{1}$$

where q is the retrieval query and the aligned patch contributes visual grounding with a weight λ that may be set to 0. At inference the query is the statement-form question,$q = \tilde{x}$, so text similarity between $\tilde{x}$ and the chunk text s dominates the ranking and aligns with the queried finding, avoiding the off-topic retrieval that whole-image similarity suffers from under modality mismatch. Both the chunk text and its aligned patch are passed to the model as a multimodal reference.

2.3 Intention-Guided OPSD

Directly porting text-only OPSD to medical RAG amplifies the two shortcuts of Section 1. Our remedy uses the *same* model as both student and teacher and controls the visibility of the retrieved reports R and the ground-truth answer a^* across teachers that each supervise a different span: the perception teacher sees only the image to combat *visual bypass*, while R is exposed only to the reasoning teacher so the model first establishes image evidence and then uses retrieval *critically* to combat *retrieval over-reliance*. Each teacher's usefulness is validated separately by ablation.

Structured rollout and parsing. The student generates $y = (y_{\mathrm{perc}}, y_{\mathrm{rea}}, y_{\mathrm{ans}})$ as three explicitly delimited segments, akin to chain-of-thought reasoning [22], sampled on-policy $y \sim p_\theta(\cdot \mid I, x, R, \pi_{\mathrm{out}})$ with output-format prompt π_{out} and never seeing the ground truth a^*. A parser extracts mutually exclusive token masks $\mathcal{T}_{\mathrm{perc}}$,$\mathcal{T}_{\mathrm{rea}}$,$\mathcal{T}_{\mathrm{ans}}$, retaining delimiter tokens but excluding them from the segment loss, and a validity flag $m_{\mathrm{inv}}(y)$: a rollout is *valid* if all three segments are present,

the perception and reasoning spans are non-empty, and the answer is parseable, where a wrong-but-parseable answer still counts as valid. For invalid rollouts the segment masks are empty and a reference teacher restores the format over the whole span $\mathcal{T}_y$.

Perception teacher. $q_{\mathrm{img},t}(\cdot) = p_{\bar{\theta}}(\cdot \mid I, h_t)$ supervises $\mathcal{T}_{\mathrm{perc}}$. It observes only the CXR image and the student prefix h_t but not R or a^*, forcing the perception span to be image-driven:

$$\mathcal{L}_{\mathrm{perc}} = \mathbb{E}_{D,p_\theta}\left[m_{\mathrm{inv}}(y) \sum_{t \in \mathcal{T}_{\mathrm{perc}}} D_\beta^{\mathrm{JSD}}(q_{\mathrm{img},t} \,\|\, p_{\theta,t})\right]. \tag{2}$$

Reasoning teacher. $q_{\mathrm{rea},t}(\cdot) = p_{\bar{\theta}}(\cdot \mid I, x, R, a^*, h_t)$ supervises $\mathcal{T}_{\mathrm{rea}} \cup \mathcal{T}_{\mathrm{ans}}$. With access to retrieval and ground truth, it teaches the student to evaluate retrieval critically against its own perception:

$$\mathcal{L}_{\mathrm{rea}} = \mathbb{E}_{D,p_\theta}\left[m_{\mathrm{inv}}(y) \sum_{t \in \mathcal{T}_{\mathrm{rea}} \cup \mathcal{T}_{\mathrm{ans}}} D_\beta^{\mathrm{JSD}}(q_{\mathrm{rea},t} \,\|\, p_{\theta,t})\right]. \tag{3}$$

Reference teacher. $q_{\mathrm{ref},t}(\cdot) = p_{\bar{\theta}}(\cdot \mid I, x, a^*, h_t)$ recovers the output format on invalid rollouts via reverse-KL and is inactive on valid ones:

$$\mathcal{L}_{\mathrm{ref}} = \mathbb{E}_{D,p_\theta}\left[m_{\mathrm{inv}}(y) \sum_{t \in \mathcal{T}_y} D_{\mathrm{KL}}(p_{\theta,t} \,\|\, q_{\mathrm{ref},t})\right]. \tag{4}$$

Overall objective. The total loss is $\mathcal{L} = \lambda_{\mathrm{perc}}\mathcal{L}_{\mathrm{perc}} + \lambda_{\mathrm{rea}}\mathcal{L}_{\mathrm{rea}} + \lambda_{\mathrm{ref}}\mathcal{L}_{\mathrm{ref}}$, and all divergences use the generalized Jensen-Shannon divergence of GKD [1]:

$$D_\beta^{\mathrm{JSD}}(q \,\|\, p) = \beta\, D_{\mathrm{KL}}(q \,\|\, m) + (1-\beta)\, D_{\mathrm{KL}}(p \,\|\, m), \quad m = \beta q + (1-\beta)p, \tag{5}$$

where $\beta \to 0$ recovers the teacher-driven forward-KL and $\beta = 1$ the reverse-KL. The perception and reasoning teachers use the symmetric JSD $\beta = 0.5$, while the reference teacher uses $\beta = 1$ to recover the format of invalid rollouts.

3 Experiments

3.1 Implementation

Datasets. We evaluate on two public chest X-ray VQA benchmarks with closed-form Yes/No questions, IU-Xray [2] and MIMIC-CXR [5], both following the data processing of MMed-RAG [24], with 2,573 and 3,470 test QA pairs. The RAG knowledge base holds 6,069 reports, and OPSD is trained on the radiology_vqa set of 4,836 samples. **Baselines.** We compare No-RAG, Vanilla RAG with unaligned BiomedCLIP [26] retrieval, the DPO-based methods RULE [23] and MMed-RAG [24], and Intent-RAG. **Metrics.** We report ACC and F1 over the Yes/No decision, with additional diagnostics defined in Section 4.1. **Model and**

Table 1. Main results on closed-form medical VQA, accuracy and F1 in percent, higher is better.

	IU-Xray		MIMIC-CXR	
Method	ACC	F1	ACC	F1
No-RAG	79.2	78.5	66.8	65.5
Vanilla RAG	75.6	74.8	58.1	56.9
RULE	82.1	81.4	73.7	73.0
MMed-RAG	83.8	83.2	74.2	74.1
Intent-RAG	**84.3**	**83.9**	**75.2**	**74.6**

retrieval settings. The base and student model is Qwen3-VL-4B-Instruct [14]. Retrieval encodes the statement-form question and the corpus chunks with the BiomedCLIP text encoder and ranks them by text similarity through FAISS [3], with a default budget of $k = 3$.

3.2 Main Results

Table 1 shows three trends. First, naive retrieval hurts: Vanilla RAG falls below No-RAG, from 66.8 to 58.1 ACC on MIMIC-CXR and 79.2 to 75.6 on IU-Xray, which we attribute to template redundancy and the modality mismatch of image-driven retrieval. Second, preference alignment helps but stays bounded: RULE and MMed-RAG reach 73.7 and 74.2 ACC on MIMIC-CXR, recovering the drop yet still trailing Intent-RAG because they correct model behavior while leaving the noisy corpus untouched. Third, Intent-RAG is best on both datasets, reaching 75.2 ACC on MIMIC-CXR and 84.3 on IU-Xray, exceeding No-RAG by 8.4 and 5.1. The margin is larger on the more heavily templated MIMIC-CXR, and F1 tracks ACC throughout, ruling out a Yes/No class bias artifact.

3.3 Ablation Studies

Table 2 starts from Vanilla RAG and adds our components in turn. Corpus pruning delivers the largest jump, lifting MIMIC-CXR ACC from 58.1 to 73.6 and IU-Xray from 75.6 to 82.4, because it replaces the noisy full corpus with a compact core corpus and aligns retrieval to the statement form of the question. Contrasted with the negative effect of Vanilla RAG in Table 1, this isolates retrieval-signal alignment, not retrieval itself, as the decisive factor. Adding the OPSD teachers lifts accuracy to 75.2 and 84.3, the perception teacher contributing the most, the reasoning teacher adding critical use of retrieval, and the reference teacher mainly restoring format on invalid rollouts.

Table 2. Ablation from Vanilla RAG through corpus pruning and the three OPSD teachers added in turn, with accuracy and F1 in percent.

	IU-Xray		MIMIC-CXR	
Configuration	ACC	F1	ACC	F1
Vanilla RAG	75.6	74.8	58.1	56.9
+ corpus pruning	82.4	81.8	73.6	72.9
+ perception teacher	83.4	82.9	74.5	73.8
+ reasoning teacher	84.1	83.6	**75.4**	74.4
+ reference teacher	**84.3**	**83.9**	75.2	**74.6**

4 Discussion

4.1 Robustness of intention-aware OPSD under corrupted retrieval

To test whether our accuracy gain reflects genuine visual grounding rather than better-retrieved text, we corrupt retrieval at inference time. For each test sample we build three corruption types, *wrong* chunks giving contradictory conclusions from other cases, *irrelevant* chunks from random reports, and *contradictory* chunks describing the opposite of what the image shows. The *Retrieval Answer Leakage Rate* measures how often a correct answer flips to the corrupted conclusion: with $\hat{y}_{\text{clean}}$ and $\hat{y}_c$ the predictions under clean and corrupted retrieval, we let $\mathcal{S} = \{i : \hat{y}_{\text{clean},i} = a_i^*\}$ collect the samples answered correctly under clean retrieval and define $\text{RALR}_c = |\{i \in \mathcal{S} : \hat{y}_{c,i} = \neg a_i^*\}|/|\mathcal{S}|$ as the fraction of $\mathcal{S}$ that flips to the opposite conclusion under corruption c, where lower is better. RALR is defined for the *wrong* and *contradictory* corruptions that push toward a specific conclusion. The *Retrieval Robustness* $\text{RR}_c = \text{ACC}_c/\text{ACC}_{\text{clean}}$, averaged over the three corruptions, is the accuracy retained under corruption, where closer to 1 is better.

Table 3 shows that Intent-RAG degrades and leaks the least, reaching an RR of 0.87 on MIMIC-CXR and 0.89 on IU-Xray against 0.72 to 0.83 for the baselines, and cutting RALR to 15.3% and 12.4% from above 18%. This matches the design, where the perception teacher keeps perception image-grounded and the reasoning teacher critical of retrieval.

4.2 Trade-off between corpus size and retrieval performance

Varying the number of clusters K yields core corpora of increasing size. On a fixed MIMIC-CXR subset of $n = 199$ samples, Table 4 traces size versus performance under question-aligned retrieval with top-$k = 3$ and shows a clear inverted-U trend: a tiny coreset at $K = 2000$, only 4.8% of the corpus, lacks coverage and drops below No-RAG, but $K = 8000$ at 19.4% reaches the peak accuracy of 73.9, edging above the full corpus (73.4) and the deduplicated pool (72.4), while enlarging further saturates and declines slightly. This confirms that template redundancy is genuine noise with a sweet spot at about 20% of the corpus.

Table 3. Counterfactual retrieval robustness under normal retrieval (Clean) and the wrong, irrelevant, and contradictory corruptions ($n = 200$ per dataset, top-$k = 3$). RR averages robustness across corruptions, and RALR is the leakage rate in percent averaged over the wrong and contradictory corruptions. Higher is better for Clean and RR, lower for RALR.

	Accuracy under retrieval condition (%)					
Method	Clean	Wrong	Irrel.	Contr.	RR	RALR (%)
MIMIC-CXR						
Vanilla RAG	58.1	41.3	43.5	40.7	0.72	38.2
RULE	73.7	56.9	60.6	56.6	0.79	23.8
MMed-RAG	74.2	59.8	62.3	59.3	0.82	22.4
Intent-RAG	**75.2**	**64.2**	**67.5**	**65.1**	**0.87**	**15.3**
IU-Xray						
Vanilla RAG	75.6	54.2	59.8	56.5	0.75	30.5
RULE	82.1	65.1	68.9	65.4	0.81	20.6
MMed-RAG	83.8	68.6	72.1	68.9	0.83	18.9
Intent-RAG	**84.3**	**74.1**	**76.8**	**74.2**	**0.89**	**12.4**

Table 4. Coreset selection on MIMIC-CXR closed-form VQA (question-aligned retrieval, top-$k = 3$, $n = 199$ samples). Size is the number of retained chunks and its ratio to the full corpus. Accuracy and macro-F1 are in percent, higher is better.

Corpus	Size	Ratio	ACC	Macro-F1
No-RAG	0	–	66.8	65.5
Coreset (K=2000)	1998	4.8%	62.8	62.8
Coreset (K=5000)	4999	12.1%	71.9	71.6
Coreset (K=8000)	8000	19.4%	**73.9**	**73.7**
Dedup only	14926	36.1%	72.4	72.1
Full corpus	41327	100%	73.4	73.0

5 Conclusion

We tackle RAG-based medical VQA from two complementary angles. On the corpus side, the data distribution of a case-based clinical corpus matters as much as the model itself. Templated clinical reports are highly redundant, and fine-grained corpus curation distills them into a compact core corpus that directly improves retrieval. On the model side, intention-guided OPSD uses differently-privileged teachers as targeted priors to elicit visual capabilities the model already possesses, keeping perception image-grounded and reasoning critical of retrieval. On MIMIC-CXR and IU-Xray, Intent-RAG surpasses the strongest baseline in accuracy by 1.0 and 0.5 while shrinking the knowledge base by more than 80%, and it degrades and leaks the least under corrupted retrieval. Future work includes scaling to larger corpora, more modalities, and open-ended VQA.

References

1. Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos, S., Geist, M., Bachem, O.: On-policy distillation of language models: Learning from self-generated mistakes. In: International Conference on Learning Representations (ICLR) (2024)
2. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., Clements, B.: Preparing a collection of radiology examinations for distribution and retrieval. J. Am. Med. Inform. Assoc. **23**(2), 304–310 (2016)
3. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library (2024)
4. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network (2015)
5. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Sci. Data **6**, 317 (2019)
6. Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., Carlini, N.: Deduplicating training data makes language models better. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL) (2022)
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)
8. 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 (NeurIPS), Datasets and Benchmarks Track (2023)
9. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
10. Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W.: A survey on hallucination in large vision-language models (2024)
11. Lopez Pellicer, A., Angelov, P., Bukhari, M., Li, Y., Soares, E., Kerns, J.: Protomedagent: Multimodal clinical interpretability via privacy-aware agentic workflows (2026)
12. Ma, Y., Liang, Z., Chen, B., et al.: Sd-search: On-policy hindsight self-distillation for search-augmented reasoning (2026)
13. Moor, M., Huang, Q., Wu, S., Yasunaga, M., Zakka, C., Dalmia, Y., Rajpurkar, P., Leskovec, J.: Med-flamingo: a multimodal medical few-shot learner. In: Conference on Health, Inference, and Learning (CHIL) (2023)
14. Qwen Team: Qwen3-vl technical report (2025)
15. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Jain, S., Safron, G., Lockett, A., Krueger, D., Mishkin, F., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML) (2021)
16. Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)
17. Shao, Z., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models (2024)

18. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Toward expert-level medical question answering with large language models. Nature Medicine (2024). https://doi.org/10.1038/s41591-024-03423-7
19. Tian, K., Liu, S., Yan, Z., Xia, S., Dong, S., Wang, Y.: Vicur: Visual cues as recoverable privilege for multimodal on-policy distillation (2026)
20. Wang, S., Liu, X., Liu, L., Han, Z.: Seeing before reasoning: Decoupling perception and reasoning for shortcut-resilient multimodal on-policy self-distillation (2026)
21. Wang, Z., Chen, Z., Ye, Z., Zhu, H., Zheng, Y., Xia, Y.: Invic: Intent-aware visual cues for medical visual question answering (2026)
22. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)
23. Xia, P., Zhu, K., et al.: Rule: Reliable multimodal rag for factuality in medical vision language models. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2024)
24. Xia, P., et al.: Mmed-rag: Versatile multimodal rag system for medical vision language models. In: International Conference on Learning Representations (ICLR) (2025)
25. Yuan, Q., Lou, J., Yu, X., Lin, H., Sun, L., Han, X., Lu, Y.: Vision-opd: Learning to see fine details for multimodal llms via on-policy self-distillation (2026)
26. Zhang, S., Xu, Y., Usuyama, N., et al.: A multimodal biomedical foundation model trained from fifteen million image–text pairs. NEJM AI **2**(1) (2025). https://doi.org/10.1056/AIoa2400640