MedEyes: Learning Dynamic Visual Focus for Medical Progressive Diagnosis

Chunzheng Zhu, Yangfang Lin, Shen Chen, Yijun Wang*, Jianxin Lin*

Hunan University
Changsha, Hunan, China
zhuchzh@hnu.edu.cn, lyfang123@hnu.edu.cn, cs05@hnu.edu.cn,
wyjun@hnu.edu.cn, linjianxin@hnu.edu.cn

Abstract

Accurate medical diagnosis often involves progressive visual focusing and iterative reasoning, characteristics commonly observed in clinical workflows. While recent vision-language models demonstrate promising chain-of-thought (CoT) reasoning capabilities via reinforcement learning with verifiable rewards (RLVR), their purely on-policy learning paradigm tends to reinforce superficially coherent but clinically inaccurate reasoning paths. We propose MedEyes, a novel reinforcement learning framework that dynamically models clinician-style diagnostic reasoning by progressively attending to and interpreting relevant medical image regions. By incorporating off-policy expert guidance, MedEyes converts expert visual search trajectories into structured external behavioral signals, guiding the model toward clinically aligned visual reasoning. We design the Gaze-guided Reasoning Navigator (GRN) to emulate the diagnostic process through a dual-mode exploration strategy, scanning for systematic abnormality localization and drilling for detailed regional analysis. To balance expert imitation and autonomous discovery, we introduce the Confidence Value Sampler (CVS), which employs nucleus sampling and adaptive termination to create diverse yet credible exploration paths. Finally, the dual-stream GRPO optimization framework decouples on-policy and off-policy learning signals, mitigating reward assimilation and entropy collapse. Experiments demonstrate that MedEyes achieves an average performance improvement of +8.5% across multiple medical VQA benchmarks, validating MedEyes's potential in building trustworthy medical AI systems.

1 Introduction

Recent breakthroughs in medical vision-language models (VLMs), such as RadFM (Wu et al. 2023), PathChat (Lu et al. 2024), and MMed-RAG (Xia et al. 2025), have demonstrated exceptional capabilities in medical question answering and report generation tasks. These models leverage large-scale multimodal pretraining to support complex diagnostic reasoning and clinical decision-making. Beyond static diagnosis, recent large multi-step reasoning models (LRMs) have demonstrated promising capabilities in Chain-of-Thought (CoT) reasoning and self-reflective inference, particularly in text-based medical image interpretation tasks (Lai et al. 2025; Pan et al. 2025). More recently, the emergence of visual CoT methods (OpenAI 2025; Zheng et al. 2025; Fan et al. 2025) has opened new directions for progressive, vision-grounded reasoning, bridging the gap between textual logic and spatial evidence.

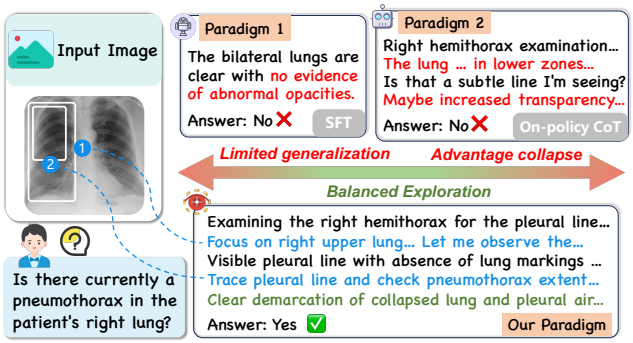

Figure 1: Comparison of medical CoT training paradigms: SFT produces overly generic responses that miss critical findings; on-policy CoT allows exploration but suffers from advantage collapse leading to incorrect reasoning; MedEyes achieves accurate pneumothorax identification through systematic visual grounding and targeted regional analysis.

Supervised fine-tuning (SFT) typically relies on large-scale annotated CoT data to learn diagnostic patterns (Yu et al. 2025). While effective in capturing task-specific knowledge, this approach often overfits to memorized reasoning trajectories, limiting its generalization and reducing the faithfulness of reasoning in unseen clinical scenarios (Zhuang et al. 2025). Recent text-based CoT methods improve interpretability by introducing intermediate reasoning steps (Lai et al. 2025), and reinforcement learning with verifiable rewards (RLVR) further refines diagnostic accuracy by optimizing reward-aligned structured reasoning processes (Pan et al. 2025). However, both approaches largely operate in the textual domain, lacking explicit grounding between reasoning steps and visual evidence. This disconnect can lead to information loss and visual hallucinations, especially in complex medical imaging tasks. As illustrated in the pneumothorax detection example (Fig. 1), SFT models tend to produce generic and vague responses, while on-policy CoT reasoning, despite exploratory analysis, easily

*Corresponding authors.

falls into the "advantage collapse" trap, generating seemingly plausible but substantively incorrect reasoning paths, ultimately reaching wrong conclusions.

These limitations raise a critical question in medical visual reasoning: *How can we enable models to acquire the progressive visual focus and iterative diagnostic refinement that characterize expert clinical workflows?*

Accurate medical diagnosis relies on expert-level reasoning supported by high-quality exploration trajectories. However, models initialized with limited capabilities often fall into local optima, resulting in repetitive, low-quality reasoning cycles, which we refer to as cognitive traps. Off-policy expert trajectories can act as cognitive anchors, guiding effective diagnostic exploration. Yet, naive behavior cloning tends to mimic action sequences without capturing their underlying reasoning logic, while overly exploring may deviate from clinically valid cognitive structures. Bridging this gap requires a hybrid learning paradigm that combines expert-guided supervision with self-driven exploration. Such a framework should support not only behavioral imitation but also knowledge-level generalization, enabling models to internalize external clinical expertise as intrinsic reasoning skills. Moreover, when reasoning steps are explicitly grounded to visual regions, the model can achieve a natural alignment between "precise observation" and "structured reasoning", thereby establishing a consistent mapping between image evidence and diagnostic descriptions.

This work introduces MedEyes, a hybrid reinforcement learning framework that captures dynamic clinician-style attention patterns during diagnostic processes through medical CoT reasoning. Our approach leverages structured off-policy expert trajectories as cognitive anchors, enabling models to internalize expert diagnostic behaviors while preventing policy collapse. Specifically, we design the Gaze-guided Reasoning Navigator (GRN), implementing dual-mode exploration: a scanning mode for identifying candidate abnormal regions, and a drilling mode for conducting focused pathological analysis. This mirrors human diagnostic workflows as captured in eye-tracking studies (Brunyé et al. 2019; Castner et al. 2020; Sultana, Qin, and Yin 2024). To balance expert imitation with autonomous exploration, the Confidence Value Sampler (CVS) generates high-quality reasoning trajectories via nucleus sampling and dynamically adjusts exploration depth based on confidence feedback. Finally, we develop a dual-stream variant of Group Relative Policy Optimization (GRPO) (Shao et al. 2024) that decouples training signals from on-policy and off-policy sources. This design mitigates reward assimilation by preventing expert trajectories from dominating autonomous learning signals and avoids entropy collapse that constrains exploration and impairs generalization to novel diagnostic scenarios.

We evaluate MedEyes on five established medical visual question answering benchmarks covering diverse imaging modalities, ranging from radiological analysis to histopathological examination. Our experiments reveal that the proposed framework successfully enables training of initially weak models, which prove intractable under pure on policy reinforcement learning, while exhibiting clinically aligned progressive visual attention and iterative diagnostic reasoning. These findings establish MedEyes as a transformative paradigm for developing interpretable medical AI systems with generalizable visual reasoning capabilities. The main contributions of this paper can be summarized as follows:

- We propose MedEyes, a dynamic focusing multi-round reasoning RL framework, which breaks through the limitations of traditional medical post-training by introducing structured off-policy expert trajectories.
- We design a collaborative mechanism between the Gaze-guided Reasoning Navigator (GRN) and Confidence Value Sampler (CVS), where the former reproduces diagnostic workflows through scanning-drilling dual-mode strategies, and the latter constructs a diverse and high-quality off-policy trajectory library.
- We use dual-stream GRPO optimization architecture to address reward assimilation and entropy collapse issues by isolating on-policy and off-policy learning components, achieving a balance between expert-level diagnostic pattern learning and task adaptability.
- Comprehensive validation on five medical visual question answering benchmarks shows that this method not only significantly outperforms existing methods, but achieves breakthroughs in clinical interpretability and visual localization accuracy, providing a new technical pathway for building trustworthy medical AI systems.

2 Methodology

2.1 Framework Overview

MedEyes presents a mixed-policy RL framework leveraging dual-stream GRPO optimization to overcome the cognitive traps and policy collapse endemic to SFT and pure on-policy paradigms through structured off-policy expert trajectory integration. The architecture contains two synergistic components: (1) an on-policy exploration stream, where the policy π_θ autonomously samples diagnostic trajectories; and (2) an off-policy guidance stream, which constructs expert trajectories via the Gaze-guided Reasoning Navigator (GRN) and Confidence Value Sampler (CVS), establishing cognitive anchors that help the agent generalize beyond initialization and support progressive refinement of diagnostic reasoning.

Gaze-guided Multi-round Reasoning Medical visual reasoning is formalized as a Markov decision process enabling progressive diagnostic attention refinement through reinforcement learning mechanisms. Given medical image I and clinical query q, policy π_θ generates diagnostic trajectory $\tau = [n_1, n_2, ..., n_T, a]$ where each reasoning step $n_t = \langle s_t, \mathcal{G}_t \rangle$ encapsulates textual cognition s_t and visual grounding $\mathcal{G}_t = \{(x_{i,1}, y_{i,1}, x_{i,2}, y_{i,2})\}$, culminating in diagnostic answer a. The reasoning process manifests as a factorized probability distribution over trajectory space:

$$\pi_\theta(\tau \mid I, q) = \prod_{t=1}^{T} \pi_\theta(n_t \mid I, q, n_{<t}) \cdot \pi_\theta(a \mid I, q, n_{\leq T}), \quad (1)$$

where policy π_θ parameterized by θ maximizes expected rewards across hybrid trajectory distributions, acquiring clini-

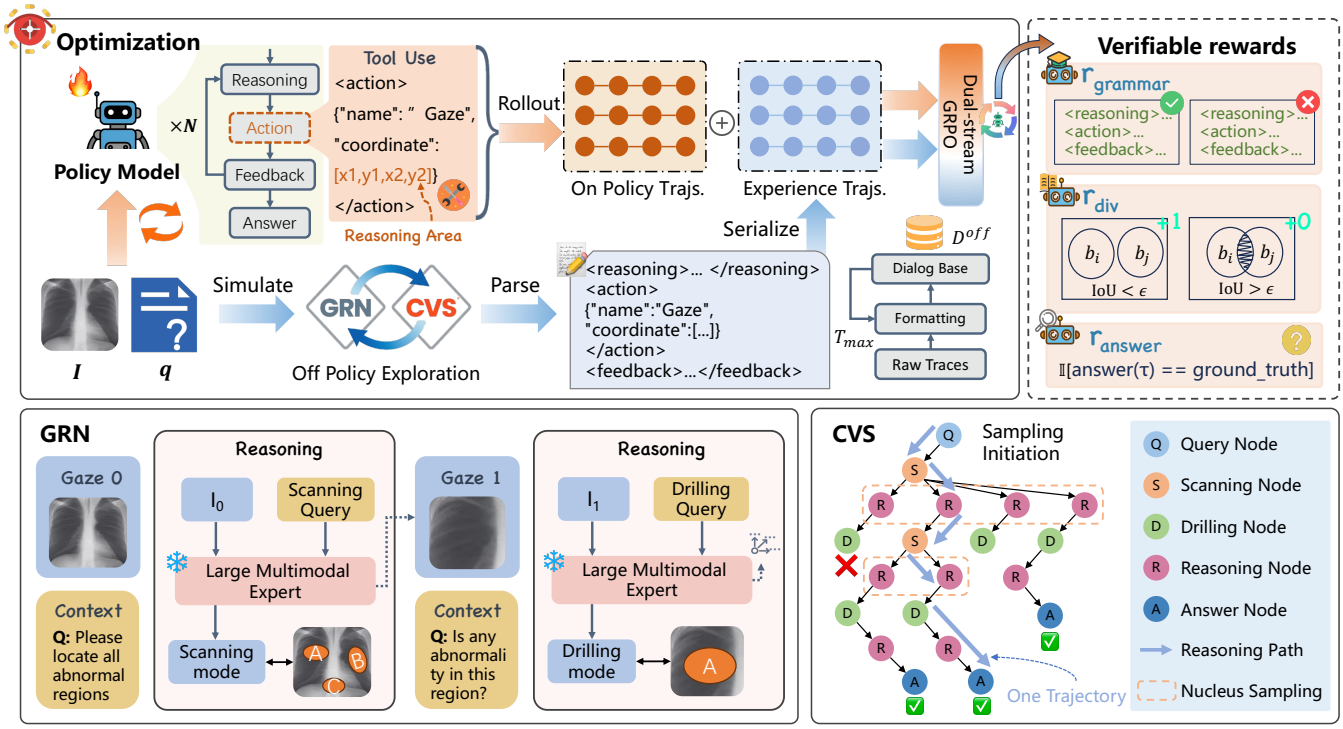

Figure 2: Overview of MedEyes. We first generate structured off-policy expert trajectories through the Gaze-guided Reasoning Navigator (GRN) and Confidence Value Sampler (CVS) to explore expert reasoning patterns, then combine these with on-policy rollouts from the policy model. The unified trajectory is subsequently for optimization via dual-stream GRPO with multi-component verifiable rewards to enhance the model's intrinsic grounding reasoning capability for medical visual understanding.

cally aligned progressive attention mechanisms that circumvent initialization bottlenecks. Progressive attention materializes through structured trajectory sequences embodying clinical diagnostic workflows: $\tau = \left[\prod_{i=1}^{T} (\theta_i \circ \alpha_i \circ o_i)\right] \circ \phi$, where concatenation operator $\circ$ links reasoning step θ_i, tool invocation α_i, and observational feedback o_i, terminating with final answer ϕ. Each reasoning-action-perception triplet $(\theta_i, \alpha_i, o_i)$ instantiates expert diagnostic protocols established in radiological interpretation studies. This architecture facilitates continuous visual-cognitive interaction, achieving diagnostic convergence through iterative grounding and reasoning refinement that mirrors expert clinician behavioral patterns documented in clinical practice.

2.2 Off-Policy Expert Trajectory Generation

To enhance policy learning from external behaviors, expert search protocols are systematically transformed into structured learning signals via the GRN for dual-mode exploration and the CVS for trajectory sampling diversification.

Gaze-guided Reasoning Navigator (GRN) Aligned with clinical eye-tracking protocols (Brunyé et al. 2019), GRN maintains ternary attention state $\psi_t = (\mathcal{R}_t, \mathcal{C}_t, \mathcal{F}_t)$ orchestrating visual exploration dynamics, where $\mathcal{R}_t$ constitutes candidate region sets generated via large-scale multimodal expert consultation through region-level VQA queries and $\langle$SEG$\rangle$ segmentation prompts, $\mathcal{C}_t$ contains corresponding confidence distributions, and $\mathcal{F}_t \in \{\text{global}, \text{local}\}$ specifies exploration modes. The state evolution follows:

$$\psi_{t+1} = \mathcal{T}(\psi_t, a_t, o_t), \tag{2}$$

where a_t is the action taken and o_t is the resulting observation. This transition controls how the model updates its attention state based on new visual diagnostic evidence.

Dual-mode Exploration Strategy The state transition $\mathcal{T}$ operates through two complementary exploration modes that reflect expert physicians' visual search patterns:

Scanning Mode: When $\mathcal{F}_t$ = global, GRN submits the prompt "Please locate all abnormal regions in the image $\langle$SEG$\rangle$" to the expert model, generating comprehensive candidate regions $\mathcal{R}_{t+1} = \{r_1, r_2, ..., r_n\}$ and populating the corresponding confidence scores $\mathcal{C}_{t+1}$ for the next state.

Drilling Mode: When $\mathcal{F}_t$ = local, GRN performs targeted analysis of the candidate regions from the current state $\mathcal{C}_t$ using the targeted diagnostic query, such as: "Please analyze the abnormality in region $\langle$region$\rangle r_i \langle$/region$\rangle$ $\langle$SEG$\rangle$". This detailed pathological analysis produces refined confidence scores that update $\mathcal{C}_{t+1}$ for the specific analyzed region r_i. The confidence evolution between consecutive states determines the subsequent exploration mode $\mathcal{F}_{t+1}$:

$$\Delta c = \frac{c_{t+1}(r_i) - c_t(r_i)}{c_t(r_i) + \epsilon}, \tag{3}$$

where $c_t(r_i) \in \mathcal{C}_t$ and $c_{t+1}(r_i) \in \mathcal{C}_{t+1}$ denote prior and refined confidence scores for region r_i, respectively, and ϵ

is a stability constant. If $\Delta c \geq \delta$, drilling mode persists. Otherwise, scanning mode resumes for broader exploration.

Confidence Value Sampler (CVS) To obtain diverse yet credible expert behaviors and capture diagnostic reasoning patterns, CVS applies nucleus sampling (Holtzman et al. 2019) to GRN's multi-round trajectories, generating multiple variable-length exploration paths. At each decision step t, the sampler selects from the top-p_0 confidence regions:

$$\mathcal{P}_{\text{nucleus}} = \left\{ a_i : \sum_{j=1}^{i} P(a_j \mid \psi_t) \leq p_0 \right\}, \tag{4}$$

where actions are ranked in descending probability order by $P(a_j \mid \psi_t)$ based on state ψ_t, with candidate region selection for $\mathcal{C}_{t+1}$ within the scanning-drilling exploration. Conditioned on GRN's region proposals, the CVS generates distinct trajectories $N_{\text{expert}} = \{\tau_1^{\text{expert}}, \tau_2^{\text{expert}}, ..., \tau_{N_{\text{expert}}}^{\text{expert}}\}$. Each trajectory sampling terminates upon satisfying convergence criteria: local confidence exceeds threshold ξ indicating diagnostic certainty, or maximum length $T_{\max}$ is reached.

Trajectory Parsing and Serialization. The raw visual exploration traces are then systematically parsed into structured dialog sequences through sequential decomposition. Each reasoning cycle is formatted using standardized tags: `<reasoning>...</reasoning>` for cognitive steps, followed by either `<action>"name": "Gaze", "coordinate": [x1,y1,x2,y2]</action>` for continued grounding operations with subsequent `<feedback>...</feedback>` containing coordinate-centered visual crops resized to target resolution, or direct termination with `<answer>...</answer>`. The serialization process converts these parsed trajectories into the multi-round format $(\theta_i, \alpha_i, o_i)$ while preserving temporal dependencies and spatial grounding relationships.

This mechanism captures diagnostic complexity variation across cases, where simple abnormalities enable immediate detection while complex scenarios demand extensive multi-region exploration. Each trajectory $\tau_k^{\text{expert}} = \{(s_1, \mathcal{G}_1), ..., (s_{T_k^*}, \mathcal{G}_{T_k^*}), a_k\}$ represents a complete reasoning path with adaptive length $T_k^* \leq T_{\max}$ determined by visual complexity rather than fixed templates. These structurally consistent yet content-diverse trajectories make up the off-policy replay buffer $\mathcal{D}^{\text{off}}$ for subsequent training.

2.3 Dual-stream GRPO Reinforcement Learning

Standard GRPO (Shao et al. 2024) employs a single advantage normalization across all trajectories, while our approach decouples the advantage computation for on-policy and off-policy data, avoiding gradient dominance that undermines mixed-policy training effectiveness.

Verifiable Reward Function The composite reward function $R(\tau) = \lambda_{\text{acc}} \cdot r_{\text{acc}}(\tau) + \lambda_{\text{grammar}} \cdot r_{\text{grammar}}(\tau) + \lambda_{\text{div}} \cdot r_{\text{div}}(\tau)$ comprises three components that address different aspects of medical visual reasoning: diagnostic accuracy, structural correctness, and exploration diversity.

Accuracy Reward Evaluates diagnostic correctness by comparing reasoning-generated answers with ground truth:

$$r_{\text{acc}}(\tau) = \mathbb{I}[\text{answer}(\tau) = \text{ground\_truth}], \tag{5}$$

where answer(τ) denotes the final answer extracted from the `<answer>...</answer>` tags from trajectory τ, and $\mathbb{I}(.)$ is an indicator function.

Grammar Reward Ensures multi-round reasoning structure correctness through format validation:

$$r_{\text{grammar}}(\tau) = \prod_{i=1}^{T} \mathbb{I}[W(\theta_i, \alpha_i, o_i)] \cdot \mathbb{I}[E(\tau)], \tag{6}$$

where $W(\theta_i, \alpha_i, o_i)$ validates each reasoning cycle format following the standardized tag structure defined in Section 2.2. The grammar reward is binary: $r_{\text{grammar}}(\tau) = 1$ if and only if all format validations succeed, and 0 otherwise, ensuring strict adherence to grammatical correctness.

Diversity Reward Incentivizes comprehensive multi-region visual exploration:

$$r_{\text{div}}(\tau) = \min\left(1, \frac{|\mathcal{U}(\tau)|}{n}\right) + \frac{1}{\binom{|\mathcal{U}(\tau)|}{2}} \sum_{i<j} \mathbb{I}[\text{IoU}(b_i, b_j) < \epsilon], \tag{7}$$

where $\mathcal{U}(\tau)$ denotes unique regions explored in trajectory τ, n matches GRN's region count limit, and b_i, b_j are bounding boxes from different tool calls. The first term rewards multiple distinct regions, while the second term encourages spatial diversity through sufficiently separated regions.

Optimization Objective We propose mixed-policy optimization that achieves progressive evolution from expert imitation to autonomous discovery by decoupling advantage normalization for on-policy exploration and off-policy guidance. Policy parameters θ are updated by maximizing:

$$\mathcal{J}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|\tau_i|} \sum_{t=1}^{|\tau_i|} \min\big(\rho_{i,t}^{\theta} A_i,\ \text{clip}(\rho_{i,t}^{\theta}, 1-\epsilon, 1+\epsilon)\, A_i\big), \tag{8}$$

where $N = N_{\text{on}} + N_{\text{off}}$ denotes total trajectory count. In our hybrid learning paradigm, we employ source-adaptive importance ratio computation: for on-policy trajectories, $\rho_i^{\theta} = \pi_{\theta}(\tau_i \mid I, q) / \pi_{\theta_{\text{old}}}(\tau_i \mid I, q)$ measures policy evolution from the previous iteration, while for off-policy expert trajectories, $\rho_i^{\theta} = \pi_{\theta}(\tau_i \mid I, q) / \pi_{\text{expert}}(\tau_i \mid I, q)$ where $\pi_{\text{expert}}(\tau_i \mid I, q) = 1$ represents the expert trajectory generation policy from CVS with nucleus sampling. This dual-stream design maintains gradient stability across heterogeneous trajectory sources while effectively mitigating potential distribution shifts between the respective signals.

Advantage Decoupling Mechanism We employ source-specific advantage normalization rather than unified normalization across all trajectories:

$$A_i = \frac{R(\tau_i) - \mu^{s(i)}}{\sigma^{s(i)} + \varepsilon}, \quad \text{where } s(i) = \begin{cases} \text{on}, & \text{if } \tau_i \in \mathcal{D}^{\text{on}} \\ \text{off}, & \text{if } \tau_i \in \mathcal{D}^{\text{off}} \end{cases}, \tag{9}$$

where $\mathcal{D}^{\text{on}}$ represents on-policy trajectories and $\mathcal{D}^{\text{off}}$ represents off-policy expert trajectories. The statistics $\mu^{\text{on/off}} = \mathbb{E}_{\tau \sim \mathcal{D}^{\text{on/off}}}[R(\tau)]$ and $\sigma^{\text{on/off}} = \sqrt{\text{Var}_{\tau \sim \mathcal{D}^{\text{on/off}}}[R(\tau)]}$ are computed independently on their respective data distributions. This decoupling strategy prevents expert trajectories from overshadowing on-policy advantages, avoiding gradient dominance that undermines autonomous learn-

Method	VQA-RAD	SLAKE	PathVQA	PMC-VQA	MMMU*	Average	Δ (%)
General Vision-Language Models							
Qwen2.5-VL-3B (Bai et al. 2025)	47.3	56.2	55.2	37.5	34.3	46.1	-19.8
GPT-4o (Hurst et al. 2024)	54.2	50.1	59.2	40.8	–	51.1	-14.8
InternVL-2 (Chen et al. 2024)	58.8	62.7	45.8	38.4	52.7	51.7	-14.2
Medical-Specific Models							
RadFM (Wu et al. 2023)	50.6	34.6	38.7	25.9	27.0	35.4	-30.5
MedVInT (Zhang et al. 2023)	45.4	43.5	54.7	23.3	28.3	39.0	-26.9
LLaVA-Med (Li et al. 2023)	51.4	48.6	56.8	24.7	36.9	43.7	-22.2
Med-Flamingo (Moor et al. 2023)	55.8	59.6	40.7	34.7	47.5	47.7	-18.2
GMAI-VL (Li et al. 2024)	64.6	71.9	47.2	52.3	51.2	57.4	-8.5
Reinforcement Learning Methods							
GRIT† (Fan et al. 2025)	54.3	57.1	43.5	42.3	49.5	49.3	-16.6
DeepEyes† (Zheng et al. 2025)	56.4	59.7	42.3	45.2	49.1	50.5	-15.4
Med-R1 (Lai et al. 2025)	55.9	55.1	53.3	45.8	32.7	48.5	-17.4
MedVLM-R1 (Pan et al. 2025)	61.4	65.9	55.2	44.8	35.5	52.5	-13.4
MedEyes (Ours)	**70.7**	**79.1**	**64.8**	**55.3**	**59.7**	**65.9**	–

Table 1: Comprehensive medical VQA benchmark performance comparison across five widely-used common open medical datasets. Best results are marked in bold. Δ indicates the performance gap (%) compared to our method.

ing. Separate normalization maintains distinct learning signals, enabling effective knowledge transfer while preserving model's reasoning adaptability to novel medical cases.

3 Experiments

3.1 Experimental Setup

Datasets We conduct comprehensive experiments on five widely adopted medical visual question answering benchmarks to evaluate the effectiveness of our method, including VQA-RAD (Lau et al. 2018), SLAKE (Liu et al. 2021), PathVQA (He et al. 2020), PMC-VQA (Zhang et al. 2023), and MMMU* (Yue et al. 2024) (* indicates the Health&Medicine subset). These datasets cover various medical imaging modalities, such as radiology (CT, MRI, X-ray), pathological slides, and multimodal medical scenarios. More details about the training datasets and statistics are provided in the supplementary materials.

Baselines We compare MedEyes against three categories of comprehensive baselines: (1) *General vision-language models*, including GPT-4o (Hurst et al. 2024), Gemini-Pro (Team et al. 2023), Qwen2.5-VL-3B (Bai et al. 2025), and InternVL-2 (Chen et al. 2024), which we use for their thinking ability to perform reasoning; (2) *Medical-specific VLMs*, including LLaVA-Med (Li et al. 2023), MedVInT (Zhang et al. 2023), Med-Flamingo (Moor et al. 2023), RadFM (Wu et al. 2023), GMAI-VL (Li et al. 2024), which leverage multimodal understanding for visual question answering; (3) *Reinforcement learning enhanced methods*, including visual CoT methods GRIT† (Fan et al. 2025) and DeepEyes† (Zheng et al. 2025) († denotes using the same dataset adapted to medical domains), and Med-R1 (Lai et al. 2025) and MedVLM-R1 (Pan et al. 2025), which leverage GRPO to generate explicit textual reasoning processes for enhanced diagnostic interpretability and reliability.

Implementation Details Our framework builds upon Qwen2.5-VL-3B (Bai et al. 2025), leveraging its strong multimodal capabilities. We adopt MedPLIB (Huang et al. 2025) as the sole visual expert to maximize task alignment. Within each reasoning step, the GRN processes $n{=}5$ candidate regions while maintaining established visual grounding formats. Mode transitions employ adaptive thresholds $\delta{=}0.15$ with stability constant $\epsilon{=}1{\times}10^{-6}$. CVS implements nucleus sampling ($p_0{=}0.9$) with termination when the confidence exceeds $\xi{=}0.85$ or reaches the maximum length $T_{\max}{=}4$, with max generation per turn set to 1024 tokens. This process generates $N_{\text{expert}}{=}6$ expert trajectories to form our off-policy replay buffer. The visual encoder processes images at a resolution of 336×336 with a patch size of 14. The dual-stream GRPO training process uses carefully balanced reward weights: $\lambda_{\text{acc}}{=}0.7$, $\lambda_{\text{grammar}}{=}0.2$, and $\lambda_{\text{div}}{=}0.1$. We optimized the model using AdamW with a learning rate of 1×10^{-6}, and conducted training over three epochs on six NVIDIA GeForce RTX 3090 GPUs. The training runs for 80 iterations with a rollout batch size of 98, generating 8 rollouts per prompt to ensure adequate exploration.

3.2 Main Results

Table 1 shows the comprehensive evaluation results of MedEyes across five medical VQA benchmarks. MedEyes achieves state-of-the-art performance with an average accuracy of 65.9, substantially outperforming all baseline categories. Specifically, it surpasses the best medical-specific model GMAI-VL by 8.5% and the strongest reinforcement learning method MedVLM-R1 by 13.4%. The superior performance demonstrates that vision-grounded chain-of-thought reasoning significantly outperforms purely textual approaches by establishing explicit visual-textual connections. On PathVQA, MedEyes achieves 64.8 accuracy, validating our dual-mode exploration strategy. On MMMU (Health & Medicine), MedEyes consistently exceeds all baselines, confirming the efficacy of dual-stream GRPO op-

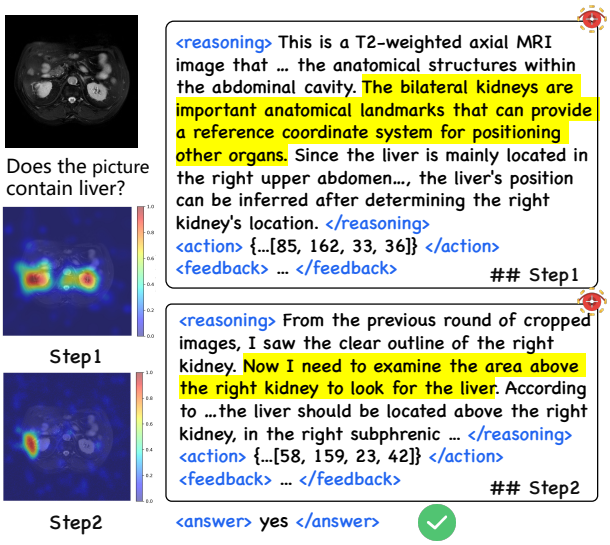

Figure 3: Diagnostic chain-of-thought example of MedEyes. Step 1 identifies bilateral kidneys as anatomical landmarks, followed by targeted liver in step 2. Heatmaps illustrate the progressive refinement process of visual attention.

timization. These results validate that off-policy expert trajectory guidance with interleaved visual-textual reasoning successfully overcomes pure on-policy learning limitations.

3.3 In-Depth Case Analysis

As illustrated in Fig. 3, the T2-weighted MRI liver detection case demonstrates MedEyes's progressive diagnostic reasoning. The framework initiates systematic anatomical localization through bilateral kidney identification as spatial reference coordinates, subsequently executing targeted exploration of the region adjacent to the right kidney. Cross-attention weight visualizations from the vision-language transformer adapter directly map reasoning tokens to corresponding image regions, revealing stepwise attention refinement from diffuse activation patterns to diagnostically salient areas. This progression validates the framework's capacity to internalize expert-level visual search hierarchies, wherein diagnostic confidence emerges through systematic evidence synthesis rather than pattern recognition heuristics.

3.4 Training Dynamics Analysis

Fig. 4 illustrates the training dynamics of MedEyes. The reward curve in (a) shows a steady increase during the training process, with the most significant improvement between steps 200 and 800, followed by stabilization. Compared to unguided training without off-policy trajectories, this validates the effectiveness of our dual-stream GRPO optimization in learning when to invoke visual tools, reflecting a balance between internal knowledge and external search. The average trajectory length (b) presents an interesting pattern: initially increasing from 2.1 steps to approximately 3.0 steps (exploration phase), then gradually decreasing to 2.6 steps (efficiency phase). This trajectory compression indicates that MedEyes learned to generate more concise reasoning chains while maintaining accuracy, directly reflecting the model's internalization of when visual grounding is critical and when internal knowledge is sufficiently adequate.

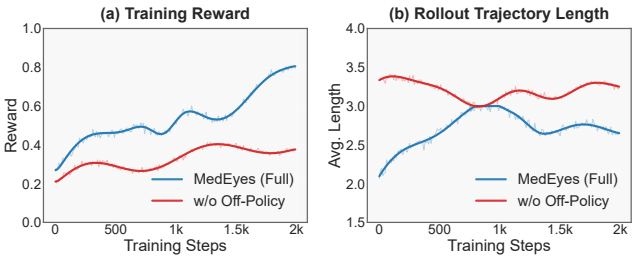

Figure 4: Training dynamics of MedEyes. (a) Reward progression highlighting the effectiveness of off-policy expert guidance. (b) Trajectory length showing exploration-efficiency transition in multi-round visual reasoning.

Components	VQA-RAD	SLAKE	PathVQA	Average
MedEyes (Full)	**70.7**$_{\pm 1.2}$	**79.1**$_{\pm 0.9}$	**64.8**$_{\pm 1.3}$	**71.5**
Main Component Ablation				
✗ GRN	62.4$_{\pm 1.4}$	69.8$_{\pm 1.1}$	56.2$_{\pm 1.5}$	62.8
✗ CVS	65.3$_{\pm 1.3}$	73.5$_{\pm 1.0}$	59.1$_{\pm 1.4}$	66.0
✗ Off-policy	61.2$_{\pm 1.5}$	67.4$_{\pm 1.2}$	54.3$_{\pm 1.6}$	61.0
GRN Design Variants				
Scanning-only	66.8$_{\pm 1.2}$	74.2$_{\pm 0.9}$	58.7$_{\pm 1.3}$	66.6
Drilling-only	64.5$_{\pm 1.3}$	71.9$_{\pm 1.0}$	60.3$_{\pm 1.2}$	65.6

Table 2: Ablation study of core components on medical VQA benchmarks. All values are reported with 95% confidence intervals. GRN: Gaze-guided Reasoning Navigator, CVS: Confidence Value Sampler.

3.5 Ablation Studies

Core Component Analysis Table 2 validates our framework's core mechanisms through systematic ablation. GRN removal results in an 8.7% performance drop, confirming that scanning-drilling strategy successfully replicates expert diagnostic workflows from systematic region identification to recursive pathological analysis. Single-mode variants demonstrate incomplete reasoning: scanning-only fails on fine-grained tasks while drilling-only lacks systematic exploration, proving both modes essential for complete medical visual reasoning. CVS removal leads to a 5.5% degradation, validating its effectiveness in generating credible exploration trajectories. Most critically, off-policy learning removal yields 10.5% degradation, directly confirming that expert trajectories provide indispensable cognitive anchors for breaking through autonomous exploration limitations.

Off-Policy Trajectory Configuration Table 3 offers important insights into how medical reasoning patterns are captured. Increasing the number of expert trajectories leads to substantial performance improvements from 2 to 6 trajectories, after which further increases provide only marginal gains, suggesting that 6 trajectories are sufficient to encode the diversity of expert diagnostic strategies. Likewise, a rea-

Config.	Value	VQA-RAD	SLAKE	PathVQA	Average
N_{expert}	2	64.2	71.8	57.3	64.4
	4	66.8	74.5	59.8	67.0
	6	**70.7**	**79.1**	**64.8**	**71.5**
	8	70.3	78.7	64.3	71.1
T_{max}	2 steps	70.5	78.8	64.3	71.2
	3 steps	**70.7**	**79.1**	**64.8**	**71.5**
	4 steps	70.6	78.2	63.9	70.9
	5 steps	68.1	76.5	62.4	69.0

Table 3: Analysis of expert trajectory count and length on model performance. Gray background and bold text indicate the selected and optimal configurations, respectively.

Config.	VQA-RAD	SLAKE	PathVQA	Average
Reward Weight Configuration ($\lambda_{acc}/\lambda_{grammar}/\lambda_{div}$)				
1.0/0.0/0.0	68.4	76.2	62.1	68.9
0.8/0.2/0.0	69.5	77.8	63.4	70.2
0.8/0.1/0.1	69.9	78.3	63.8	70.7
0.7/0.2/0.1	**70.7**	**79.1**	**64.8**	**71.5**
0.6/0.2/0.2	69.8	78.5	64.2	70.8
0.5/0.3/0.2	68.3	77.1	63.0	69.5

Table 4: Effect of our reward function design. Configurations selected for final use are marked with gray shading, and optimal values are shown in bold.

soning sequence length of 3 steps achieves the best performance, whereas longer sequences tend to accumulate errors, resulting in decreased accuracy. These findings highlight the balance between trajectory quantity and sequence length in effectively modeling expert reasoning.

Reward Function Design Table 4 presents a detailed multi-component reward analysis, showing that optimizing for accuracy alone fails to capture structured diagnostic reasoning. Incorporating both grammar and diversity produces coherent reasoning–action–perception cycles. Grammar enforces strict syntactic grounding, while diversity promotes broader and more comprehensive exploration of diagnostically relevant regions. The optimal rewards configuration further encourages thorough and fully comprehensive exploration of diagnostically relevant regions.

Off-Policy Sampling Analysis Table 5 analyzes the impact of different off-policy trajectory sources. We compare random sampling, external trajectories from DeepSeek-R1, and historical replay, which selects 8 samples based on either recency (most recent) or reward (highest-reward). Results show that while all alternatives consistently demonstrate the benefits of off-policy learning, trajectory quality remains crucial for achieving optimal performance.

3.6 Failure Case Study

Fig. 5 demonstrates two representative error scenarios. Quantitative measurement tasks suffer from flawed relative scaling that substitutes anatomical proportion estimation for pixel-to-centimeter calibration, introducing variance incompatible with clinical requirements. Additionally, MedEyes faces challenges in fine-grained concept differentiation, occasionally conflating pathological subtypes with shared morphological features, such as "aneurysmal changes" versus "dissecting aneurysms". These limitations underscore the necessity of richer tools and deeper expertise understanding for future medical agents.

Strategy/Config	VQA-RAD	SLAKE	PathVQA	Average
GRN+CVS	**70.7**	**79.1**	**64.8**	**71.5**
Baseline Sampling Methods				
Random	58.3	65.7	52.1	58.7
DeepSeek-R1	65.5	71.2	53.9	63.9
Historical Replay Strategies				
Recency-based	63.2	70.8	56.4	63.5
Reward-oriented	65.1	72.9	58.7	65.6

Table 5: Off-policy trajectory quality and sampling strategy analysis. Gray shading indicates selected configurations.

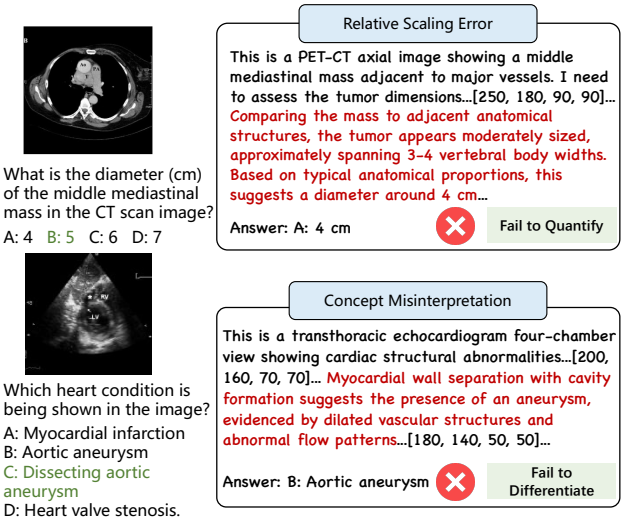

Figure 5: Failure cases analysis. quantitative measurement errors in tumor sizing (top) and pathological concept misinterpretation in ultrasound aneurysm identification (bottom).

4 Conclusion

This paper presents MedEyes, a novel medical visual Chain-of-Thought (CoT) framework that enables diagnostic evolution through dynamic focus learning. By combining the scanning-drilling dual-mode exploration of the gaze-guided reasoning navigator with end-to-end reinforcement learning using dual-stream GRPO, MedEyes effectively replicates the progressive visual focusing process of expert clinicians. Experimental results show that analysis of MedEyes's dynamic focus trajectories reveals a clear shift from casual exploration to efficient visual utilization, and that it substantially outperforms existing RL methods on multiple medical VQA benchmarks. These findings demonstrate that the system can autonomously develop effective and reliable visual reasoning and offer a novel technical pathway toward building agent-driven, reasoning-enabled medical frameworks.

Acknowledgments

This research was partially supported by grants from the National Natural Science Foundation of China (Grants No. 62472157, No. 62202158, No. 62206089), and the Science and Technology Innovation Program of Hunan Province (Grant No. 2023RC3098).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Brunyé, T. T.; Drew, T.; Weaver, D. L.; and Elmore, J. G. 2019. A review of eye tracking for understanding and improving diagnostic interpretation. *Cognitive research: principles and implications*, 4(1): 7.

Castner, N.; Kuebler, T. C.; Scheiter, K.; Richter, J.; Eder, T.; Hüttig, F.; Keutel, C.; and Kasneci, E. 2020. Deep semantic gaze embedding and scanpath comparison for expertise classification during OPT viewing. In *ACM symposium on eye tracking research and applications*, 1–10.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Fan, Y.; He, X.; Yang, D.; Zheng, K.; Kuo, C.-C.; Zheng, Y.; Narayanaraju, S. J.; Guan, X.; and Wang, X. E. 2025. GRIT: Teaching MLLMs to Think with Images. *arXiv preprint arXiv:2505.15879*.

He, X.; Zhang, Y.; Mou, L.; Xing, E.; and Xie, P. 2020. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Huang, X.; Shen, L.; Liu, J.; Shang, F.; Li, H.; Huang, H.; and Yang, Y. 2025. Towards a multimodal large language model with pixel-level insight for biomedicine. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 3779–3787.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

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*.

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

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

Li, T.; Su, Y.; Li, W.; Fu, B.; Chen, Z.; Huang, Z.; Wang, G.; Ma, C.; Chen, Y.; Hu, M.; et al. 2024. GMAI-VL & GMAI-VL-5.5 M: A Large Vision-Language Model and A Comprehensive Multimodal Dataset Towards General Medical AI. *arXiv preprint arXiv:2411.14522*.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, 1650–1654. IEEE.

Lu, M. Y.; Chen, B.; Williamson, D. F.; Chen, R. J.; Zhao, M.; Chow, A. K.; Ikemura, K.; Kim, A.; Pouli, D.; Patel, A.; et al. 2024. A multimodal generative AI copilot for human pathology. *Nature*, 634(8033): 466–473.

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.

OpenAI. 2025. Thinking with Images. https://openai.com/index/thinking-with-images/. Accessed: 2025-07-08; Release date: April 16, 2025.

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. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 337–347. Springer.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sultana, J.; Qin, R.; and Yin, Z. 2024. Seeing Through Expert's Eyes: Leveraging Radiologist Eye Gaze and Speech Report with Graph Neural Networks for Chest X-ray Image Classification. In *Proceedings of the Asian Conference on Computer Vision*, 2579–2595.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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 e-prints*, arXiv–2308.

Xia, P.; Zhu, K.; Li, H.; et al. 2025. MMed-RAG: Versatile Multimodal RAG System for Medical Vision Language Models. In *ICLR*.

Yu, H.; Cheng, T.; Cheng, Y.; and Feng, R. 2025. FineMedLM-o1: Enhancing the Medical Reasoning Ability of LLM from Supervised Fine-Tuning to Test-Time Training. *arXiv preprint arXiv:2501.09213*.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; 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*, 9556–9567.

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*.

Zheng, Z.; Yang, M.; Hong, J.; Zhao, C.; Xu, G.; Yang, L.; Shen, C.; and Yu, X. 2025. DeepEyes: Incentivizing" Thinking with Images" via Reinforcement Learning. *arXiv preprint arXiv:2505.14362*.

Zhuang, B.; Song, C.; Lu, H.; Qiao, J.; Liu, M.; Yu, M.; Hong, P.; Li, R.; Song, X.; Xu, X.; et al. 2025. WiNGPT-3.0 Technical Report. *arXiv preprint arXiv:2505.17387*.