MedGR^2: Breaking the Data Barrier for Medical Reasoning via Generative Reward Learning

Weihai Zhi[1*], **Jiayan Guo**[2*], **Shangyang Li**[1†]

[1]Guangdong Institute of Intelligence Science and Technology, Hengqin, Zhuhai, Guangdong, China
[2]School of Intelligence Science and Technology, Peking University, Beijing, China

Abstract

The application of vision-language models in medicine is critically hampered by the scarcity of high-quality, expert-annotated data. Supervised fine-tuning on existing datasets often leads to poor generalization on unseen modalities and tasks, while reinforcement learning, a promising alternative, is stymied by the lack of reliable reward signals in this data-scarce domain. To address this challenge, we propose a Generative Reward Learning framework that establishes a self-improving training cycle. The framework jointly develops a data generator and a reward model, enabling the automated and continuous creation of high-quality multi-modal medical data that serves as an effective training source for post-training. Our experiments demonstrate that supervised fine-tuning using the generated data already surpasses models trained on large-scale human-curated datasets. More importantly, when the generated data is further leveraged for reinforcement learning via Group Relative Policy Optimization, the resulting model achieves state-of-the-art cross-modality and cross-task generalization, significantly outperforming specialized reinforcement-learning-based methods. Notably, a compact model trained under this framework attains performance competitive with foundation models containing more than an order of magnitude more parameters. These results suggest a new paradigm for data-efficient learning in high-stakes medical domains, shifting the bottleneck from data scarcity to data generation and unlocking the potential of reinforcement learning for building robust and generalizable medical AI systems.

Code — https://github.com/HelmarZhi/MedGR-Square

Introduction

Vision-language models (VLMs) hold great promise for transforming clinical reasoning, from multimodal diagnosis to radiology report generation. Their ability to jointly process and reason over medical images and texts offers an unprecedented opportunity to improve decision support, reduce clinician burden, and enable generalizable AI in high-stakes domains. Yet this vision is fundamentally bottlenecked by a key obstacle: the scarcity of high-quality, expert-annotated medical data (Wang, Lu, and Li 2025; Lu and Li 2025). Compared to general-domain tasks, acquiring medical data is constrained by privacy, annotation cost (Guo, Li, and Zhang 2023), and the domain-specific nature of expertise (Li and Guo 2025), which results in narrow task coverage and limited generalization across unseen modalities and tasks.

To address this challenge, the research community has explored two dominant paths, each with clear limitations. The first approach focuses on large-scale data curation and supervised fine-tuning (SFT). Some systems (Chen et al. 2024) attempt to clean and reformat noisy biomedical corpora such as PubMed using large vision-language models (e.g., GPT-4o), producing instruction-style datasets for downstream training. While effective at increasing data scale, these methods remain fundamentally limited by the quality of their source material and the domain alignment of the models used for cleaning. More importantly, as recent studies highlight, SFT-trained models often memorize training distributions and struggle with generalization to new modalities and tasks. This constitutes a critical gap for real-world clinical deployment. (Zhao et al. 2024).

The second path is reinforcement learning (RL), which offers a more powerful paradigm for learning generalizable reasoning strategies. For instance, some works (Lai et al. 2025; Pan et al. 2025) demonstrate that training with RL using small, expert-annotated preference datasets significantly improves cross-task and cross-modal performance over SFT baselines. However, RL in medicine faces a structural limitation: it relies on reward signals that are expensive to obtain, requiring human expert judgments for reward modeling or preference annotation. This creates a paradox: RL requires high-quality supervision to improve generalization, but such supervision is itself too costly to scale.

To break this deadlock, we propose **Generative Reward Learning for Medical Reasoning (MedGR^2)**, a novel self-improving framework that reframes the problem from one of data scarcity to one of data generation. Rather than curating or cleaning existing datasets, MedGR^2 proactively creates new multimodal training samples through a sequential pipeline: a data generator first produces candidate samples (e.g., image-question-answer triples), which are then filtered by a reward model trained to reflect expert-derived criteria such as clinical plausibility, logical consistency, and reason-

*These authors contributed equally.
†Corresponding author: lishangyang@gdiist.cn

ing depth. The resulting high-quality samples are used to train a supervised model (via SFT), which in turn provides a stronger initialization for downstream RL. This pipeline establishes a step-wise process that transitions from raw generation to filtered supervision, and ultimately to generalizable policy learning, effectively enabling scalable and clinically robust medical reasoning. Unlike static preference annotations, our reward model evolves alongside the reasoning policy, enabling progressively refined supervision that scales without additional human effort. Our contributions are as follows:

- We propose MedGR^2, a novel generative reward learning framework that establishes a virtuous cycle to autonomously synthesize high-quality reasoning data for medical understanding.
- We show that our framework achieves state-of-the-art results on OmniMedVQA. SFT on our generated data already surpasses prior fine-tuned methods, and evaluations on the MMMU Health & Medicine track in our ablations further validate its effectiveness. The full MedGR^2 with RL improves performance even more.
- We show unprecedented parameter efficiency, where our compact MedGR^2 model substantially outperforms foundation models over 10x its size (e.g., achieving a +19.7% absolute gain over the 72B-parameter Qwen2-VL-72B), highlighting its immense potential for real-world clinical deployment and service.

Related Work

Data Construction and Curation for Medical VLMs

Constructing high-quality, large-scale datasets remains a central bottleneck in developing medical VLMs. Manual annotation is prohibitively expensive and labor-intensive, prompting a surge of semi-automated solutions. HuatuoGPT-Vision (Chen et al. 2024) reformats PubMed-derived image-text pairs into instruction-following VQA samples using GPT-4V. LLaVA-Med (Li et al. 2023a) synthesizes multimodal QA pairs by prompting ChatGPT with biomedical figures. PMC-VQA (Zhang et al. 2023) leverages large language models and rule-based heuristics to mine high-quality triplets from scientific publications. Flamingo-CXR (Tanno et al. 2025) demonstrates the potential of radiologist-validated datasets, yet the extensive human oversight required severely limits scalability. While these approaches offer valuable contributions, they remain constrained by source data noise, domain coverage gaps, and alignment inconsistencies. In contrast, our framework abandons static corpora altogether, employing a self-improving generative pipeline to synthesize clinically meaningful, task-aligned multimodal instruction data from scratch.

Reinforcement Learning for VLM Generalization

Although SFT has facilitated early progress in VLMs, it often encourages rote pattern learning over robust generalization (Chu et al. 2025). This limitation is especially detrimental in medical AI, where models must generalize to novel modalities and diagnostic scenarios. Recent studies in general domains, including One-Shot RLVR (Wang et al. 2025), LIMO (Ye et al. 2025), and LIMR (Li, Zou, and Liu 2025), illustrate RL's capacity for data-efficient learning and complex reasoning. In the medical domain, (Lai et al. 2025) and MedVLM-R1 (Pan et al. 2025) demonstrate that RL—specifically GRPO—substantially improves generalization when guided by high-quality preferences. However, these methods rely on the availability of curated reward datasets, a costly and unsustainable requirement in clinical contexts. Our approach addresses this shortfall by integrating a generative-reward pipeline that produces instruction samples and corresponding reward signals automatically.

Methodology

To address the dual challenges of *data scarcity* and *reasoning generalization* in medical VLMs, we propose **MedGR^2**, a unified, self-improving framework that couples scalable instruction generation with reinforcement-based policy optimization.Rather than a static pipeline, MedGR^2 functions as a dynamic ecosystem where a *Multimodal Generator* G_θ and a *Reward Model* R_ϕ co-evolve to fuel a two-stage learning process for our final Reasoning Policy π_ψ. This process begins with reward-filtered SFT for base alignment, followed by RL for superior generalization. An overview is shown in Figure 1.

Multimodal Generator G_θ: Engineering the Data Engine

The generator G_θ is responsible for synthesizing diverse and clinically meaningful VQA triplets (I, q, a), where I is a medical image, q is a diagnostic or descriptive instruction, and a is a rationale-grounded answer. Unlike classical data augmentation methods, G_θ generates *semantically novel* instruction-answer pairs guided by domain priors and prompt structure.

Prompt-Driven Controllable Generation. We implement G_θ using a powerful multimodal backbone (e.g., Gemini-2.0-Flash) and explore three prompting paradigms: (1) direct instruction, (2) step-by-step reasoning, and (3) meta-cognitive introspection. Prompt (3), which explicitly prompts the model to reflect on question types and expected answer structure, consistently yields more diverse and clinically aligned outputs. Full templates are provided in the supplementary.

Self-Updating Generator. The generator is periodically fine-tuned on high-reward examples filtered by R_ϕ, allowing it to incrementally adapt to downstream reward signals and policy behaviors. This results in a self-enhancing data engine where generated samples become increasingly relevant and high-quality over iterations.

Reward Model R_ϕ: The Co-Evolving Expert

The reward model R_ϕ serves as an automated domain expert, assigning scalar rewards $r \in [-6, 10]$ to candidate triplets (I, q, a). We implement R_ϕ as a lightweight vision-language model followed by a regression head that predicts

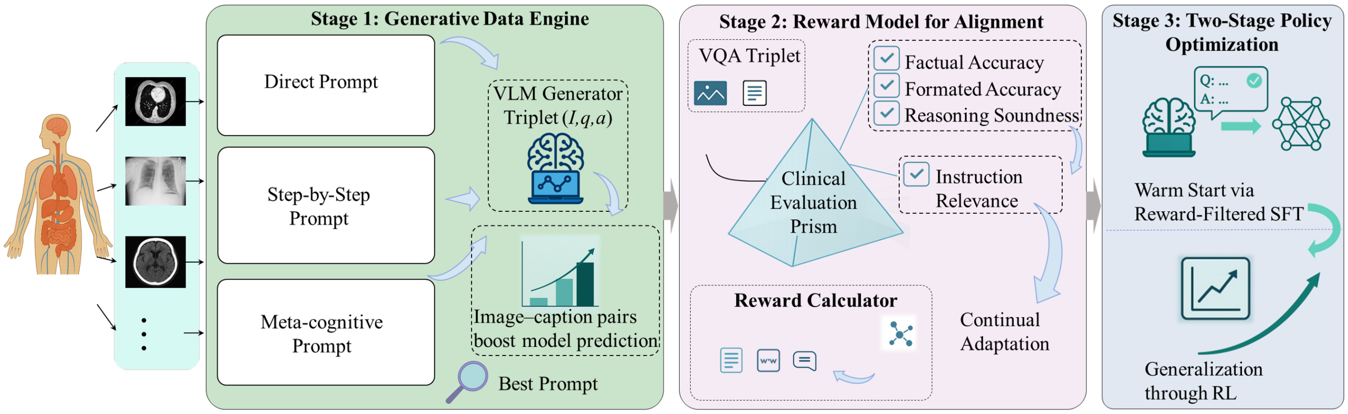


Figure 1: **Overview of the MedGR^2 framework.** Our self-improving framework operates in three sequential stages. Stage 1: VQA Generation. A VLM generator, guided by one of three prompt engineering strategies (Direct, Step-by-Step, Meta-cognitive), synthesizes multimodal VQA triplets (I, q, a) from various medical images. Our findings indicate the Meta-cognitive prompt yields the highest quality data. Stage 2: Reward Model for Alignment. A dynamic Reward Model, visualized as a "Clinical Evaluation Prism," assesses each generated triplet based on four key criteria (Factual Accuracy, Formatted Accuracy, Reasoning Soundness, Instruction Relevance). This model continually adapts using new data, providing a robust quality signal. Stage 3: Two-Stage Policy Optimization. The final reasoning policy is trained in a two-stage process. First, it receives a Warm Start via Reward-Filtered SFT using high-quality data from Stage 2. Second, it is further optimized to achieve superior Generalization through RL (GRPO), creating a highly capable and robust medical reasoning model.

a reward score reflecting multiple clinically-relevant dimensions:

- Factual Accuracy: Is the answer clinically plausible given the image?
- Reasoning Soundness: Is the rationale coherent and logically consistent?
- Instruction Relevance: Does the question pertain to meaningful clinical tasks (e.g., diagnosis, description, follow-up)?

Multi-Grade Supervision. Unlike standard pairwise preference training, we design a four-grade reward dataset $\mathcal{D}_{\text{grade}}$ to supervise R_ϕ with finer control:

- Grade 1 (score = 10): Original reports written by radiologists.
- Grade 2 (score = 6): Lightly perturbed versions using WordNet-based synonym replacements.
- Grade 3 (score = 0): Incoherent answers created by random phrase deletions.
- Grade 4 (score = -6): Irrelevant or hallucinated responses.

We minimize a mean squared error loss:

$$\mathcal{L}_{\text{RM}} = \mathbb{E}_{(x,r)\sim\mathcal{D}_{\text{grade}}}\left[(R_\phi(x) - r)^2\right] \quad (1)$$

Continual Adaptation. Unlike static RMs trained once and deployed, R_ϕ in MedGR^2 is designed for continual adaptation. It is initialized on a small seed dataset (e.g., distilled GPT-4o preferences or expert-annotated pairs) and is periodically retrained on increasingly challenging examples generated by the improving reasoning policy π_ψ. This co-evolution enables R_ϕ to serve as a progressively more discerning judge, aligned with the distribution shift and complexity escalation in the generated tasks.

Two-Stage Policy Optimization

We train the final reasoning policy π_ψ in two distinct but complementary stages, using the high-quality dataset $\mathcal{D}_{\text{gen}}$ filtered by R_ϕ (where $R_\phi(x) > \tau$).

Stage 1: Warm Start via Reward-Filtered SFT The first stage provides a warm start by aligning π_ψ with proven high-quality reasoning patterns. We train on $\mathcal{D}_{\text{gen}}$ with the next-token prediction objective:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(I,q,a)\sim\mathcal{D}_{\text{gen}}}\left[\log \pi_\psi(a \mid I, q)\right] \quad (2)$$

The primary motivation for this SFT stage is to mitigate the cold-start problem for RL. By first aligning the policy with a corpus of high-quality demonstrations, we anchor it in a region of the policy space characterized by coherent reasoning and correct formatting. This significantly stabilizes the subsequent RL exploration phase and accelerates convergence. This phase bootstraps a robust baseline model to comprehend clinically valid reasoning structure and content.

Stage 2: Generalization through Reinforcement Learning While SFT effectively aligns the model with expert demonstrations, it often results in overly deterministic behavior, limiting the policy's robustness and adaptability. To overcome this, we introduce a RL stage that promotes exploration beyond the demonstration distribution. Rather than imitating a single ground truth, the policy π_ψ is trained to optimize a reward signal, enabling it to uncover a broader spectrum of clinically valid reasoning trajectories. This reward-driven optimization is essential for enhancing generalization to out-of-distribution scenarios and unforeseen clinical inputs.

We employ GRPO (Pan et al. 2025) to further refine π_ψ under a composite reward that captures both reasoning qual-

Algorithm 1: MedGR^2 Closed-Loop Training

Input: Seed dataset $\mathcal{D}_{\text{seed}}$, threshold τ, generator $G_\theta^{(0)}$, reward model $R_\phi^{(0)}$, overlap flag `allow_overlap`
Output: Optimized reasoning policy π_ψ

```
Initialize D_gen ← D_seed
for t = 1, 2, ... do
    Generate candidate samples: D_cand ← G_θ^(t)
    Score all samples: r(x) ← R_φ^(t)(x), ∀x ∈ D_cand
    Filter high-quality samples:
        D_high = {x ∈ D_cand | r(x) > τ}
    if allow_overlap is False then
        Remove overlaps: D_high ← D_high \ D_gen
    end if
    Update training set: D_gen ← D_gen ∪ D_high
    SFT: Train π_ψ on D_gen using loss L_SFT
    RL: Fine-tune π_ψ via GRPO using reward r(a')
    Generate preferences: D_pref ← Sample from π_ψ^(t)
    Update reward model: Train R_φ^(t+1) on D_pref
    Optionally update generator G_θ^(t+1) via distillation
end for
```

ity and answer correctness. Specifically, the reward function is defined as:

$$r(a') = \alpha \cdot R_\phi(I, q, a') + \beta \cdot \mathbb{I}[\texttt{extract}(a') = \texttt{extract}(a)] \tag{3}$$

where R_ϕ evaluates the semantic and clinical plausibility of the generated answer, and the indicator function $\mathbb{I}[\cdot]$ enforces alignment with the correct answer entity extracted via $\texttt{extract}(\cdot)$. This hybrid dense-plus-sparse reward structure encourages the policy to generate coherent multi-step rationales while preserving factual consistency in the final prediction.

The Iterative MedGR^2 Loop

MedGR^2 operates as a closed, iterative loop:

1. **Generate:** The generator $G_\theta^{(t)}$ produces a new batch of candidate samples.
2. **Filter:** The reward model $R_\phi^{(t)}$ evaluates the candidates; high scorers augment $\mathcal{D}_{\text{gen}}$.
3. **Train Policy:** $\pi_\psi^{(t)}$ is updated via the two-stage SFT+RL process.
4. **Evolve:** $\pi_\psi^{(t+1)}$ generates new preference pairs to refine $R_\phi^{(t+1)}$ and optionally update $G_\theta^{(t+1)}$.

This co-evolutionary loop progressively enhances both data generation and reasoning in complex clinical domains.

Experiments & Results

To evaluate the effectiveness of MedGR^2 in enhancing generalization and data quality for medical visual reasoning, we conduct comprehensive experiments on the OmniMedVQA benchmark.

Research Questions

Our experimental design is structured to answer four key research questions that progressively validate our framework's effectiveness:

- **RQ1: State-of-the-Art Generalization.** Does our full MedGR^2 framework achieve state-of-the-art generalization performance against leading medical VLMs?
- **RQ2: The Synergy of Data and Reinforcement Learning.** What are the respective contributions of our high-quality generated data (via SFT) and the subsequent reinforcement learning stage (via GRPO)?
- **RQ3: Beyond Accuracy: A Granular Analysis of Generalization.** How does MedGR^2 transfer knowledge across domains, and where does it specifically correct the errors of strong baseline models?

Datasets. We use OmniMedVQA (Hu et al. 2024), a large-scale, multi-specialty medical VQA benchmark. It contains over 300k triplets covering radiology, dermatology, pathology, and ophthalmology, with diverse reasoning types (e.g., diagnosis, procedural, descriptive). We use the official train/val/test splits for all experiments.

Task Setting. Our evaluation employs two distinct settings. First, for our main SOTA comparison (RQ1, RQ2), we follow the standard protocol of training models on the full, mixed-modality OmniMedVQA training set and evaluating their performance on the test set for each specific modality. Second, to specifically analyze knowledge transfer (RQ3), we adopt the stricter cross-domain protocol from Med-R1 (Lai et al. 2025), where models are trained on a single source domain (e.g., only MRI data) and evaluated on unseen target domains in a zero-shot manner.

Baseline Methods. We compare MedGR^2 against a comprehensive suite of baseline models, including both zero-shot and fine-tuned paradigms:

- **General-Purpose VLMs (Zero-shot):** To assess the performance ceiling of large-scale vision-language models without any domain-specific tuning, we evaluate **BLIP-2** (Li et al. 2023c), **InstructBLIP** (Dai et al. 2023), **LLaVA** (Li et al. 2024), **MiniGPT-4** (Zhu et al. 2023), and **LLaMA Adapter v2** (Gao et al. 2023), as well as strong recent models like **Qwen2-VL-7B** and **Qwen2-VL-72B** (Wang et al. 2024), and their upgraded versions **Qwen2.5-VL-7B/72B** (Bai et al. 2025).
- **Medical VLMs (Zero-shot):** We include specialized models trained on biomedical data, such as **LLaVA-Med** (Li et al. 2023b), **RadFM** (Wu et al. 2023), **Med-Flamingo** (Moor et al. 2023), and **MedVInT** (Zhang et al. 2024), to benchmark domain-specific zero-shot capabilities.
- **Fine-tuned VLMs:** This group contains all models that are fine-tuned on the OmniMedVQA training split.

Methods	Modality								Overall
	CT	MRI	X-Ray	US	Der	FP	OCT	Micro	
Zero-shot VLMs									
BLIP-2†	56.74	41.32	67.58	37.27	40.65	46.24	68.08	50.40	51.04
InstructBLIP†	28.72	33.15	61.04	41.25	62.22	50.31	42.59	46.29	45.70
LLaVA†	17.73	26.72	30.70	18.66	49.74	47.11	33.73	28.87	31.66
LLaMA Adapter v2†	21.41	36.65	46.44	34.31	44.13	51.34	30.15	33.17	37.20
MiniGPT-4†	22.81	27.48	38.30	25.50	40.52	33.40	31.36	36.23	32.54
Qwen2-VL-7B	61.46	40.34	64.27	51.11	60.02	63.70	53.34	53.64	56.40
Qwen2-VL-72	67.97	69.39	72.11	51.39	65.31	72.58	72.76	67.84	68.05
Qwen2.5-VL-7B	60.44	58.44	73.99	50.46	62.48	67.66	67.40	61.87	60.29
Qwen2.5-VL-72B	66.18	63.64	79.81	69.85	69.75	71.04	69.22	69.37	67.71
Zero-shot Medical VLMs									
LLaVA-Med†	18.69	27.47	30.68	29.88	44.95	39.03	34.61	33.29	32.33
RadFM†	27.56	24.06	30.95	16.57	39.21	36.86	32.80	27.97	29.50
Med-Flamingo†	31.28	26.34	44.01	31.69	48.56	41.26	25.16	30.03	34.29
MedVInT†	40.74	43.10	55.10	41.26	29.11	31.84	23.26	32.02	37.05
Fine-tuned VLMs									
Qwen2.5-VL-3B (SFT)	66.00	60.81	69.37	45.11	62.11	63.95	65.66	60.83	61.73
Qwen2.5-VL-7B (SFT)	97.75	60.15	69.37	45.11	62.11	63.95	65.66	60.83	81.32
HuatuoGPT-34B	60.80	66.50	83.80	81.70	74.00	85.50	90.00	67.40	76.70
HealthGPT-3.8B	-	-	-	-	-	-	-	-	68.50
HealthGPT-14B	-	-	-	-	-	-	-	-	74.40
Med-R1-2B (Think)	66.30	71.61	84.52	57.31	72.33	71.33	71.96	70.80	70.77
Med-R1-3B (Nothink)	69.89	72.91	84.52	43.91	73.62	80.10	84.18	71.44	72.57
MedGR^2(SFT)	99.86	64.57	87.52	47.14	74.34	73.07	95.62	71.20	<u>85.59</u>
MedGR^2(SFT+GRPO)	99.58	65.00	92.65	46.47	74.55	95.55	94.42	67.86	**87.45**

Table 1: Comprehensive performance comparison of VLMs on the eight modalities of the OmniMedVQA benchmark. The table is segmented into three categories: zero-shot general-purpose VLMs, zero-shot medical VLMs, and fine-tuned VLMs. All fine-tuned models were trained on the OmniMedVQA training set for the respective modality.

- **SFT Baselines:** We fine-tune strong general-purpose models such as **Qwen2.5-VL-3B** and **Qwen2.5-VL-7B** using standard supervised instruction tuning. We also consider curated data generation pipelines like **HuatuoGPT-Vision** (Chen et al. 2024) and **HealthGPT-Vision**(Lin et al. 2025) when available.
- **RL Baselines:** We include **Med-R1** (Lai et al. 2025) variants (2B and 3B, with and without reasoning modules) as state-of-the-art reward-learning based VLMs, and compare directly against our own model variants: **MedGR^2 (SFT)** and **MedGR^2 (SFT+GRPO)**.

Evaluation Metrics. We report top-1 Accuracy as the primary metric for all VQA tasks. To provide a more nuanced view on class-imbalanced diagnostic tasks, we also report Balanced Accuracy and AUROC where applicable.

Implementation Details. Our reasoning policy π_ψ is based on the Qwen2.5-VL-7B architecture. The generator G_θ utilizes the Gemini-2.0-Flash API with the meta-cognitive prompt identified as most effective. The MedGR^2 framework was run for 3 iterative cycles, generating approximately 50k VQA samples per cycle. The reward model R_ϕ is initialized from a pre-trained 3B VLM and fine-tuned on an initial seed set of 2k preference pairs distilled from GPT-4o. For RL, we use the GRPO implementation from Med-R1 with reward coefficients $\alpha = 0.8$ and $\beta = 0.2$. All models were trained on a cluster of 8x NVIDIA L40 GPUs. Further hyperparameters are detailed in the Appendix.

RQ1: State-of-the-Art Performance Across Medical Modalities

To answer our first research question, we conduct a large-scale comparison of MedGR^2 against a diverse set of leading VLMs on the OmniMedVQA benchmark. Table 1 reports the detailed accuracy across eight clinically diverse imaging modalities.

Overall State-of-the-Art Performance. Our full model MedGR^2 achieves the highest overall accuracy of 87.45%, outperforming all baselines by a significant margin. Compared to the 72B-parameter foundation model Qwen2.5-VL-72B (67.71%), our 7B MedGR^2 yields a +19.74% absolute gain, demonstrating that a well-designed generative and reward-guided framework can surpass models more than 10× larger in scale.

Cross-Specialty Generalization. MedGR^2 excels across a wide range of modalities with top-tier performance in highly heterogeneous domains such as X-Ray (92.65%), Fundus Photography (95.55%), and OCT (94.42%). This consistent dominance reflects our method's ability to generate semantically rich, clinically grounded supervision signals that generalize beyond narrow modality-specific distributions.

The Power of Synthetic Data. Even without reinforcement learning, MedGR^2 (SFT)—trained purely on generated data—achieves 85.59% accuracy, outperforming all fine-tuned baselines, including HuatuoGPT-34B (76.70%) and the RL-based Med-R1-3B (72.57%). This validates that our generator-reward pipeline produces training samples of higher fidelity and diversity than human-curated datasets.

RL Amplifies Generalization. Applying GRPO on top of the SFT model yields a further boost to 87.45%, highlighting the synergistic effect of reinforcement learning when trained in a reward-informed data ecosystem. Unlike static preference tuning, our reward model evolves alongside the reasoning policy, enabling progressively refined supervision without additional human effort.

These results challenge the "scale-is-all-you-need" hypothesis in medical AI. MedGR^2 demonstrates that generalization can be achieved more efficiently through intelligent data generation and optimization rather than brute-force scaling, thereby paving the way for practical, high-performing VLMs deployable in real-world healthcare settings.

RQ2: The Synergy of High-Quality Data and Reinforcement Learning

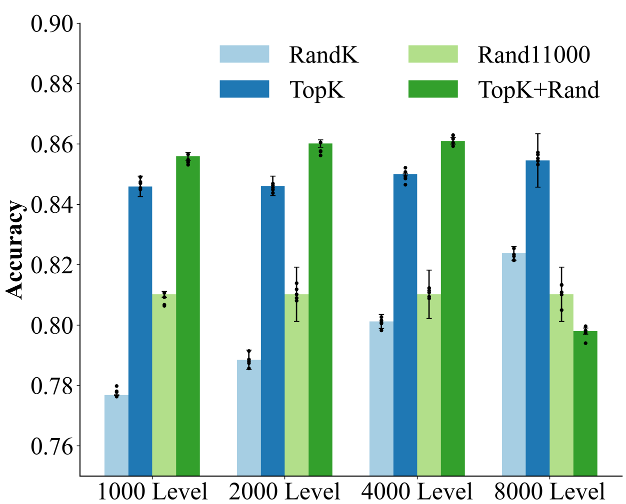


Figure 2: The synergy of reward-filtered data and Reinforcement Learning across varying data scales. We compare four training strategies: 'RandK' (SFT on K randomly sampled generated data); 'TopK' (SFT on the top K samples selected by our reward model; 'Rand11000' (a baseline SFT on 11,000 unfiltered samples); and 'TopK+Rand' (MedGR^2 with GRPO applied on the 'TopK' data).

To better understand the components behind MedGR^2's superior performance, we investigate RQ2 by analyzing the respective contributions of our reward-filtered data and the RL stage. Our goal is to characterize how each component contributes to performance gains across both varying data scales and datasets.

Models	OmniMed			MMMU		
	Acc	F1	AUC	Acc	F1	AUC
Qwen2.5-VL	73.63	68.10	69.55	69.42	67.35	71.88
+ SFT (raw GT answer)	81.32	80.21	76.74	73.18	71.02	75.43
+ SFT (rewarded pairs)	85.56	84.19	83.92	76.83	74.95	78.30
+ GRPO	**87.45**	-	-	**78.66**	**76.51**	**79.88**

Table 2: Ablation study on OmniMed and MMMU datasets. We evaluate models using Accuracy, F1, and AUC. Reward-based SFT and GRPO provide consistent gains over raw SFT baselines.

Reward-Filtered Data Enables Extreme Data Efficiency. Figure 2 evaluates SFT models trained on different data subsets. We compare: (i) randomly selected K samples ('RandK'), (ii) top-K high-reward samples ('TopK'), and (iii) 11,000 raw samples ('Rand11000'). Across all sizes (1k–8k), TopK substantially outperforms RandK. Remarkably, 'TopK' with only 2,000 samples achieves 84.6% accuracy—exceeding the large-scale 'Rand11000' baseline. This demonstrates the high quality and efficiency of our generative reward filtering, which selects valuable training samples automatically.

Reinforcement Learning Unlocks Robust Reasoning. Our framework leverages RL via GRPO to further enhance generalization. Table 2 presents ablation results across two benchmarks: OmniMedVQA and MMMU.

We observe clear step-wise gains: SFT on raw GT pairs yields 81.32% on OmniMedVQA. Using reward-filtered pairs improves to 85.56%, and applying GRPO further raises it to 87.45%, confirming strong synergy. On MMMU, the full pipeline achieves 78.66%, outperforming raw SFT by 5.48%. These results validate that reward-driven generation and RL form a complementary, self-reinforcing loop, boosting overall performance.

When Does RL Matter Most? Figure 2 further reveals an interesting trend: RL provides the most benefit when the amount of reward-filtered data is small to moderate (e.g., 1k–4k). At 8k, however, SFT-only models slightly outperform their RL-augmented counterparts. This suggests a potential saturation point where sufficient high-quality data enables effective imitation learning, and the marginal benefits of further RL diminish. Thus, we recommend applying GRPO in scenarios with limited, high-quality supervision, which is common in many specialized medical domains.

RQ3: A Granular Analysis of Generalization and Its Origins

To further understand the sources of MedGR^2's effectiveness, we investigate how it generalizes across tasks and modalities. Our analysis focuses on two key factors: (1) transferability from imbalanced training data and (2) fine-grained error correction dynamics. Figure 3 shows the training data distribution: modalities like MRI (42.8%) and CT

(20.6%) dominate, while Microscopy (3.86%) is underrepresented. Despite this imbalance, MedGR^2 exhibits non-trivial transfer patterns.

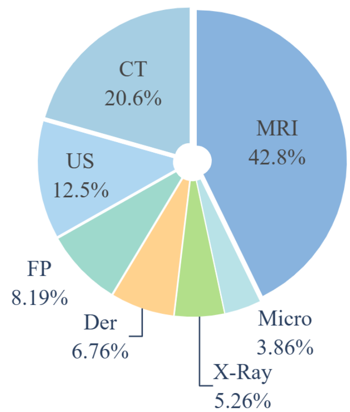
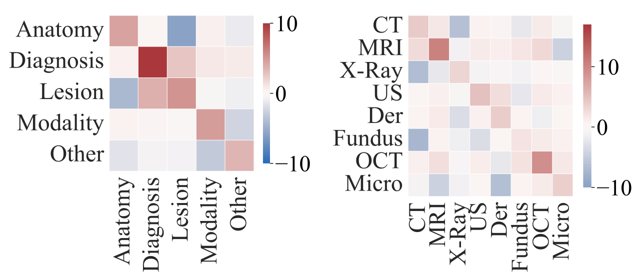


Figure 3: Distribution of modalities in all training dataset.



(a) Cross-task transfer performance.

(b) Cross-modality transfer performance.

Figure 4: Cross-task and cross-modality transferability. Warmer colors indicate positive transfer, while cooler tones show negative effects.

Transferability Across Tasks and Modalities Figure 3 shows the training data distribution: modalities like MRI (42.8%) and CT (20.6%) dominate, while Microscopy (3.86%) is underrepresented. Despite this imbalance, MedGR^2 exhibits non-trivial transfer patterns.

The task transfer heatmap (Figure 4a) reveals that training on Diagnosis leads to broad positive transfer across tasks, suggesting it captures foundational reasoning skills. Meanwhile, the modality transfer heatmap (Figure 4b) shows that frequent modalities like MRI enable strong transfer to others like CT and Microscopy, indicating that MedGR^2 learns generalizable features rather than memorizing frequency.

Error Correction Breakdown To analyze how these capabilities translate into performance gains, we perform a head-to-head comparison with a strong SFT baseline. Figure 5 visualizes the transitions: correct-to-correct (green), wrong-to-correct (blue), correct-to-wrong (orange), and wrong-to-wrong (red).

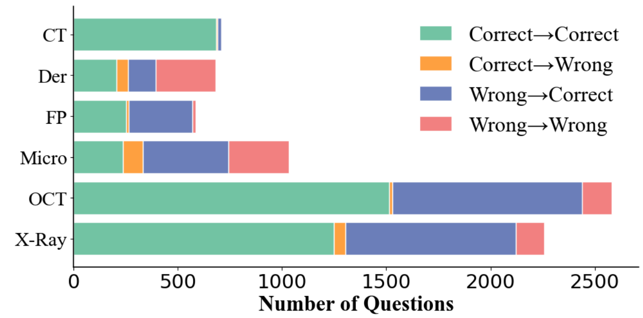


Figure 5: Head-to-Head Error Transition Analysis. Each bar dissects the entire question set for a modality into four transition flows based on the outcomes of the two models. The four flows are: Correct→Correct (green), where both models succeed; Wrong→Correct (blue), where MedGR^2 corrects the baseline's error; Correct→Wrong (orange), where our model makes a new error on a question the baseline answered correctly; and Wrong→Wrong (red), where both models fail.

MedGR^2 consistently yields more corrected errors (blue) than new ones (orange) across all modalities. Particularly in high-resource domains like X-Ray and OCT, the number of baseline errors fixed by MedGR^2 is substantial. Even in low-resource settings (e.g., Dermatology), it shows meaningful improvements. These results confirm that MedGR^2 not only generalizes well but also systematically reduces error rates, showcasing its robust and transferable reasoning ability.

Conclusion

In this work, we introduced MedGR^2, a novel framework addressing data scarcity and generalization challenges in medical VLM via a self-improving virtuous cycle of data generation and reward modeling. Our experiments show MedGR^2 achieves new state-of-the-art performance on the OmniMedVQA benchmark, outperforming strong SFT and RL baselines, and even substantially larger foundation models, demonstrating remarkable parameter efficiency. By transforming the core problem from data curation to intelligent, automated data generation, MedGR^2 paves the way for more robust, generalizable, and scalable AI for high-stakes clinical reasoning.

Acknowledgments

This research was Supported by Guangdong S&T Program under Grant 2023B0303040002, and the Young Scientists Fund of the National Natural Science Foundation of China (C-Class, No. 32500997). The authors would like to express their sincere gratitude to Prof. Lirong Zheng and Prof. Yuxiang Huan for their constructive discussions and valuable feedback. The authors also thank all members of the lab for their continuous support.

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Chen, J.; Gui, C.; Ouyang, R.; Gao, A.; Chen, S.; Chen, G.; Wang, X.; Cai, Z.; Ji, K.; Wan, X.; et al. 2024. Towards injecting medical visual knowledge into multimodal llms at scale. 7346–7370.

Chu, T.; Zhai, Y.; Yang, J.; Tong, S.; Xie, S.; Schuurmans, D.; Le, Q. V.; Levine, S.; and Ma, Y. 2025. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Guo, J.; Li, S.; and Zhang, Y. 2023. An Information Theoretic Perspective for Heterogeneous Subgraph Federated Learning. *Proceedings of the International Conference on Database Systems for Advanced Applications (DASFAA)*, 745–760.

Hu, Y.; Li, T.; Lu, Q.; Shao, W.; He, J.; Qiao, Y.; and Luo, P. 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*, 22170–22183.

Lai, Y.; Zhong, J.; Li, M.; Zhao, S.; and Yang, X. 2025. Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv preprint arXiv:2503.13939*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023b. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR.

Li, S.; and Guo, J. 2025. Subgraph Federated Learning with Information Bottleneck Constrained Generative Learning. *ACM Trans. Knowl. Discov. Data*, 19(6).

Li, X.; Zou, H.; and Liu, P. 2025. Limr: Less is more for rl scaling. *arXiv preprint arXiv:2502.11886*.

Lin, T.; Zhang, W.; Li, S.; Yuan, Y.; Yu, B.; Li, H.; He, W.; Jiang, H.; Li, M.; Song, X.; et al. 2025. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*.

Lu, T.; and Li, S. 2025. Harnessing Pre-trained Language Models for EEG-based Epilepsy Detection. *Proceedings of the 2025 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H. B.; Chen, C.; Ouyang, C.; and Rueckert, D. 2025. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*.

Tanno, R.; Barrett, D. G.; Sellergren, A.; Ghaisas, S.; Dathathri, S.; See, A.; Welbl, J.; Lau, C.; Tu, T.; Azizi, S.; et al. 2025. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2): 599–608.

Wang, K.; Lu, T.; and Li, S. 2025. Empowering Cross-Patient Adaptive-Length Epilepsy Diagnosis with ECNorm: A Channel-wise Approach. *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, Y.; Yang, Q.; Zeng, Z.; Ren, L.; Liu, L.; Peng, B.; Cheng, H.; He, X.; Wang, K.; Gao, J.; et al. 2025. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*.

Wu, C.; Zhang, X.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *arXiv preprint arXiv:2308.02463*.

Ye, Y.; Huang, Z.; Xiao, Y.; Chern, E.; Xia, S.; and Liu, P. 2025. Limo: Less is more for reasoning. *arXiv preprint arXiv:2502.03387*.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2023. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhang, X.; Wu, C.; Zhao, Z.; Lin, W.; Zhang, Y.; Wang, Y.; and Xie, W. 2024. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1): 277.

Zhao, W.; Wang, Y.; Li, Z.; Lin, K.; and Chen, J. 2024. Towards Robust Multimodal Pretraining for Medical VQA: Benchmarking and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.