VRUGA: Visual and Reasoning Uncertainty Guided Active Learning for Medical MLLM Post-training

Boyi Xiao[1,2], Jianghao Wu[4], Rongzhao Zhang[2], Jie Xu[2], Feng Zhao[1†], Guotai Wang[2,3†], and Shaoting Zhang[2,3]

[1] Department of Automation, University of Science and Technology of China, Hefei, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China
[4] Department of Data Science & AI, Monash University, Melbourne, Australia
{fzhao956@ustc.edu.cn, guotai.wang@uestc.edu.cn} [†] Corresponding authors.

Abstract. Medical Multimodal Large Language Models (MLLMs) hold significant potential for clinical visual question answering (VQA), yet their reliability is often compromised by unstable visual grounding and flawed reasoning logic. While reinforcement learning-based post-training, such as Group Relative Policy Optimization (GRPO), offers a path to steer model behavior, its efficacy is severely constrained by the high cost of expert supervision. To address this, we propose VRUGA, an active learning framework designed to maximize GRPO post-training efficiency under a constrained expert budget. Specifically, we introduce Dual-Uncertainty Quantification (DUQ) to identify high-risk failure modes by jointly evaluating visual grounding stability and reasoning trajectory consistency. Logic-Aware Diversity Sampling (LADS) is subsequently employed to select a representative active set by leveraging decoder hidden-state embeddings to capture a diverse spectrum of latent reasoning patterns. Finally, a Cross-Threshold Pseudo-Labeling (CTPL) mechanism stabilizes the alignment process through tiered supervision, integrating expert-corrected samples with majority-voted pseudo-label anchors to mitigate policy overfitting. Extensive evaluations on VQA-RAD, SLAKE, and PathVQA benchmarks demonstrate that VRUGA achieves a 75.2% average accuracy, significantly outperforming state-of-the-art active learning baselines while requiring only a 10% expert annotation budget.

Keywords: Medical Multimodal Large Language Models · Active Learning · Uncertainty Quantification · Reinforcement Learning

1 Introduction

Medical Multimodal Large Language Models (MLLMs) have shown strong potential for medical visual question answering (VQA) [33,30], but their relia-

bility remains constrained by unstable visual grounding and multi-step reasoning errors [28,26]. Recently, reinforcement-learning-based post-training (e.g., Group Relative Policy Optimization, GRPO [5]) has emerged as a promising way to steer MLLMs toward more accurate and stable behavior in medical VQA [8,15,24]. However, such post-training critically depends on high-quality expert references, which are scarce and expensive in medicine. This shifts the challenge from how to optimize with labels to which cases to label: under a fixed expert budget, the goal is to identify high-value samples for annotation while maintaining stable post-training and avoiding policy drift.

Active learning (AL) is a natural strategy for reducing expert labeling cost by selecting the most valuable unlabeled samples for annotation [20,17]. However, most existing AL methods are developed for classification or segmentation, where uncertainty [12,21,31,16] and sample similarity [13,23,29,4,3] are relatively well defined. These assumptions do not transfer well to generative medical VQA with MLLMs, where correctness depends not only on the final answer, but also on whether the model is grounded on consistent visual evidence and follows a stable reasoning process [11]. In this setting, the key challenge is to distinguish genuine clinical failure risk from superficial output variability. For generative medical VQA, predictive variation may arise from decoding randomness, or from deeper instability in visual grounding and reasoning. Therefore, effective querying should measure uncertainty in terms of evidence consistency and reasoning reliability, rather than final-answer confidence alone.

Even with improved uncertainty estimates, selecting the top-ranked uncertain samples can still be inefficient due to redundancy [2,19]. High-uncertainty cases may cluster around recurring failure patterns, such as similar question types or challenging imaging conditions, leading to diminishing returns when expert effort is spent on near-duplicate samples. Moreover, in RL-based post-training, aligning only on a limited set of difficult cases can induce policy drift, as the model may over-specialize to sparse corrective signals and lose stability on routine cases [14]. Under limited supervision, it is therefore crucial to improve high-risk failure cases while preserving reliable behavior on routine examples.

To address these challenges, we propose VRUGA, an active learning framework for efficient post-training of medical MLLMs. First, VRUGA introduces Dual-Uncertainty Quantification (DUQ), which goes beyond text-only uncertainty by jointly measuring visual grounding instability and reasoning trajectory uncertainty. By detecting cross-modal misalignments and logical hesitation, DUQ identifies clinically risky samples that standard logit-based metrics often overlook. Second, we propose Logic-Aware Diversity Sampling (LADS) to reduce redundancy by selecting samples that cover diverse reasoning patterns in the decoder hidden-state space. Third, we propose Cross-Threshold Pseudo-Labeling (CTPL) to stabilize GRPO-based post-training by combining expert corrections on high-uncertainty cases with pseudo-label anchors on high-confidence samples, effectively preventing policy over-specialization on limited active sets. Experiments show that VRUGA achieves 75.2% average accuracy, outperforming the state-of-the-art DEUCE (72.5%). Notably, on SLAKE, it recovers 97.8% of

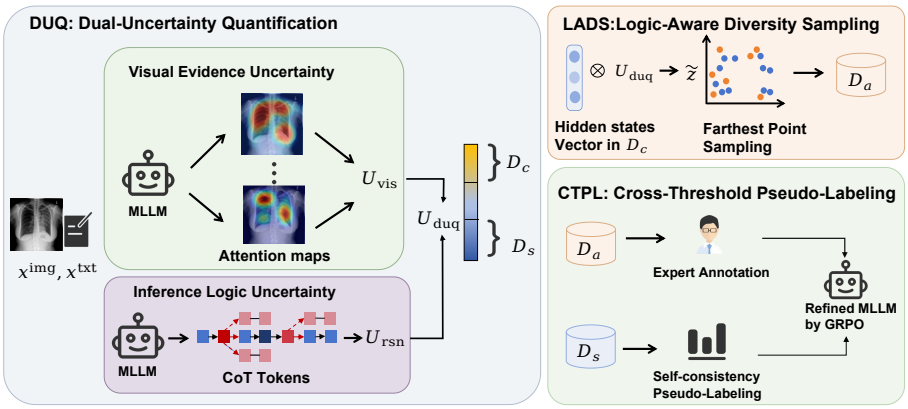

Fig. 1. Overview of our VRUGA framework that integrates DUQ, LADS, and CTPL for active sample selection and stable post-training for medical MLLM.

the full-data performance using only 10% expert labels, significantly reducing annotation costs.

2 Method

2.1 Problem Settings and Method Overview

In the medical MLLM post-training, we consider a pre-trained multimodal model $\mathcal{M}_\theta$, a limited expert annotation budget $B \ll N$, and an unlabeled medical VQA dataset $\mathcal{D}_u = \{(\mathbf{x}_i^{\text{img}}, \mathbf{x}_i^{\text{txt}})\}_{i=1}^{N}$. Here, $\mathbf{x}_i^{\text{img}}$ denotes a medical image and $\mathbf{x}_i^{\text{txt}}$ denotes the corresponding question. The objective is to post-train $\mathcal{M}_\theta$ with GRPO [5] by selecting the most informative samples in $\mathcal{D}_u$ for expert annotation, thereby improving VQA performance under a limited labeling budget.

As illustrated in Fig. 1, VRUGA integrates three components to query informative samples under a constrained annotation budget: DUQ identifies high-uncertainty cases by measuring grounding instability and reasoning uncertainty; LADS selects a diverse active set in the decoder-state embedding space; and CTPL balances expert corrections with high-confidence pseudo-label anchors for stable post-training.

2.2 Dual-Uncertainty Quantification (DUQ)

Standard output-level metrics for vision or language tasks, such as logit entropy, often fail to capture the instability of internal decision-making. To better

measure prediction uncertainty in VQA tasks, we instead quantify uncertainty through two complementary perspectives: visual grounding and reasoning logic.

Visual Evidence Uncertainty (U_{vis}). To assess whether the model relies on consistent visual evidence when generating the answer, we measure the stability of attention from generated text tokens to visual tokens under stochastic generation. Given an input $x = (\mathbf{x}^{\text{img}}, \mathbf{x}^{\text{txt}})$, we sample K responses with temperature $T > 0$. For the k-th trial, let $\mathbf{y}^{(k)} = \{y_1^{(k)}, y_2^{(k)}, \ldots, y_{L_k}^{(k)}\}$ be the sequence of L_k generated answer tokens. For each token $y_j^{(k)}$, we extract an attention vector $\mathbf{a}_j^{(k)} \in \mathbb{R}^{N_v}$ from the last multimodal fusion layer, where N_v is the number of visual tokens (corresponding to image patches). Each entry in $\mathbf{a}_j^{(k)}$ indicates the attention weight assigned to a specific visual token when predicting $y_j^{(k)}$. We aggregate these weights into a trial-specific evidence map $\mathbf{M}^{(k)} \in \mathbb{R}^{N_v}$:

$$\mathbf{M}^{(k)} = \frac{1}{L_k} \sum_{j=1}^{L_k} \mathbf{a}_j^{(k)}. \tag{1}$$

We then define the salient patch set $\mathcal{S}_k$ as the set of indices corresponding to the top-m highest values in $\mathbf{M}^{(k)}$ (with $m \ll N_v$), capturing the primary diagnostic regions. The visual uncertainty U_{vis} is computed as the complement of the average pairwise Jaccard similarity across K trials:

$$U_{\text{vis}} = 1 - \frac{2}{K(K-1)} \sum_{1 \le i < j \le K} \frac{|\mathcal{S}_i \cap \mathcal{S}_j|}{|\mathcal{S}_i \cup \mathcal{S}_j|}, \tag{2}$$

where high U_{vis} indicates non-convergent evidence localization; i.e., even if the generated text is consistent across trials, it is supported by inconsistent visual evidence.

Reasoning Trajectory Uncertainty (U_{rsn}). To quantify logical hesitation, we measure the average next-token entropy along the Chain-of-Thought (CoT) tokens (excluding the final answer span). Let $\mathcal{T} = \{w_1, w_2, \ldots, w_{L_{\text{CoT}}}\}$ denote the CoT token sequence generated on-policy. We define

$$U_{\text{rsn}} = -\frac{1}{L_{\text{CoT}}} \sum_{t=1}^{L_{\text{CoT}}} \sum_{v \in \mathcal{V}} P(v \mid x, w_{<t}) \log P(v \mid x, w_{<t}), \tag{3}$$

where $\mathcal{V}$ is the vocabulary and $P(\cdot \mid x, w_{<t})$ is the model's on-policy next-token distribution under the fixed decoding configuration. Unlike final-answer entropy, U_{rsn} captures uncertainty accumulated throughout the reasoning trajectory; persistently high entropy suggests unstable reasoning. We compute the final DUQ score as:

$$U_{\text{duq}} = \alpha \cdot U_{\text{vis}} + (1 - \alpha) \cdot U_{\text{rsn}}, \tag{4}$$

where $\alpha \in [0, 1]$ is a hyperparameter that balances the contribution of visual grounding stability and reasoning logic consistency.

DUQ-Guided Data Partition. Based on U_{duq}, we partition the unlabeled pool into three functional subsets. The candidate pool $\mathcal{D}_{\text{c}} = \{x \mid U_{\text{duq}}(x) > \tau_{\text{high}}\}$ identifies highly uncertain cases that require manual expert intervention for ground-truth labeling. Conversely, the stable set $\mathcal{D}_s = \{x \mid U_{\text{duq}}(x) < \tau_{\text{low}}\}$ contains samples where the model is confident enough for reliable self-supervised pseudo-labeling. Samples with intermediate uncertainty are omitted from training to avoid ambiguous supervision and ensure the stability of the knowledge anchors.

2.3 Logic-Aware Diversity Sampling (LADS)

While high DUQ scores identify informative failures, selection based solely on uncertainty leads to logical redundancy due to sample clustering. To ensure broad coverage of distinct reasoning patterns, we propose Logic-Aware Diversity Sampling (LADS). LADS integrates DUQ scalars with logical embeddings to perform weighted diversity sampling within the candidate pool $\mathcal{D}_{\text{c}}$.

To achieve this, we construct an uncertainty-weighted logic vector $\tilde{\mathbf{z}}_i$:

$$\tilde{\mathbf{z}}_i = U_{\text{duq},i} \cdot \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2}, \quad \text{where} \quad \mathbf{z}_i = \frac{1}{|\mathcal{T}_i|} \sum_{t \in \mathcal{T}_i} \mathbf{h}_{i,t} \tag{5}$$

where $\mathbf{h}_{i,t}$ is the last-layer decoder hidden state of the CoT token $t \in \mathcal{T}_i$. The logic embedding $\mathbf{z}_i$, obtained by mean-pooling $\mathbf{h}_{i,t}$ over the CoT tokens, captures a coarse representation of the reasoning trajectory, while the DUQ score $U_{\text{duq},i}$ acts as a scaling factor indicating the estimated risk of visual grounding or reasoning failure. To select the final active set $\mathcal{D}_a$ of size B, we employ an iterative farthest point sampling strategy in this weighted space. Starting with the sample with the highest U_{duq}, we iteratively select the next sample $i^* \in \mathcal{D}_{\text{c}}$ that maximizes its distance to the current set $\mathcal{D}_a$:

$$i^* = \arg \max_{i \in \mathcal{D}_{\text{c}} \setminus \mathcal{D}_a} \left(\min_{j \in \mathcal{D}_a} \|\tilde{\mathbf{z}}_i - \tilde{\mathbf{z}}_j\|_2 \right) \tag{6}$$

This strategy ensures that expert intervention is focused on samples that are both highly informative and logically distinct, preventing the budget from being consumed by repetitive error patterns and ensuring a more comprehensive coverage of the cognitive failure space.

2.4 Post-training with Cross-Threshold Pseudo-Labeling (CTPL)

To prevent overfitting on the limited active set $\mathcal{D}_a$, CTPL employs a tiered GRPO reward strategy. By combining expert supervision on high-uncertainty samples with pseudo-label anchoring on high-confidence ones, it maintains robust performance across routine clinical scenarios while leveraging high-precision guidance.

We optimize GRPO on the expert-labeled active set $\mathcal{D}_a$ and the pseudo-labeled anchor set $\mathcal{D}_s$. For each sample $x_i \in \mathcal{D}_a$, we use its expert reference y_i^{exp};

for each sample $x_i \in \mathcal{D}_s$, we use its majority-voted pseudo-label $\hat{y}_i^{\text{psd}}$ obtained from K stochastic trials. The total objective is

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{x_i \in \mathcal{D}_a} \left[\mathcal{L}_{\text{GRPO}}(x_i, y_i^{\text{exp}}) \right] + \gamma \, \mathbb{E}_{x_i \in \mathcal{D}_s} \left[\mathcal{L}_{\text{GRPO}}(x_i, \hat{y}_i^{\text{psd}}) \right]. \tag{7}$$

where γ controls the trade-off between high-precision expert supervision and the stability provided by high-confidence pseudo-label anchors.

3 Experiments

3.1 Experimental Setup

Datasets and Backbone We evaluate VRUGA on closed-ended subsets of three medical VQA benchmarks: **VQA-RAD**[9] (2,248 QAs) contains radiology images (CT, MRI, X-ray) spanning head, chest, and abdominal pathologies; **SLAKE**[10] (4,923 English QAs) features multi-organ CT and MRI scans for abnormality detection across various body regions; and **PathVQA**[7] (16,334 QAs) provides histopathology images for tissue-level disease classification. Together, these datasets cover a wide range of clinical scenarios in radiology and pathology. We adopt **Qwen2.5-VL-7B**[27] as the foundational MLLM.

Implementation Details For each dataset, τ_{high} and τ_{low} were set as the 85th and 50th percentiles of U_{duq}, respectively. The active set size B is 10% of total samples, and the stable set $\mathcal{D}_s$ comprises the bottom 50%. We follow the official train/test splits of each benchmark and draw the active learning pool from the training set only. For U_{vis}, cross-attention is extracted from the last multimodal fusion layer with head-averaged weights, and m is set to top-10% of N_v. We use Chain-of-Thought (CoT) prompting [25] to elicit reasoning trajectories. All experiments use PyTorch on 2 × NVIDIA H200 GPUs. For GRPO, we set group size $G = 8$, KL penalty 0.01, and 200 global steps with AdamW (lr 1×10^{-5}, cosine decay). DUQ uses $K = 8$ trials at $T = 0.7$ with $\alpha = 0.5$; CTPL anchor weight $\gamma = 0.1$. All results are averaged over 3 random seeds. Following [9,1], exact-match accuracy is the primary metric.

3.2 Comparison with State-of-the-art Methods

We evaluate VRUGA against random sampling and five state-of-the-art active learning (AL) methods. As shown in Table 1, VRUGA consistently outperforms traditional heuristics. In terms of average accuracy, VRUGA improves over Random sampling (70.7%) and uncertainty-based Entropy [18] (71.2%) by 4.5 and 4.0 percentage points, respectively. This suggests that simple logit-based uncertainty is insufficient to capture the multimodal failure modes of medical VQA. Compared with diversity-oriented strategies such as VOTE-k [22] (71.9%) and PATRON [32] (71.8%), VRUGA further achieves average gains of 3.3 and 3.4 percentage points, respectively. The advantage is particularly evident on PathVQA,

Table 1. Accuracy (%) on medical VQA benchmarks under 10% expert budget. The first two lines represent the upper bound (100% data) and lower bound (zero-shot), respectively. Best results are **bolded**; second-best are underlined.

Method	Training Data	VQA-RAD	SLAKE	PathVQA	Avg. ↑
Qwen2.5-VL-7B (GRPO)	100% Expert	70.5	79.3	82.8	77.5
Qwen2.5-VL-7B (Base)	0% (Zero-shot)	67.3	71.6	65.5	68.1
Random	10% Expert	68.1	73.2	70.7	70.7
Entropy [18]	10% Expert	68.2	74.1	71.2	71.2
Core-set [19]	10% Expert	68.3	73.9	71.5	71.2
VOTE-k [22]	10% Expert	68.5	74.4	72.8	71.9
PATRON [32]	10% Expert	68.4	74.8	72.2	71.8
DEUCE [6]	10% Expert	68.6	75.5	73.5	72.5
VRUGA (w/o Unl.)	10% Expert	68.6	76.2	74.3	73.0
VRUGA (Ours)	10% Exp. + 50% Unl.	**68.8**	**77.6**	**79.1**	**75.2**

where VRUGA reaches 79.1% accuracy, outperforming VOTE-k and PATRON by 6.3 and 6.9 percentage points, respectively. Furthermore, VRUGA surpasses the previous state of the art, DEUCE [6] (72.5%), by 2.7 percentage points in average accuracy. Most notably, on the SLAKE benchmark, VRUGA attains 77.6% accuracy, recovering 97.8% of the full-data GRPO upper bound (79.3%) while requiring only 10% expert annotations, together with additional unlabeled anchor samples for CTPL.

Scalability is validated in Fig. 2(a), where VRUGA maintains a superior margin across all annotation budgets (2%–20%). The advantage is most pronounced in scenarios with extremely low annotation ratios (2%–5%), where its steeper learning curve demonstrates the efficiency of dual-uncertainty guidance in identifying informative samples.

3.3 Ablation and Parameter Analysis

Strategy Effectiveness. Table 2 confirms the synergy of VRUGA's components. In Part (A), the incremental addition of DUQ and LADS boosts average accuracy to 73.0%, while the inclusion of CTPL yields the largest gain (+2.2%). This significant jump underscores that high-precision expert labels require stable knowledge anchors to prevent policy drift during reinforcement learning, effectively bridging the gap between error correction and routine stability. Regarding uncertainty strategies in Part (B), individual visual (U_{vis}) and reasoning (U_{rsn}) metrics outperform text-only entropy, but their integration in DUQ is vital for mitigating hallucinations through complementary grounding views. Finally, Part (C) shows that LADS using decoder hidden states outperforms ViT-based or LLM-based features. This superiority confirms that the most informative diversity for medical VQA resides in the model's internal reasoning trajectories rather than in superficial feature distances, indicating that diversity sampling is most effective in a latent space that has already integrated cross-modal information.

Table 2. Ablation study of VRUGA: (A) module contributions; (B) uncertainty metrics; (C) diversity sampling latent spaces.

Block Variant Configuration	VQA-RAD	SLAKE	PathVQA	Avg. ↑
(A) Incremental Module Analysis				
Zero-shot Baseline	67.3	71.6	65.5	68.1
Random Selection	68.1	73.2	70.7	70.7
+ DUQ	68.4	74.5	72.1	71.7
+ DUQ + LADS	68.6	76.2	74.3	73.0
Full VRUGA (+CTPL)	**68.8**	**77.6**	**79.1**	**75.2**
(B) Uncertainty Strategy Analysis (Fixed LADS + CTPL)				
Logit Entropy (Text-only)	68.2	74.1	71.2	71.2
U_{vis} Only (Perceptual)	68.4	75.2	73.4	72.3
U_{rsn} Only (Cognitive)	68.5	75.8	72.8	72.4
DUQ (Full Perceptual + Cognitive)	**68.8**	**77.6**	**79.1**	**75.2**
(C) Diversity Selection Space Analysis (Fixed DUQ + CTPL)				
No Diversity (Random)	68.4	74.5	72.1	71.7
Visual Feature (ViT-based)	68.5	75.3	73.1	72.3
Textual Feature (LLM-based)	68.6	75.5	73.4	72.5
LADS (Decoder Hidden States)	**68.8**	**77.6**	**79.1**	**75.2**

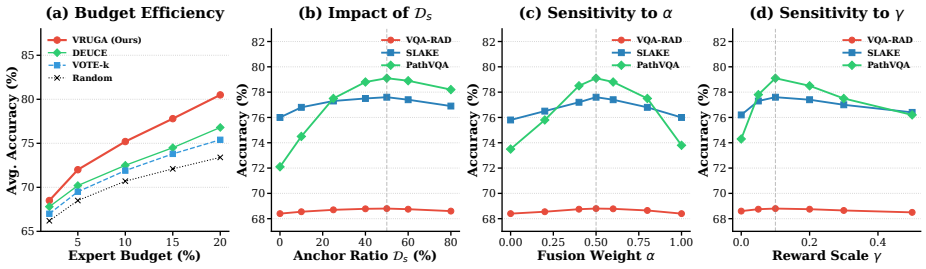

Fig. 2. Sensitivity and efficiency analysis: (a) budget efficiency; (b) anchor ratio $\mathcal{D}_s$; (c) fusion weight α; and (d) reward scale γ.

Hyperparameter Sensitivity. Fig. 2(b)-(d) evaluates hyperparameter sensitivity across all datasets. Accuracy universally peaks at an anchor ratio $\mathcal{D}_s = 50\%$ (Fig. 2b), as excessive ratios introduce reward-contaminating noise while lower ratios limit anchor diversity. For uncertainty fusion, $\alpha = 0.5$ (Fig. 2c) achieves the optimal balance between visual grounding and reasoning consistency. Finally, a reward scale $\gamma = 0.1$ (Fig. 2d) provides the best trade-off by leveraging unlabeled anchors without over-relying on pseudo-labels. The consistency of these optimal settings validates the generalizability of VRUGA.

4 Conclusion

We present VRUGA, a dual-uncertainty-guided active learning framework for medical MLLM post-training. By combining multimodal uncertainty quantification with logic-aware diversity sampling, VRUGA recovers 97.8% of the full-data

GRPO performance using only 10% expert annotations. CTPL further stabilizes RL via automated anchors. We acknowledge two limitations: (i) U_{vis} relies on attention as a grounding proxy, and verbalized CoT may not perfectly reflect internal reasoning; nonetheless, our ablations show both signals correlate with downstream error. (ii) Evaluation uses exact-match accuracy on closed-ended VQA; assessing grounding faithfulness with open-ended metrics is left for future work. VRUGA offers a scalable, cost-effective solution for aligning medical MLLMs with clinical reasoning demands.

Acknowledgments. This work was supported by the National Natural Science Foundation of China under Grant 62271115, Shanghai Artificial Intelligence Laboratory.

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

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
2. Caramalau, R., Bhattarai, B., Kim, T.K.: Sequential graph convolutional network for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9583–9592 (2021)
3. Chen, J., Ma, B., Cui, H., Xia, Y.: Think twice before selection: Federated evidential active learning for medical image analysis with domain shifts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11439–11449 (2024)
4. Elhamifar, E., Sapiro, G., Yang, A., Sasrty, S.S.: A convex optimization framework for active learning. In: Proceedings of the IEEE international conference on computer vision. pp. 209–216 (2013)
5. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint (2025)
6. Guo, J., Philip Chen, C., Li, S., Zhang, T.: Deuce: Dual-diversity enhancement and uncertainty-awareness for cold-start active learning. Transactions of the Association for Computational Linguistics **12**, 1736–1754 (2024)
7. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering. arXiv preprint arXiv:2003.10286 (2020)
8. Hu, Y., Xu, C., Lin, B., Yang, W., Tang, Y.Y.: Medical multimodal large language models: A systematic review. Intelligent Oncology (2025)
9. 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)
10. 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)

11. Liu, R., Yu, D., Zheng, T., Dai, R., Li, Z., Yu, W., Liang, Z., Song, L., Mi, H., Tokekar, P., et al.: Vogue: Guiding exploration with visual uncertainty improves multimodal reasoning. arXiv preprint arXiv:2510.01444 (2025)
12. Monarch, R.M.: Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Simon and Schuster (2021)
13. Nguyen, H.T., Smeulders, A.: Active learning using pre-clustering. In: Proceedings of the twenty-first international conference on Machine learning. p. 79 (2004)
14. Noukhovitch, M., Lavoie, S., Strub, F., Courville, A.C.: Language model alignment with elastic reset. Advances in Neural Information Processing Systems **36**, 3439–3461 (2023)
15. Pan, J., Liu, C., Wu, J., Liu, F., Zhu, J., Li, H.B., Chen, C., Ouyang, C., Rueckert, D.: Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 337–347. Springer (2025)
16. Parvaneh, A., Abbasnejad, E., Teney, D., Haffari, G.R., Van Den Hengel, A., Shi, J.Q.: Active learning by feature mixing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12237–12246 (2022)
17. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM computing surveys (CSUR) **54**(9), 1–40 (2021)
18. Schröder, C., Niekler, A., Potthast, M.: Revisiting uncertainty-based query strategies for active learning with transformers. In: Findings of the Association for Computational Linguistics: ACL 2022. pp. 2194–2203 (2022)
19. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. arXiv preprint arXiv:1708.00489 (2017)
20. Settles, B.: Active learning literature survey (2009)
21. Shannon, C.E.: A mathematical theory of communication. The Bell system technical journal **27**(3), 379–423 (1948)
22. Su, H., Kasai, J., Wu, C.H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N.A., et al.: Selective annotation makes language models better few-shot learners. arXiv preprint arXiv:2209.01975 (2022)
23. Urner, R., Wulff, S., Ben-David, S.: Plal: Cluster-based active learning. In: Conference on learning theory. pp. 376–397. PMLR (2013)
24. Wang, L., Qin, Y., Yang, H., Li, X.: Proactive reasoning-with-retrieval framework for medical multimodal large language models. arXiv preprint arXiv:2510.18303 (2025)
25. Wei, J., Wang, X., Schuurmans, D., Maarten, B., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
26. Xiao, J., Li, Q., Yang, Y., Qiu, L., Yao, A.: Unleashing the power of llms for medical video answer localization. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 669–679. Springer (2025)
27. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F.: Qwen2 technical report. arXiv preprint (2024)
28. Yang, X., Miao, J., Yuan, Y., Wang, J., Dou, Q., Li, J., Heng, P.A.: Medical large vision language models with multi-image visual ability. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 402–412. Springer (2025)

29. Yang, Y., Ma, Z., Nie, F., Chang, X., Hauptmann, A.G.: Multi-class active learning by uncertainty sampling with diversity maximization. International Journal of Computer Vision **113**(2), 113–127 (2015)
30. Ye, J., Tang, H.: Multimodal large language models for medicine: A comprehensive survey. arXiv preprint arXiv:2504.21051 (2025)
31. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 93–102 (2019)
32. Yu, Y., Zhang, R., Xu, R., Zhang, J., Shen, J., Zhang, C.: Cold-start data selection for better few-shot language model fine-tuning: A prompt-based uncertainty propagation approach. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 2499–2521 (2023)
33. Zhou, H., Liu, F., Gu, B., Zou, X., Huang, J., Wu, J., Li, Y., Chen, S.S., Zhou, P., Liu, J., et al.: A survey of large language models in medicine: Progress, application, and challenge. arXiv preprint arXiv:2311.05112 (2023)