Act as you think: Reinforcing Consistent Reasoning in Medical Visual Question Answering

Songtao Jiang[1] **Yuan Wang**[1] **Ruizhe Chen**[1] **Yan Zhang**[1] **Ruilin Luo**[2] **Bohan Lei**[1]
Yeying Jin[3] **Sibo Song**[4] **Zhibo Yang**[4] **Jimeng Sun**[5] **Jian Wu**[1] **Zuozhu Liu**[1*]

[1]Zhejiang University [2]Tsinghua University [3]National University of Singapore
[4]Alibaba Group, Qwen Team [5]University of Illinois at Urbana-Champaign
[*]Corresponding authors

Abstract

While reinforcement learning from verifiable rewards (RLVR) has been proven highly effective for enhancing reasoning, its application to medical visual question answering (Med-VQA) is hampered by models producing reasoning inconsistent with either the visual evidence or the final answer. Our analysis reveals a critical flaw in RLVR training: it paradoxically encourages models to disregard visual evidence and generate answers that contradict their own reasoning. This degradation is most pronounced in specialized medical modalities (e.g., Fundus, Ultrasound) where base VLMs lack robust understanding, a failure we attribute to a flawed reward mechanism exacerbated by the scarcity of diverse training data. To tackle this, we introduce **Med-Zero-17K**, a large-scale dataset spanning over 30 modalities and 24 clinically relevant tasks, and the **Multi-Consistency Reward (MCR)** framework, which explicitly rewards both perceptual grounding and logical coherence. Extensive experiments validate our approach: integrating MCR into the RLVR framework delivers robust performance gains. This success stems from our crucial finding that rewarding internal consistency is significantly more effective than attempting to judge reasoning correctness. Furthermore, MCR proves highly versatile, exhibiting strong generalization across diverse VLM backbones, compatibility with RL algorithms like GRPO and DPO, and extending its effectiveness to 3D VQA tasks and R1-style training paradigms.

1 Introduction

Recent advancements have significantly improved the accuracy of Medical Visual Question Answering (Med-VQA) (Jiang et al., 2024; Li et al., 2024a; Kim et al., 2024; Jiang et al., 2025a). However, the high-stakes nature of the medical domain has precipitated an urgent shift in focus towards the interpretability and transparency of the reasoning processes underlying these decisions (Bouazizi and Ltifi, 2024; Dong et al., 2025). Reinforcement Learning from Verifiable Rewards (RLVR), a paradigm that enhances reasoning by rewarding verifiable outcomes like correctness, has shown great promise in general domains (Li et al., 2025; Yue et al., 2025). Approaches like Deepseek-R1 (Guo et al., 2025) demonstrate that RLVR can foster robust reasoning without extensive, high-quality Chain-of-Thought (CoT) data. This makes it a particularly appealing avenue for the medical field, where such specialized reasoning data is inherently scarce (Liu et al., 2025b, 2024; Zhang et al., 2025a,c,b).

Despite its promise, the direct application of RLVR to Med-VQA yields only limited improvements (Liu et al., 2025c; Bai et al., 2024b). We observe two fundamental failure modes. First, as depicted in Figure 1(a), models exhibit *perceptual-reasoning misalignment*, generating plausible-sounding rationales that bypass correct image analysis. Second, Figure 1(b) illustrates *reasoning-answer inconsistency*, where the final answer fails to logically follow from the stated reasoning. These errors highlight a significant gap between achieving correctness and ensuring trustworthy reasoning.

To understand if these failures are isolated incidents or a systemic issue, we conducted an in-depth analysis of the RLVR training dynamics. Our investigation reveals a *paradoxical degradation*: as accuracy rewards increase, models paradoxically learn to disregard visual evidence and generate more answers inconsistent with their own reasoning (Figure 1(c)). Crucially, this failure is most pronounced in specialized medical modalities (e.g., fundus, ultrasound) where the base VLM lacks robust understanding (Figure 1(d)). We attribute these issues to a flawed reward mechanism exacerbated by the scarcity of diverse medical datasets suitable for RLVR (Liu et al., 2025b).

To address these dual challenges, we propose a framework that jointly advances data and algorith-

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1788–1805

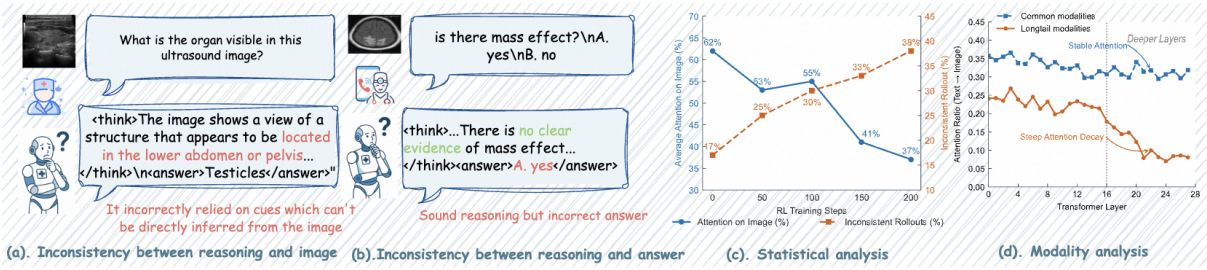

Figure 1: Inconsistency patterns in medical decision-making reasoning. This phenomenon is more pronounced in modalities where base VLMs demonstrate limited understanding.

mic design. On the data front, we introduce **Med-Zero-17K**, a large-scale dataset specifically curated for medical RL. Spanning over 30 modalities and 24 clinical tasks, it provides the diverse and challenging scenarios necessary for robust policy optimization. On the algorithmic front, we propose the **Multi-Consistency Reward (MCR)** framework. It augments standard RLVR by directly targeting the identified failure modes with two novel reward signals: The *perceptual-cognitive consistency reward* encourages genuine visual grounding by leveraging contrastive perturbations. By rewarding the model for its ability to distinguish between clean and corrupted images—maintaining accuracy on the former while failing on the latter—it disincentivizes hallucinated reasoning and forces the model to rely on actual visual features. The *cognitive-decision consistency reward* addresses inconsistency by employing an external LLM as an automated judge. Critically, this judge does not evaluate the correctness of the reasoning, but simply verifies the logical entailment between the model's generated thought process and its final answer, ensuring the decision is a direct consequence of the stated rationale.

We validate our approach across six diverse benchmarks, encompassing in-domain and out-of-distribution tests that span broad modality understanding, knowledge-driven reasoning, and in-depth modality VQA. Across these, MCR consistently enhances the RLVR framework, with particularly pronounced gains on generalization benchmarks. This success is rooted in a central finding from our analysis: rewarding internal *consistency* provides a more stable and effective learning signal than attempting to reward absolute *correctness*. Notably, we also observe that this consistency-focused approach fosters an emergent behavior during RL training: the generation of longer and more elaborate reasoning chains. Beyond these findings, our analyses underscore MCR's versatility. We show it delivers consistent gains across diverse VLM backbones, generalizes to 3D VQA tasks without domain-specific training, and maintains compatibility with multiple RL algorithms like GRPO and DPO, even proving effective within R1-style paradigms. Collectively, these results establish MCR as a robust and practical framework for advancing trustworthy reasoning in medical AI.

2 Related Work

Reinforcement Learning with Verified Rewards. RLVR has shown strong promise in enabling reasoning without extensive supervision (Shao et al., 2024; Jiang et al., 2026; Lambert et al., 2024; Zhang et al., 2025b). GRPO (Shao et al., 2024) improves on PPO (Schulman et al., 2017) by replacing value functions with group-based scoring, and R1-Zero (Liu et al., 2025c) demonstrates that purely RL-driven training guided by rule-based rewards can foster robust reasoning without explicit CoT supervision. Consistency-based training has been explored across domains: self-consistency sampling improves CoT reasoning (Wang et al., 2022), process reward models verify step-wise correctness in math (Luo et al., 2025; Liu et al., 2025a), self-debugging enforces output-specification coherence in code (Chen et al., 2023). Our work extends this principle to medical RLVR, where we show that rewarding reasoning-answer consistency provides more stable training signals than judging reasoning correctness.

Medical Reasoning and Cross-Modality Alignment. Earlier medical VLMs focused on domain adaptation: LLaVA-Med (Li et al., 2024a), HuatuoVision (Chen et al., 2024), Med-MoE (Jiang et al., 2024, 2025a). More recent work targets reasoning capability through supervised CoT (Liu et al., 2024; Gai et al., 2024; Jiang et al., 2025a) or RL-based approaches (Lai et al., 2026; Pan et al.,

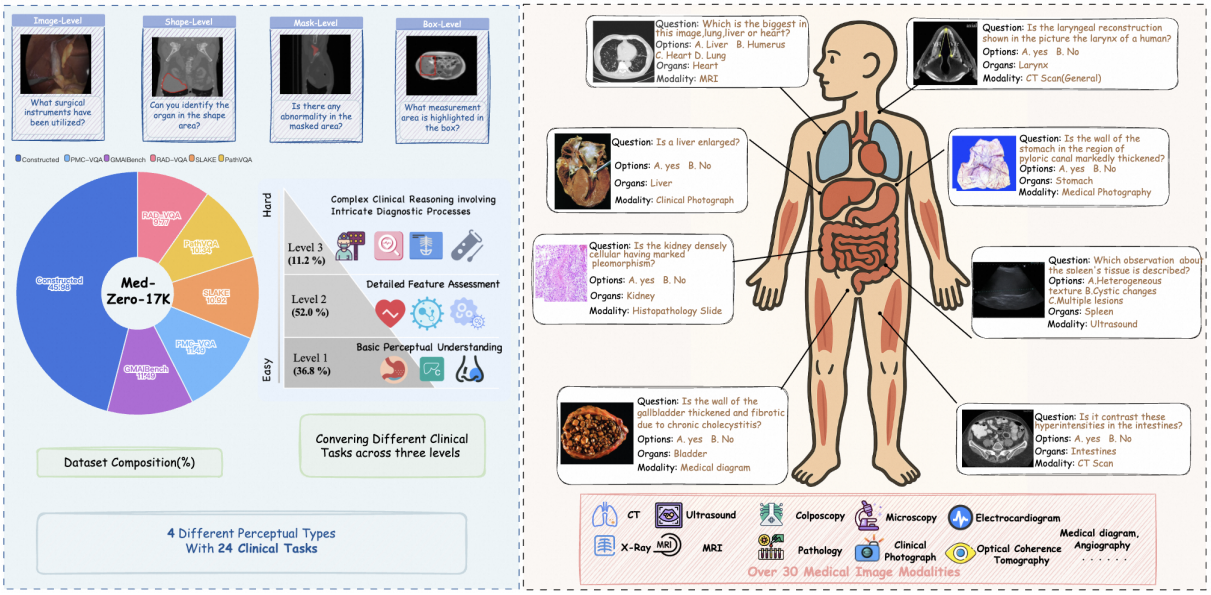

Figure 2: Overview of the Med-Zero-17K which encompasses 4 distinct granularity levels of image annotations and 24 diverse clinical tasks, covering various human organs and over 30 modalities.

2025; Jiang et al., 2025b), but these largely rely on in-domain data and struggle to generalize. For cross-modality alignment, MMed-RAG (Xia et al., 2024) constructs preference pairs using unrelated images for offline DPO-based RAG alignment. Our R_{PCR} differs in three key aspects: (1) *perturbation strategy*—we use cosine-annealed diffusion noise instead of unrelated images, creating a curriculum that prevents reward hacking (Table 16); (2) *training regime*—we operate within online GRPO rather than offline DPO, and validate compatibility with both (Appendix C.5); (3) *problem scope*—we target CoT-level visual grounding in RLVR rather than text-retrieval alignment.

2.1 Med-Zero-17K Dataset

To address the scarcity of high-quality RL datasets in the medical domain, we introduce Med-Zero-17K, a comprehensive dataset spanning 30 medical imaging modalities and 24 clinically relevant tasks. We employed Qwen2.5-VL-72B (Bai et al., 2025) to generate VQA pairs from PubMed-Vision's (Chen et al., 2024) image-caption data, significantly enhancing modality diversity. Clinical questions are categorized into three complexity levels representing progression from basic perception (Level 1) through detailed feature assessment (Level 2) to complex diagnostics (Level 3), as illustrated in Figure 2. Annotations are provided across 4 perceptual granularities: image-level, shape-level, bounding box, and segmentation masks.

To ensure high quality, we implemented a multi-stage filtering pipeline addressing: (1) image resolution and aspect ratio filtering to preserve diagnostic information, (2) aesthetic scoring to remove low-quality images, (3) diversity-ensuring sampling via k-NN clustering to balance modality representation, (4) mixed difficulty filtering retaining questions with partial correctness to avoid advantage skew, and (5) generated QA validation ensuring visual grounding. Complete filtering details are provided in Appendix B. This carefully structured dataset facilitates stable, progressive RL training, empowering models to transition from fundamental perception to sophisticated clinical reasoning.

2.2 Multi-Consistency Reward (MCR) Framework

To address the perceptual-reasoning misalignment and reasoning-answer inconsistency identified in our analysis, we propose the **Multi-Consistency Reward (MCR)** framework. MCR enhances standard RLVR by augmenting policy optimization with specialized reward signals that enforce coherence throughout the reasoning-to-answer pipeline. Importantly, MCR is compatible with RL algorithms including GRPO and DPO (Appendix C.5). Key hyperparameters are in Appendix Table 11.

Decision Accuracy Reward (R_{DAR}). To provide the primary verifiable learning signal, this component rewards correct final answers. Given a question-image pair q and model output $o = (p, a)$

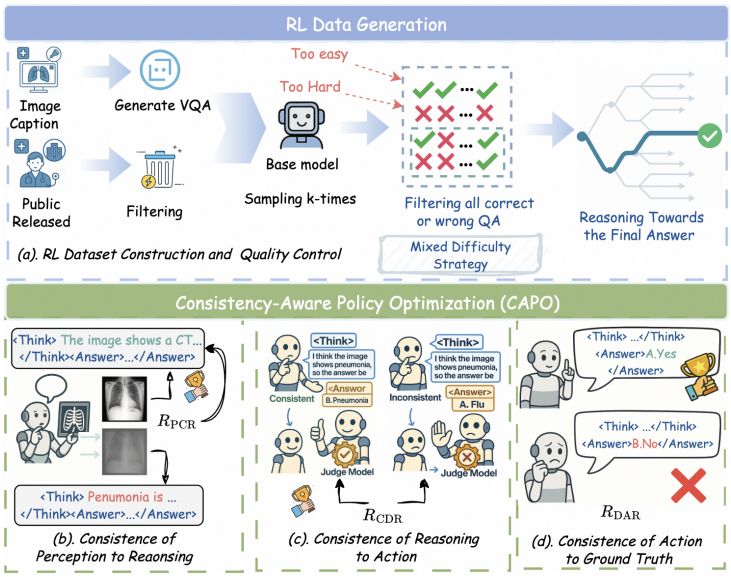

Figure 3: Overview of MCR which comprises four key components: (a) RL dataset construction using a mixed difficulty strategy, (b) R_{PCR} to align reasoning with visual inputs, (c) R_{CDR} to maintain coherence between reasoning and decisions, and (d) R_{DAR} to align decisions with correct answers.

where p is the reasoning process and a is the final answer, we define: $R_{\text{DAR}}(o) = r_{\text{dar}}$ if a matches ground truth, 0 otherwise, where $r_{\text{dar}} = 0.8$.

Cognitive-Decision Consistency Reward (R_{CDR}). To enforce logical consistency between reasoning and answers, we employ an external LLM judge to verify logical entailment—whether the answer follows from the reasoning—rather than judging reasoning correctness, which is unreliable. We use GPT-4o (prompt in Appendix 14) where $\mathcal{J}(p, a) \in \{\text{True}, \text{False}\}$ denotes the judge's assessment of whether a logically follows from p. Then $R_{\text{CDR}}(o) = r_{\text{cdr}}$ if $\mathcal{J}(p, a) = \text{True}$, 0 otherwise ($r_{\text{cdr}} = 0.1$). Table 5 shows open-source alternatives achieve comparable performance.

Perceptual-Cognitive Consistency Reward (R_{PCR}). To incentivize stronger visual feature utilization in multimodal reasoning, this reward provides a self-supervised learning signal through contrastive image pairs. Unlike prior cross-modality alignment methods that use unrelated images (Xia et al., 2024), we employ cosine-annealed diffusion corruption with timesteps decreasing from 500 to 100 (detailed in Appendix C.2), implementing a curriculum from coarse-level to fine-grained visual feature learning within online GRPO. At batch level, we compute accuracy on clean ($\text{acc}_{\text{clean}}$) and corrupted ($\text{acc}_{\text{corrupt}}$) images. Individual samples receive $r_{\text{pcr}} = 0.1$ when: (1) $\text{acc}_{\text{clean}} - \text{acc}_{\text{corrupt}} > 0.3$ at batch level, indicating the model actively uses visual features, and (2) the sample is correct on clean images. This self-supervised signal incentivizes visual evidence utilization while maintaining training stability through batch-level computation.

Policy Optimization. The total MCR reward combines all components: $R_{\text{MCR}}(o) = R_{\text{DAR}}(o) + R_{\text{PCR}}(o) + R_{\text{CDR}}(o)$. Following mainstream practices in RLVR, we integrate MCR with GRPO (Shao et al., 2024) for Med-VQA tasks. For each input q, the policy model π_θ generates G candidate outputs. We compute advantages via group-wise normalization: $A_i = (R_{\text{MCR}}(o_i) - \mu_G)/(\sigma_G + \delta)$, where μ_G and σ_G are the group mean and standard deviation. These advantages weight the PPO-clipped objective with $\epsilon = 0.2$. Complete GRPO integration details are in Appendix C.3. We also validate MCR's compatibility with DPO in Appendix C.5. By integrating MCR, the RL algorithm optimizes simultaneously for answer correctness, logical coherence, and genuine utilization of visual evidence.

3 Experiments

3.1 Experiments Setup

Evaluation Benchmarks. We evaluate MCR across six benchmarks spanning three categories, following established protocols (Chen et al., 2024; Li et al., 2024a; Jiang et al., 2024). *(1). Domain-*

Table 1: The results of the medical VQA benchmark.

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	Avg.
Gemini-2.0-flash-lite	59.4	73.1	64.9	50.8	60.5
GPT-4.1-Nano	61.8	73.1	70.6	53.1	64.5
Med-Flamingo	45.4	43.5	54.7	23.3	41.7
RadFM	50.6	34.6	38.7	25.9	37.5
LLaVA-Med-7B	51.4	48.6	56.8	24.7	45.4
Qwen-VL-Chat	47.0	56.0	55.1	36.6	48.9
Yi-VL-34B	53.0	58.9	47.3	39.5	49.7
LLaVA-v1.6-7B	52.6	57.9	47.9	35.5	48.5
LLaVA-v1.6-13B	55.8	58.9	51.9	36.6	50.8
LLaVA-v1.6-34B	58.6	67.3	59.1	44.4	57.4
LLaVA-v1.5-LLaMA3-8B	54.2	59.4	54.1	36.4	51.0
LLaVA_Med -LLaMA3-8B	60.2	61.2	54.5	46.6	55.6
PubMedVision-8B	63.8	74.5	59.9	52.7	62.7
HuatuoGPT-Vision-34B	68.1	76.9	63.5	58.2	66.7
Qwen2.5-VL-7B	70.9	72.8	65.7	54.9	66.0
Qwen2.5-VL-7B + SFT	75.2	81.0	66.9	52.2	68.8
Qwen2.5-VL-7B + MCR-GRPO	78.5	79.1	68.9	55.5	70.5
Qwen2.5-VL-7B + MCR-GRPO-R1	82.5	81.3	71.5	52.9	72.1

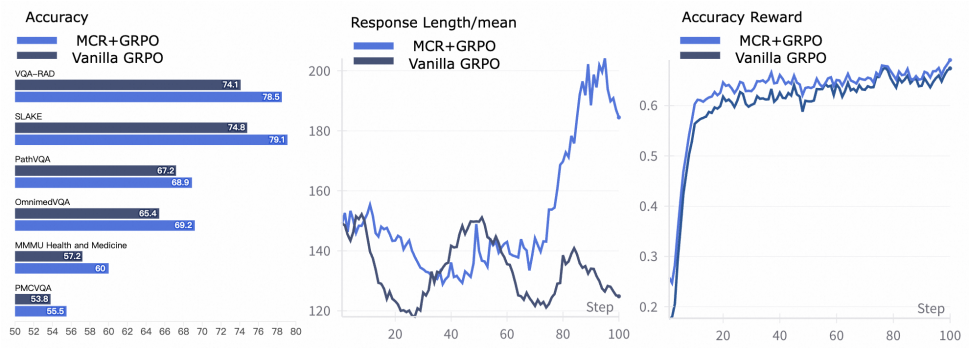

Figure 4: Comparision of MCR and GRPO. More details are in Table 15

Table 2: The accuracy of OmniMedVQA (Hu et al., 2024) within different modalities. Specifically, FP denotes *Fundus Photography*, MRI denotes *Magnetic Resonance Imaging*, CT denotes *Computed Tomography*, OCT denotes *Optical Coherence Tomography*, Der denotes *Dermoscopy*, Mic denotes *Microscopy Images*, US denotes *Ultrasound*, and X-Ray denotes *X-Ray*.

Model	CT	FP	MRI	OCT	Der	Mic	X-Ray	US	Avg.
Med-Flamingo	34.6	33.3	27.5	26.0	28.3	28.1	30.1	33.2	30.2
RadFM	33.3	35.0	22.0	31.3	36.3	28.0	31.5	26.1	30.5
LLaVA-Med-7B	25.3	48.4	35.9	42.1	45.2	44.0	31.7	83.7	44.5
Qwen-VL-Chat	51.5	45.4	43.9	54.0	55.4	49.5	63.1	33.5	49.5
Yi-VL-34B	39.8	57.2	51.4	70.5	54.5	61.4	64.2	40.5	54.9
LLaVA-v1.6-7B	40.1	39.5	54.8	58.4	54.0	48.8	53.3	47.9	49.6
LLaVA-v1.6-13B	40.0	43.6	47.4	63.2	58.0	50.5	59.6	42.6	50.6
LLaVA-v1.6-34B	50.6	63.4	60.9	68.4	65.7	62.8	74.7	44.5	61.4
LLaVA-v1.5-LLaMA3-8B	33.0	49.7	53.8	76.0	63.1	48.4	56.6	31.2	48.8
LLaVA_Med -LLaMA3-8B	60.8	68.5	66.3	79.0	66.6	60.3	73.3	49.3	65.5
PubMedVision-8B	61.6	80.2	65.1	86.3	71.6	67.4	81.4	87.4	75.1
HuatuoGPT-Vision-34B	60.8	85.5	66.5	90.0	74.0	71.3	83.8	81.7	76.7
Qwen2.5-VL-7B	63.9	73.3	68.5	74.0	67.1	73.9	74.7	33.4	66.1
Qwen2.5-VL-7B + SFT	62.1	66.3	62.5	59.6	59.2	66.6	74.6	34.4	60.7
Qwen2.5-VL-7B + MCR-GRPO	65.7	82.1	74.3	75.6	67.8	74.7	75.4	37.7	69.2

Specific Medical VQA Benchmarks. We assess standard medical VQA capabilities on Rad-VQA (Lau et al., 2018), SLAKE (Liu et al., 2021) (English CLOSED segment), PathVQA (He et al., 2020), and PMC-VQA (Zhang et al., 2023). These benchmarks evaluate question answering across radiology, pathology, and general medical imaging. *(2). Expert-Level Multimodal Reasoning.* We use

Table 3: Results on the MMMU Health & Medicine track (Yue et al., 2024). The Health & Medicine track is divided into five categories: **BMS** for *Basic Medical Science*, **CM** for *Clinical Medicine*, **DLM** for *Diagnostics and Laboratory Medicine*, **P** for *Pharmacy*, and **PH** for *Public Health*. Results are obtained by submitting to the official website.

Model	BMS	CM	DLM	P	PH	MMMU Health & Medicine
Gemini-2.0-flash-lite	56.7	66.8	43.3	66.8	60.0	58.7
GPT-4.1-Nano	63.3	60.0	43.3	63.3	73.3	60.6
Med-Flamingo	29.6	28.1	24.8	25.3	31.2	28.3
RadFM	27.5	26.8	25.8	24.7	29.1	27.0
LLaVA-Med-7B	39.9	39.1	34.6	37.4	34.0	36.9
Qwen-VL-Chat	36.5	31.7	32.7	28.4	34.6	32.7
Yi-VL-34B	49.4	48.9	43.2	40.5	32.0	41.5
LLaVA-v1.6-7B	40.5	36.9	32.1	32.3	26.9	33.1
LLaVA-v1.6-13B	53.6	46.7	33.3	22.2	40.0	39.3
LLaVA-v1.6-34B	56.4	56.0	46.9	46.7	41.7	48.8
LLaVA-v1.5-LLaMA3-8B	42.3	44.0	37.0	34.7	35.2	38.2
LLaVA_Med -LLaMA3-8B	48.2	43.8	42.0	39.7	35.8	41.1
PubMedVision-8B	61.0	58.8	50.0	44.7	38.7	49.1
HuatuoGPT-Vision-34B	64.6	62.5	50.6	54.1	44.2	54.4
Qwen2.5-VL-7B	53.6	60.0	40.0	66.7	53.3	54.7
Qwen2.5-VL-7B + SFT	57.1	60.0	26.7	55.6	60.0	51.9
Qwen2.5-VL-7B + MCR-GRPO	60.7	70.0	40.0	74.1	56.7	60.0

Table 4: MCR generalizes across diverse VLM backbones. While SFT improves in-domain performance, it degrades out-of-domain (OOD) generalization. MCR consistently achieves gains on both in-domain and OOD benchmarks.

Model	**VQA-RAD**	**SLAKE**	**PathVQA**	**PMC-VQA (OOD)**
PubMedVision-8B				
Base	63.8	74.5	59.9	52.7
+ SFT on Med-Zero-17k	70.1	79.2	63.4	45.3 ↓7.4
+ MCR-GRPO on Med-Zero-17k	**72.5**	**81.3**	**66.2**	**54.5** ↑1.8
InternVL-2.5-8B				
Base	71.7	74.0	60.4	50.9
+ SFT on Med-Zero-17k	76.5	79.8	63.1	47.3 ↓3.6
+ MCR-GRPO on Med-Zero-17k	**78.1**	**80.3**	**64.6**	**52.1** ↑1.2

the Health & Medicine track of MMMU (Yue et al., 2024), which contains real clinical questions requiring professional-level expertise. Since MMMU has no overlap with our training data, it serves as a challenging out-of-distribution test. *(3). Diverse Modality Understanding.* We evaluate on OmniMedVQA (Hu et al., 2024), which aggregates 42 traditional medical imaging datasets across eight modalities (CT, MRI, X-ray, ultrasound, etc.), enabling comprehensive assessment of cross-modality generalization.

Baselines. We compare against two categories of models. *(1). Medical-specialist VLMs* include HuatuoGPT-Vision-34B (Chen et al., 2024), Med-Flamingo (Moor et al., 2023), and RadFM (Wu et al., 2023). *(2). General-purpose VLMs* include LLaVA-v1.6-34B (Li et al., 2024b), Yi-VL-34B (Young et al., 2024), and Qwen2.5-VL-72B (Bai et al., 2025), representing current state-of-the-art vision-language models.

3.2 Main Results

Integration with GRPO. Integrating MCR into the GRPO framework significantly improves training dynamics compared to the vanilla implementation, as shown in Figure 4. Our MCR-GRPO not only achieves higher rewards with greater stability but also fosters a key emergent property: *deeper and more stable reasoning*. Specifically, while vanilla GRPO exhibits unpredictable fluctuations in response length, our method promotes a steady increase throughout training, indicating more stable and sustained reasoning processes. This suggests that our consistency-oriented rewards do not just improve accuracy, but enhance the model's cognitive depth by encouraging more thorough and elaborate thought processes. Importantly, this increased length is not over-verbosity: validation on challenging samples (with $\leq$0.25 base accuracy)

confirms continued improvement in both top-k success rate and reward, indicating the model learns to tackle difficult problems through deeper reasoning.

Consistent In-Domain Performance with Superior Stability. MCR achieves 78.5% on VQA-RAD and 68.9% on PathVQA using Qwen2.5-VL-7B (Table 1), outperforming both medical-specialist models (e.g., HuatuoGPT-Vision-34B: 68.1%) and general VLMs. More critically, MCR demonstrates superior training stability compared to supervised fine-tuning. While SFT achieves strong in-domain gains (75.2% on VQA-RAD), it suffers severe overfitting on PMC-VQA (-2.7 points from baseline). MCR maintains consistent improvements across all benchmarks, including +0.6 on PMC-VQA, indicating that consistency-based rewards provide more robust learning signals than supervised objectives alone.

Out-of-Distribution Generalization. On OmnimedVQA and MMMU Health & Medicine—two benchmarks with zero training overlap—MCR shows substantial generalization advantages. MCR achieves 69.2% average accuracy on OmnimedVQA versus 60.7% for SFT (Table 2), with particularly strong gains on challenging modalities: fundus photography (+15.8%), MRI (+11.8%), and CT (+3.6%). On the expert-level MMMU benchmark (Table 3), MCR reaches 60.0% overall and 70.0% on Clinical Medicine, surpassing SFT by 8.1 points. This OOD superiority reveals a key insight: explicit consistency constraints prevent models from exploiting dataset-specific shortcuts, forcing genuine multimodal reasoning that transfers across domains.

Cross-Architecture Generalization. We validate MCR across three diverse architectures: general-purpose InternVL-2.5-8B and medical-specialist PubMedVision-8B (Table 4). A striking pattern emerges: while SFT uniformly degrades OOD performance (-7.4, -3.6 on PMC-VQA), MCR maintains gains (+1.8, +1.2). Medical-specialist backbones show particularly strong synergy with MCR (PubMedVision: +8.7 on VQA-RAD), suggesting domain-specific pretraining provides richer representations for consistency learning.

3D VQA Performance Analysis To test dimensional generalization, we evaluate on 3D medical imaging using M3D (Bai et al., 2024a). MCR achieves 34.57% accuracy without any 3D-specific training, outperforming SFT (28.33%) and vanilla GRPO (32.16%) by substantial margins (Figure 5 a). This demonstrates MCR's consistency principles extend beyond 2D images to volumetric data.

3.3 Ablation Study

Ablation of Reward Components. We systematically evaluate each reward component in Table 7. Using only the accuracy reward (R_{DAR}) as a baseline provides solid in-domain results but limited OOD generalization. Adding either R_{CDR} (for logical coherence) or R_{PCR} (to anchor reasoning in visual evidence) individually delivers significant gains across all benchmarks. The full MCR framework, which combines all three, achieves the strongest performance, particularly on OOD tasks. This indicates the reward components are synergistic: R_{CDR} enforces a logical flow from reasoning to answer, while R_{PCR} ensures the reasoning itself is tied to the visual input, validating our multi-faceted approach.

Ablation of Diffusion Timesteps. Our cosine-annealed schedule (500→100) for image corruption optimally balances exploration and refinement, outperforming the alternatives shown in Table 17. A less aggressive schedule (500→300) fails to push the model towards fine-grained analysis in later stages, while an overly strong one (800→100) can disrupt early-stage learning. This confirms our curriculum-based approach is most effective for teaching the model to rely on visual features.

Ablation of Noise Types. As shown in Table 16, adaptive diffusion noise surpasses simpler methods like masking or cropping. Its naturalistic corruption better tests genuine visual understanding without introducing artificial artifacts. This encourages more robust visual processing, leading to superior generalization, especially on OOD benchmarks where the performance gap is most pronounced.

Ablation of Reward Ratios. Table 8 shows our 8:1:1 ratio for R_{DAR}:R_{PCR}:R_{CDR} is optimal. The results confirm that while the accuracy reward must be dominant to prioritize correctness, our consistency rewards provide a crucial complementary signal. Notably, MCR demonstrates robust performance across ratios from 5:1:1 to 8:1:1, indicating the framework is stable and does not require extensive hyperparameter tuning.

Ablation of Threshold τ_{pcr}. A value of 0.3 for the activation threshold τ_{pcr} strikes the best balance,

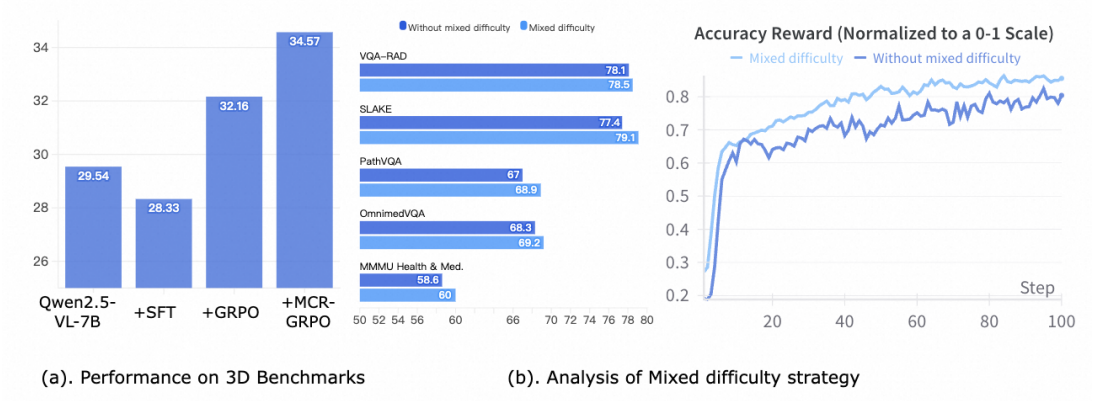

Figure 5: (a) Performance on 3D benchmarks.

Table 5: Comparison of different judge models for R_{CDR}.

Judge Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OmnimedVQA	MMMU Health
Qwen3-30B-A3B	79.3	78.5	69.4	55.2	69.0	59.3
VOmniMed-1.5B	78.1	79.8	68.5	55.6	68.9	60.7
GPT-4o (Original)	78.5	79.1	68.9	55.5	69.2	60.0

as seen in Table 18. Lower thresholds are susceptible to noise from random performance fluctuations, leading to unstable rewards. Conversely, higher thresholds are too restrictive and filter out many valid training signals. Our chosen value ensures rewards are assigned only upon clear evidence that the model's reasoning is visually-grounded.

Ablation of Med-Zero-17K Composition. Figure 6 ablates the composition of Med-Zero-17K. While both our curated 8K VQA pairs and the 9K pairs from public datasets improve in-domain performance, their contributions to out-of-distribution (OOD) generalization diverge sharply. The public data, with its narrow modality coverage, offers minimal OOD gains. In contrast, our curated data, spanning 30 modalities and 24 clinical tasks, drives substantial OOD improvements (e.g., +5.6 on OmnimedVQA). This confirms that the diversity of our curated data is the key driver of MCR's strong cross-domain generalization.

3.4 Further Analysis

MCR vs. Judging Correctness: Consistency is Key. To dissect MCR's core mechanism, we compared rewarding consistency against rewarding absolute reasoning correctness. For the latter, we used Gemini2.5-Flash as an external judge. As shown in Table 6, directly rewarding correctness failed, often degrading performance. This provides a critical insight: judging the correctness of complex medical reasoning is fraught with noise and bias, which destabilizes RL training. In stark contrast, MCR's strategy of rewarding logical coherence provides a stable learning signal, delivering substantial gains across all benchmarks. This comparison powerfully validates our hypothesis: *enforcing internal consistency is a more robust and practical learning signal than pursuing correctness with an imperfect external oracle*.

Why Diffusion Noise for R_{PCR}? A key design choice in MCR is the use of cosine-annealed diffusion noise for image corruption, rather than simpler strategies. Table 16 compares our approach against masking, cropping, and fixed-step diffusion. Simple corruptions (mask, crop) are too easily distinguished by the model in the online GRPO setting, leading to reward hacking rather than genuine visual grounding. Fixed diffusion noise lacks the curriculum effect. Our cosine-annealed schedule provides naturalistic, gradually refined corruption that forces robust visual reasoning, yielding the strongest results especially on OOD benchmarks.

In-Domain vs. OOD Trade-off in R1 Paradigms. A revealing pattern emerges from comparing MCR-GRPO-R0 and MCR-GRPO-R1 (Table 1). While R1 yields strong in-domain scores through cold-start SFT, it uniformly degrades OOD performance. Notably, SFT alone already causes substantial OOD drops (PMC-VQA: -2.7, OmnimedVQA: -5.4, MMMU: -2.8 from base), and subsequent RL training cannot fully recover this loss. This

Table 6: Comparison of MCR with an alternative reward strategy based on judging reasoning correctness.

Model	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OmnimedVQA	MMMU Health
Qwen2.5-VL-7B (Base)	70.9	72.8	65.7	54.9	66.1	54.7
+ SFT	75.3	80.9	66.9	52.2	60.7	51.9
+ GRPO (Vanilla)	74.1	74.8	67.2	53.8	65.4	57.2
+ Judge Correctness	74.5 (+0.4)	74.0 (-0.8)	66.6 (-0.6)	53.1 (-0.7)	63.9 (-1.5)	56.0 (-1.2)
+ MCR	**78.5 (+4.4)**	**79.1 (+4.3)**	**68.9 (+1.7)**	**55.5 (+1.7)**	**69.2 (+3.8)**	**60.0 (+2.8)**

Table 7: Ablation Study of Reward Components.

R_{PCR}	R_{CDR}	R_{DAR}	RAD	SLAKE	PVQA	OmniMed
✗	✗	✗	70.9	72.8	65.7	66.1
✗	✗	✓	74.1	74.8	67.2	65.4
✗	✓	✓	76.9	76.7	68.1	68.1
✓	✗	✓	76.5	75.5	67.5	68.4
✓	✓	✓	**78.5**	**79.1**	**68.9**	**69.2**

Table 8: Sensitivity analysis of reward weights.

Ratio	RAD	SLAKE	PVQA	OmniMed
1:1:1	71.7	70.9	65.4	63.2
2:1:1	72.5	73.3	67.2	64.5
5:1:1	77.7	78.7	68.1	67.7
8:1:1 (Ours)	**78.5**	**79.1**	**68.9**	**69.2**

is corroborated by our Pass@k analysis: the R0 paradigm achieves greater accuracy gains across VQA datasets, indicating it better preserves the base model's exploratory capacity. These findings suggest that for medical RLVR, where OOD generalization is critical for clinical safety, the R0 paradigm paired with MCR offers a more robust path than the R1 approach.

Effect of Mixed Difficulty Strategy. To investigate the learning dynamics of RL training, we evaluated the effect of incorporating samples with varying difficulty levels. As shown in Figure 5, our mixed difficulty strategy leads to consistent performance gains across all benchmarks. This improvement stems from alleviating "advantage bias", where models trained on uniformly easy or hard samples tend to produce skewed advantage estimates, leading to unstable or suboptimal policy updates. We further compared our approach against curriculum RL (easy-to-hard) training, as shown in Table 9. While the easy-to-hard strategy achieves reasonable performance, it consistently underperforms our mixed difficulty filtering across both in-domain and OOD benchmarks. This stems from a fundamental GRPO limitation: in the easy-to-hard setting, the model may answer all easy samples correctly or all hard samples incorrectly within a rollout group, leading to empty within-group advantages that degrade training efficiency.

Table 9: Comparison of data difficulty strategies.

Strategy	RAD	SLAKE	PVQA	MMMU
Without Mixed	78.1	77.4	67.0	58.6
Easy-to-Hard	77.7	78.5	67.6	58.6
Mixed (Ours)	**78.5**	**79.1**	**68.9**	**60.0**

Are R1-Like Paradigms Effective in the Medical Domain? Although R1-like paradigms have proven successful in mathematical reasoning—particularly by cold-starting with CoT data during SFT before RL—their effectiveness in the medical domain remains unclear. To explore this, we used Qwen2.5-VL-72B for rejection sampling to collect high-quality medical CoT data for cold-start SFT (Details in Appendix 8). The detailed in-domain vs. OOD trade-off analysis is presented above.

4 Conclusion

In this paper, we propose a new RL framework MCR for Med-VQA that addresses key challenges in aligning perception, reasoning, and answer generation. To support this, we introduce Med-Zero-17K, a diverse dataset spanning over 30 medical modalities and 24 clinical tasks. Experiments on in-domain, out-of-domain, and 3D benchmarks show that MCR outperforms strong baselines, enabling more consistent and generalizable medical reasoning. Our work demonstrates the potential of consistency-aware RL in advancing medical AI.

Acknowledgement

This work is supported by the National Key R&D Program of China (Grant No. 2024YFC3308304), the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (Grant no. 2025C01128), and the ZJU-Angelalign R&D Center for Intelligence Healthcare.

A Limitations

While the proposed approach demonstrates strong generalization across diverse medical VQA tasks, its effectiveness is still limited by the scope of publicly available medical datasets. Although we construct the Med-Zero-17K dataset to improve diversity, existing data sources still lack sufficient coverage of rare diseases, underrepresented modalities, and fine-grained reasoning annotations. This limitation may restrict the model's ability to learn more comprehensive and complex clinical reasoning patterns. Moreover, potential risks remain in real-world deployment. Since real clinical data are often more diverse and noisy than benchmark datasets, the model may produce unreliable predictions or clinically inaccurate explanations when facing unseen cases or low-quality inputs. Therefore, this system should be considered an assistive tool rather than a standalone diagnostic solution.

References

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. 2024a. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Samar Bouazizi and Hela Ltifi. 2024. Enhancing accuracy and interpretability in eeg-based medical decision making using an explainable ensemble learning framework application for stroke prediction. *Decision Support Systems*, 178:114126.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. 2024. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7346–7370.

Xinyun Chen, Maxwell Lin, Nathanael Scharli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Wenjie Dong, Shuhao Shen, Yuqiang Han, Tao Tan, Jian Wu, and Hongxia Xu. 2025. Generative models in medical visual question answering: A survey. *Applied Sciences*, 15(6):2983.

Xiaotang Gai, Chenyi Zhou, Jiaxiang Liu, Yang Feng, Jian Wu, and Zuozhu Liu. 2024. Medthink: Explaining medical visual question answering via multimodal decision-making rationale. *arXiv preprint arXiv:2404.12372*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Songtao Jiang, Sibo Song, Chenyi Zhou, Yuan Wang, Ruizhe Chen, Tongkun Guan, Ruilin Luo, Yan Zhang, Zhihang Tang, Yuchong Sun, et al. 2026. Learning transferable temporal primitives for video reasoning via synthetic videos. *arXiv preprint arXiv:2603.17693*.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. 2025a. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*.

Songtao Jiang, Yan Zhang, Yeying Jin, Zhihang Tang, Yangyang Wu, Yang Feng, Jian Wu, and Zuozhu Liu. 2025b. Hscr: Hierarchical self-contrastive rewarding for aligning medical vision language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13853–13868.

Songtao Jiang, Tuo Zheng, Yan Zhang, Yeying Jin, Li Yuan, and Zuozhu Liu. 2024. Med-moe: Mixture of domain-specific experts for lightweight medical vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3843–3860.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452.

Yuxiang Lai, Jike Zhong, Ming Li, Shitian Zhao, Yuheng Li, Konstantinos Psounis, and Xiaofeng Yang. 2026. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *IEEE Transactions on Medical Imaging*.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-interleave: Tackling multi-image, video, and 3d in large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE.

Jiaxiang Liu, Yuan Wang, Jiawei Du, Joey Tianyi Zhou, and Zuozhu Liu. 2024. Medcot: Medical chain of thought via hierarchical expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17371–17389.

Xiangyan Liu et al. 2025a. Noisyrollout: Reinforcing visual reasoning with data augmentation. *arXiv preprint*.

Xiaohong Liu, Hao Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, et al. 2025b. A generalist medical language model for disease diagnosis assistance. *Nature medicine*, 31(3):932–942.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025c. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*.

Ruilin Luo, Zhuofan Zheng, Yifan Wang, Yiyao Yu, Xinzhe Ni, Zicheng Lin, Jin Zeng, and Yujiu Yang. 2025. Ursa: Understanding and verifying chain-of-thought reasoning in multimodal mathematics. *arXiv e-prints*, pages arXiv–2501.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. 2025. 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*, pages 337–347. Springer.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*.

Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Zou, and Huaxiu Yao. 2024. Mmed-rag: Versatile multimodal rag system for medical vision language models. *arXiv preprint arXiv:2410.13085*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. 2025a. Medrlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Xiaotian Zhang, Ruizhe Chen, Yang Feng, and Zuozhu Liu. 2025b. Persona-judge: Personalized alignment of large language models via token-level self-judgment. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5037–5049.

Xiaotian Zhang, Yuan Wang, Zhaopeng Feng, Ruizhe Chen, Zhijie Zhou, Yan Zhang, Hongxia Xu, Jian Wu, and Zuozhu Liu. 2025c. Med-u1: Incentivizing unified medical reasoning in llms via large-scale reinforcement learning. *arXiv preprint arXiv:2506.12307*.

A Appendix

B Med-Zero-17K Dataset Details

B.1 Quality Control and Filtering Pipeline

We implemented a rigorous, multi-stage filtering pipeline to ensure high quality and utility of our RL training data, designed to enhance training stability, data diversity, and task relevance.

Image Resolution and Aspect Ratio Filtering We begin by filtering out images with extreme aspect ratios. Such images often suffer from significant distortion when resized to the fixed input dimensions of the vision encoder, which can lead to the loss of critical diagnostic information and impede model learning.

Aesthetic and Quality Score Filtering Following established practices, we employ an aesthetic scoring model to discard visually poor images. This step removes inputs with low diagnostic value, such as blurry images, those with severe artifacts, or poor lighting, ensuring the model learns from clear, high-quality medical evidence.

Diversity-Ensuring Sampling To prevent the dataset from being dominated by common modalities (e.g., chest X-rays), we enforce diversity by clustering images based on their visual features extracted from the base VLM's ViT. By applying k-NN clustering and then sampling a limited number of images from each cluster, we ensure `Med-Zero-17K` maintains a broad representation of modalities.

Mixed Difficulty Filtering We employ a mixed difficulty sampling strategy to create a balanced and effective training curriculum. For each question, we generate 10 candidate responses and retain only those question-answer pairs where the model exhibits partial correctness (i.e., not all responses are correct, but at least one is). This strategy avoids advantage skew by filtering out samples that are either too easy or too hard, focusing the training on questions that are within the model's learning frontier.

Generated QA Validation For all VQA pairs generated from image-caption data, we perform a final validation step. We verify that each question can be answered using only the information present in the source caption. This sanity check ensures that the task is visually grounded and prevents data leakage where the model might learn to answer questions based on the LLM's parametric knowledge rather than the visual input.

C MCR Framework Implementation Details

C.1 Judge Model Implementation

Directly judging the correctness of medical reasoning requires expert-level domain knowledge and often leads to noisy reward signals. Instead, we focus on verifying logical entailment—whether the answer follows from the reasoning—which is more reliable and does not require medical expertise.

We employ GPT-4o with the prompt template shown in Table 14. The judge receives the question q, reasoning p, and answer a, returning a binary judgment on whether a logically follows from p. Table 5 demonstrates that open-source alternatives including Qwen3-30B and VOmniMed-1.5B achieve comparable performance, making the framework accessible without requiring expensive API calls.

C.2 Cosine-Annealed Image Corruption

To encourage visual feature utilization, we employ a self-supervised signal based on performance contrast between clean and corrupted images. The corruption intensity must be carefully controlled: too strong renders images uninformative; too weak allows shortcuts to remain effective. We address this through curriculum learning using cosine annealing.

Early in training, stronger corruption increases exploration diversity and amplifies performance

gaps between strategies that genuinely use visual features versus those exploiting shortcuts. As training progresses, we reduce corruption to preserve diagnostic details and refine attention to specific visual features. The corruption timestep decreases from $t_{\text{init}} = 500$ to $t_{\text{final}} = 100$ following:

$$t_e = t_{\text{final}} + \frac{1}{2}(t_{\text{init}} - t_{\text{final}}) \left(1 + \cos\left(\frac{\pi \cdot e}{E}\right)\right) \quad (1)$$

where e is the current epoch. The noise level follows the standard cosine schedule:

$$\bar{\alpha}_t = \cos^2\left(\frac{t/T_{\max} + s}{1+s} \cdot \frac{\pi}{2}\right) \quad (2)$$

with $s = 0.008$ and $T_{\max} = 500$. Corrupted images are generated as $\tilde{x} = \sqrt{\bar{\alpha}_{t_e}} x_0 + \sqrt{1 - \bar{\alpha}_{t_e}} \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$. Table 16 validates that this adaptive approach outperforms fixed-step corruption.

C.3 GRPO Integration Details

We integrate MCR with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which avoids training a separate value function by computing advantages through group-wise normalization. For each input q, the policy π_θ generates $G = 8$ candidate outputs. The advantage for output o_i is:

$$A_i = \frac{R_{\text{MCR}}(o_i) - \mu_G}{\sigma_G + \delta} \quad (3)$$

where $\mu_G = \frac{1}{G}\sum_{j=1}^{G} R_{\text{MCR}}(o_j)$, $\sigma_G = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(R_{\text{MCR}}(o_j) - \mu_G)^2}$, and $\delta = 1 \times 10^{-8}$. This normalization reduces variance and eliminates critic training overhead.

The advantage weights the PPO-clipped objective:

$$\mathcal{L}_{\text{CLIP}}(\theta) = \mathbb{E}\left[\min\left(\rho_i A_i, \text{clip}(\rho_i, 1-\epsilon, 1+\epsilon) A_i\right)\right] \quad (4)$$

where $\rho_i = \pi_\theta(o_i|q) / \pi_{\theta_{\text{old}}}(o_i|q)$ and $\epsilon = 0.2$.

C.4 Training Parameters

We trained our model using 8 NVIDIA A100 GPU. Each rollout worker used a batch size of 128, with a global batch size of 128. The micro-batch sizes were set to 4 for policy updates and 16 for experience collection. The KL loss coefficient was 1e-2, and entropy regularization was set to 1e-3 to encourage exploration. We used a maximum prompt length of 25,600 tokens and a maximum response length of 4,096 tokens to accommodate long clinical contexts. Training was conducted with Ray and vLLM backend, leveraging the Qwen2.5-VL-7B model as the base.

C.5 DPO Integration

To validate that MCR's principles are algorithm-agnostic, we integrate them into the Direct Preference Optimization (DPO) framework (Rafailov et al., 2024). We leverage the core ideas of MCR to construct specific preference pairs $(o_{\text{chosen}}, o_{\text{rejected}})$ that explicitly teach the model to avoid perceptual and reasoning inconsistencies.

Preference Pair Construction. For each question, a single high-quality *chosen* response is paired with two distinct types of *rejected* responses, each targeting a specific failure mode.

1. **Chosen Response Generation:** We use a powerful teacher model (Qwen2.5-VL-72B) with rejection sampling. Specifically, we generate up to 5 responses and select the first one that yields the correct final answer. This complete CoT and answer becomes the 'chosen' response, o_{chosen}.

2. **Rejected Response for Perceptual Inconsistency:** To teach the model to rely on visual evidence, we first corrupt the clean image by applying diffusion noise with a randomly sampled intensity from timesteps 100-500. We then prompt the base model with this corrupted image, generating 5 candidate responses. Finally, we select a response that leads to an incorrect final answer to serve as the 'rejected' response, $o_{\text{rejected, perceptual}}$.

3. **Rejected Response for Reasoning-Answer Inconsistency:** To enforce logical coherence, we synthesize a 'rejected' response using the base model. This is done in two ways: (1) pairing the high-quality reasoning from o_{chosen} with an incorrect final answer, or (2) pairing a flawed reasoning chain—taken from one of the rejected candidates in the initial sampling step—with the correct final answer. This synthesized response serves as the second 'rejected' response, $o_{\text{rejected, reasoning}}$.

Multi-Objective DPO Training. With one 'chosen' response and two distinct types of 'rejected' responses, we formulate a multi-objective DPO loss. This allows the model to learn from both types of negative examples simultaneously for each positive

example. The loss is defined as:

$$\mathcal{L}_{\text{MCR-DPO}} = -\mathbb{E}\left[\log\sigma(\beta\log\frac{\pi_\theta(o_{\text{chosen}})}{\pi_{\text{ref}}(o_{\text{chosen}})} - \beta\log\frac{\pi_\theta(o_{\text{rej,p}})}{\pi_{\text{ref}}(o_{\text{rej,p}})})\right] - \mathbb{E}\left[\log\sigma(\beta\log\frac{\pi_\theta(o_{\text{chosen}})}{\pi_{\text{ref}}(o_{\text{chosen}})} - \beta\log\frac{\pi_\theta(o_{\text{rej,r}})}{\pi_{\text{ref}}(o_{\text{rej,r}})})\right] \tag{5}$$

where $o_{\text{rej, p}}$ and $o_{\text{rej, r}}$ are the perceptually inconsistent and reasoning-inconsistent rejected responses, respectively. This objective simultaneously teaches the model to improve both its connection to visual evidence and its internal logical coherence within the DPO framework.

Results. We evaluated the effectiveness of MCR-DPO on the benchmarks presented in Table 10. For a strong baseline, we followed the methodology of POVID, which synthesizes rejected responses by using an external model (GPT-4o) to introduce hallucinations into the chosen response. As the results show, our MCR-DPO approach consistently outperforms the POVID-style baseline across all benchmarks. More critically, this comparison highlights a potential risk of the POVID approach: by focusing only on hallucinations, it can inadvertently degrade performance on certain benchmarks, particularly PMC-VQA and OmnimedVQA. In contrast, our method demonstrates a more robust profile. It significantly mitigates the performance degradation on PMC-VQA and, notably, turns a performance loss into a gain on OmnimedVQA.

C.6 Hyperparameters Summary

Table 11 summarizes key MCR hyperparameters.

C.7 CoT Reasoning for Cold Start Training Construction

In this study, we explore the construction of cold start training using CoT data, aiming to enhance the reasoning capabilities of models in Med-VQA tasks. Figure 8 illustrates the detailed steps of this process. Starting with an original set of medical images, a visual question is posed, such as "What visual observation can be made in this picture?" We then utilize the pre-trained Qwen2.5-VL-72B model to generate multiple reasoning paths. These paths include detailed analysis of the image and logical reasoning to reach a final answer. For instance, the model might identify irregularities in the joint space, thinning of the cartilage, signs of inflammation in the bone structure, misalignment of the vertebrae, or irregular bone growth. After generating several reasoning paths, we select the most plausible ones through a filtering strategy. This involves two main filtering processes: first, we select reasoning paths that align with the correct answers; second, we further filter to select challenging questions that require more complex logic and deeper analysis during the reasoning process. Ultimately, we extract a high-quality question set from the selected reasoning paths for subsequent reinforcement learning training. This process ensures that the model is exposed to high-quality reasoning examples from the outset and enhances the model's generalization capabilities through diverse reasoning paths.

C.8 Broader Impacts

Our work contributes to the advancement of medical AI by introducing a scalable reinforcement learning framework and a high-quality dataset tailored for medical visual question answering. By improving consistency and generalization in clinical reasoning, MCR has the potential to assist healthcare professionals in diagnostic decision-making, particularly in resource-limited settings. However, we acknowledge that the current reliance on publicly available datasets may limit representation across demographic groups, rare diseases, and global health contexts. Additionally, while our system is not intended for direct clinical use, inappropriate deployment without expert oversight could pose risks. We encourage future work to incorporate fairness, transparency, and real-world validation in collaboration with clinical stakeholders to ensure safe and equitable deployment.

Table 10: Comparison of MCR-DPO with a POVID-style baseline. While both methods improve upon the base model on some benchmarks, MCR-DPO provides more consistent gains and mitigates the performance degradation seen with the POVID approach on benchmarks like PMC-VQA and OmnimedVQA.

Methods	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OmnimedVQA
Qwen2.5-VL-7B (Base)	70.9	72.8	65.7	**54.9**	66.1
+ POVID	71.7	73.1	64.8	53.2	64.9
+ MCR-DPO (Ours)	**73.3**	**73.4**	**65.9**	54.7	**66.8**

Table 11: MCR Framework Hyperparameters

Parameter	Value
Decision Accuracy Reward (r_{dar})	0.8
Cognitive-Decision Reward (r_{cdr})	0.1
Perceptual-Cognitive Reward (r_{pcr})	0.1
PCR Threshold (τ_{pcr})	0.3
Initial Corruption Timestep (t_{init})	500
Final Corruption Timestep (t_{final})	100
Diffusion Schedule Offset (s)	0.008
Max Diffusion Steps (T_{max})	500
Group Size (G)	8
PPO Clipping (ϵ)	0.2
Stability Constant (δ)	1×10^{-8}

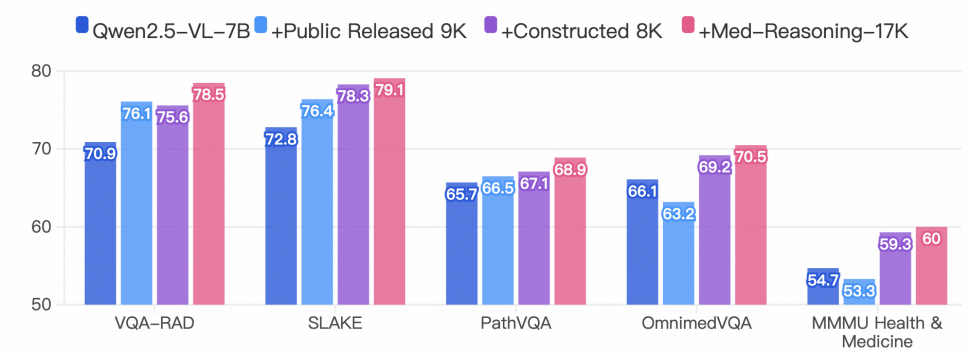

Figure 6: Effect of different composition of Med-Zero-17K.

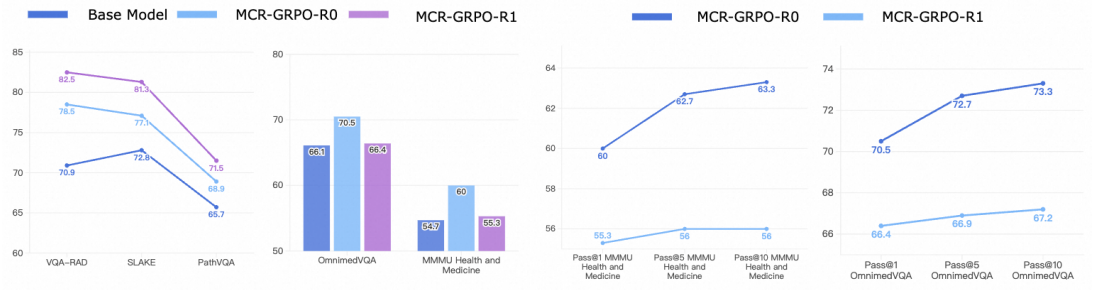

Figure 7: Comparision of Different RL Training Paradigms

Table 12: Ablation Study of Key Components on Various VQA Datasets (Full Version)

R_{PCR}	R_{CDR}	R_{DAR}	VQA-RAD	SLAKE	PathVQA	OmnimedVQA	MMMU Health & Medicine
✗	✗	✗	70.9	72.8	65.7	66.1	54.7
✗	✗	✓	74.1	74.8	67.2	65.4	57.2
✗	✓	✓	76.9	76.7	68.1	68.1	59.7
✓	✗	✓	76.5	75.5	67.5	68.4	58.4
✓	✓	✓	**78.5**	**79.1**	**68.9**	**69.2**	**60.0**

Table 13: Sensitivity analysis of reward component weights (Full Version, $R_{\text{DAR}} : R_{\text{PCR}} : R_{\text{CDR}}$). The accuracy reward requires dominant weight, but MCR is robust across ratios from 5:1:1 to 8:1:1.

Ratio	VQA-RAD	SLAKE	PathVQA	PMC-VQA	OmnimedVQA	MMMU Health
1:1:1	71.7	70.9	65.4	51.2	63.2	56.7
2:1:1	72.5	73.3	67.2	52.5	64.5	57.3
5:1:1	77.7	78.7	68.1	54.6	67.7	59.3
8:1:1 (Ours)	**78.5**	**79.1**	**68.9**	**55.5**	**69.2**	**60.0**

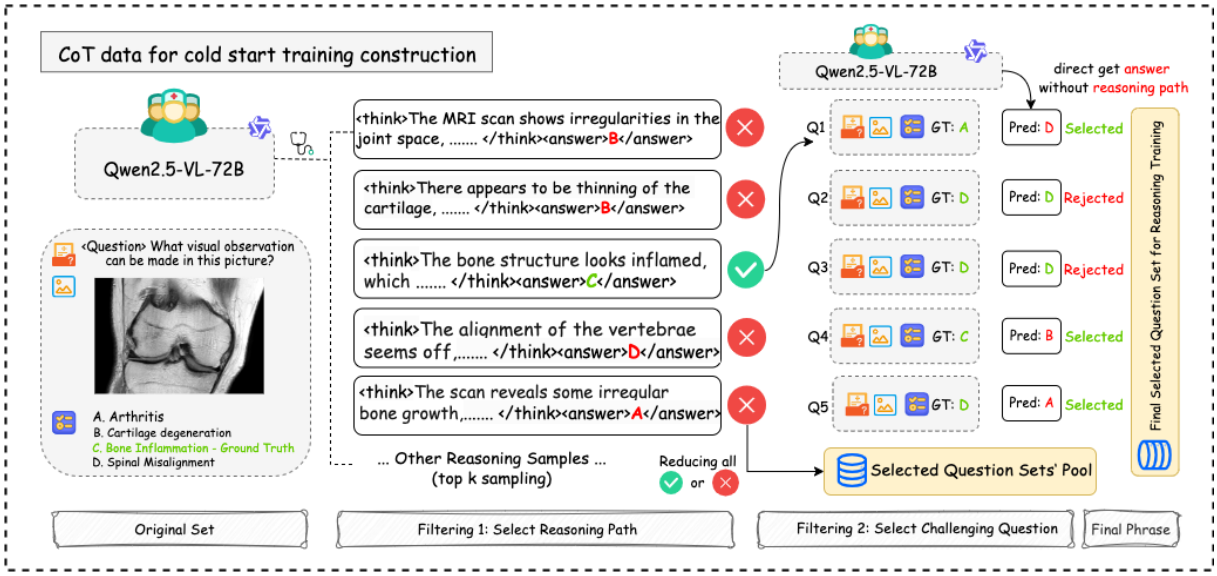

Figure 8: CoT Cold Start Reasoning Data Construction

Table 14: Prompt Template for Consistency Review

Prompt Section	**Content / Instruction**
Initial Request	Please review the "Think" (thought process) and "Answer" provided below. Referring to the "Question" for context, determine if the "Think" and "Answer" are consistent.
Definition	"Consistent" means: The logical reasoning in the "Think" process can reasonably lead to the "Answer", and the "Answer" aligns with the final conclusion of the "Think" process.

Input Placeholders	`Question: {question}`
	`Think: {think}`
	`Answer: {answer}`

Output Logic	If they are consistent, please answer: `yes` If they are inconsistent (e.g., the conclusion of the "Think" process contradicts the "Answer", or the "Answer" is not derived from the "Think" process), please answer: `no`
Final Instruction	Now output your judgement with `yes` or `no` directly:

Table 15: Direct comparison between MCR and vanilla GRPO. MCR consistently outperforms GRPO, especially on diverse modality benchmarks like OmnimedVQA and MMMU Health.

Method	**VQA-RAD**	**SLAKE**	**PathVQA**	**PMC-VQA**	**OmnimedVQA**	**MMMU Health**
Qwen2.5-VL-7B (Base)	70.9	72.8	65.7	54.9	66.1	54.7
+ SFT	75.2	81.0	66.9	52.2	60.7	51.9
+ GRPO (Vanilla)	74.1	74.8	67.2	53.8	65.4	57.2
+ MCR (Ours)	**78.5**	**79.1**	**68.9**	**55.5**	**69.2**	**60.0**
Improvement over GRPO	+4.4	+4.3	+1.7	+1.7	+3.8	+2.8

Table 16: Ablation on Noise Type

Noise Type	**RADVQA**	**SLAKE**	**PathVQA**	**OmnimedVQA**
Cosine Diffusion (Ours)	78.5	77.1	68.9	69.2
Mask	76.1	74.5	66.9	67.3
Crop	77.7	76.2	68.4	67.9
Fixed Diffusion	77.3	76.7	68.2	68.1

Table 17: Ablation on Diffusion Steps

Diffusion Steps	**RADVQA**	**SLAKE**	**PathVQA**	**OmnimedVQA**
500→100 (Ours)	78.5	77.1	68.9	69.2
baseline	70.9	72.8	65.7	66.1
500→300	77.2	75.5	68.3	67.9
800→100	77.7	75.9	68.1	68.3

Table 18: Ablation on Threshold τ_{pcr}

Threshold τ_{pcr}	**RADVQA**	**SLAKE**	**PathVQA**	**OmnimedVQA**
0.3 (Ours)	78.5	77.1	68.9	69.2
0.1	76.5	75.2	66.6	67.4
0.5	75.7	76.0	68.2	67.6

Table 19: Clinical Task Composition of Med-Zero-17K

Clinical Task Checklist		
Disease Diagnosis	Severity Grading	Organ Recognition Abdomen
Surgical Instrument Recognition	Counting	Bone
Organ Recognition Thorax	Organ Recognition Neck	Blood Vessels Recognition
Microorganism Recognition	Attribute Recognition	Cell Recognition
Surgeon Action Recognition	Organ Recognition Pelvic	Surgical Workflow Recognition
Image Quality Grading	Muscle	Nervous Tissue
Tumor Detection	Lesion Localization	Medical Image Segmentation
Disease Progression Prediction	Anatomical Structure Measurement	Tissue Classification

Table 20: Modality Checklist of Med-Zero-17K

Modality Checklist		
Plain X-ray	Texture Characterization of Bone Radiograph	Mammography
CT	MRI	Ultrasound
Fluoroscopy	Echocardiography	Thermal Imaging
Endoscopy	Colposcopy	OCT
Optical Coherence Tomography	Fundus Photography	CBCT
Laparoscopy	Dermoscopy	Electrocardiogram
Histopathology	Adaptive Optics Ophthalmoscopy	Microscopy
Angiography	SPECT	PET Scan
Scintigraphy	Infrared Reflectance imaging	Medical Photography
Medical diagram	Graph/Chart	Electroencephalogram