NOISE-ROBUST REINFORCEMENT LEARNING WITH VERIFIABLE REWARDS FOR OPEN-ENDED MEDICAL MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Reinforcement learning with verifiable rewards (RLVR) has shown strong potential for improving the reasoning ability of multimodal large language models. However, most existing RLVR methods assume reliable labels for reward computation, making them vulnerable to annotation noise. This challenge is particularly prevalent in medical domains, where expert annotations are costly and often affected by diagnostic uncertainty, inter-observer variability, and complex clinical interpretation. In this work, we investigate noise-robust RLVR for open-ended medical multimodal reasoning and propose **N**oise-**R**obust **RLVR** (NR-RLVR), a framework that dynamically refines supervision signals during training. NR-RLVR consists of two key components. First, **R**eference-aware **P**seudo-label **G**eneration (RPG) improves pseudo-label reliability by aggregating predictions from multiple model rollouts and a reference-guided reasoning trajectory. Second, **C**onfidence-gated **S**upervision **R**efinement (CSR) resolves conflicts between reference labels and pseudo-labels by selecting the more reliable supervision signal according to model confidence. Extensive experiments on text-only and vision-language medical benchmarks show that NR-RLVR consistently improves reasoning accuracy and robustness across different levels of label noise. Notably, under the challenging 90% noise setting, NR-RLVR improves the average performance of Qwen2.5-VL-3B by 7.47% over standard GRPO on vision-language medical benchmarks. Code is available in the supplementary material.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has recently emerged as a promising paradigm for enhancing the reasoning capabilities of multimodal large language models (MLLMs) Zheng et al. (2025b); Ma et al. (2026); Zhang et al. (2025). Representative approaches such as GRPO Shao et al. (2024); Guo et al. (2025a) and its variants Zheng et al. (2025a); Yu et al. (2026c) leverage multiple sampled reasoning trajectories and compute rewards through verifiable rules, enabling policy optimization without critic models Schulman et al. (2017). RLVR effectively improves reasoning performance while reducing the complexity of traditional pipelines.

Despite this progress, a critical assumption underlying most RLVR methods is the reliability of labels used for reward computation Mullappilly et al. (2026); Yu et al. (2026b). In practice, this assumption is often violated, especially in medical scenarios, where expert annotations are costly and often affected by diagnostic uncertainty, inter-observer variability, and complex clinical interpretation Guo et al. (2025b); Wei et al. (2024). As shown in Figure 1(a), unreliable labels lead to inaccurate reward estimation and misguide policy optimization, resulting in degraded reasoning performance. Existing approaches to learning with noisy labels have been extensively studied in classical discriminative tasks, including classification, detection, and segmentation Collier et al. (2021); Hendrycks et al. (2018); Li & Zhu (2024). However, these methods are primarily designed for discriminative settings with predefined label sets and relatively constrained output spaces, limiting their direct applicability to generative reasoning tasks.

In this work, we focus on a more challenging and realistic setting: *how to enable effective RLVR of MLLMs on open-ended medical question answering data with noisy supervision*. Recent stud-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

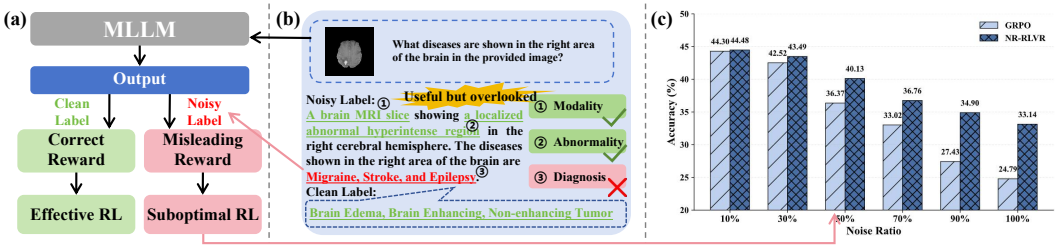

Figure 1: (a) Noisy annotations introduce misleading reward signals and hinder RLVR training of medical MLLMs. (b) Medical annotation noise is often structured, where partial clinical information is preserved, while high-level semantic attributes such as diagnostic conclusions are more vulnerable to annotation errors. (c) NR-RLVR consistently improves average accuracy across multimodal medical benchmarks under different levels of annotation noise.

ies Yang et al. (2026); Xu et al. (2026) on noisy RLVR setting primarily identify unreliable supervision through statistical signals, such as sample uncertainty or entropy-based heuristics. However, these approaches largely overlook the valuable information contained in reference labels. In real-world medical scenarios, annotation noise is often not random corruption but exhibits structured semantic deviations Abdalla & Fine (2023); Maier-Hein et al. (2018); Warfield et al. (2004); Irvin et al. (2019); Wang et al. (2017). As shown in Figure 1(b), even noisy medical annotations may preserve reliable clinical information, such as imaging modality and anatomical details, while high-level semantic attributes, including diagnostic conclusions, are more vulnerable to annotation uncertainty and expert disagreement. This suggests that noisy annotations may provide useful information for model reasoning. Moreover, existing methods fail to exploit model-generated reasoning trajectories for supervision refinement. Beyond statistical cues derived from model behaviors, reasoning trajectories provide complementary semantic evidence that can be leveraged to identify and refine noisy supervision. Meanwhile, most existing methods are developed on closed-form or restricted-output tasks, such as multiple-choice question answering and short-form numerical prediction, where candidate outputs have explicit forms and can be directly compared. In contrast, open-ended medical reasoning requires the generation of free-form responses with diverse linguistic expressions, making label reliability assessment and supervision refinement substantially more challenging.

Based on this insight, we construct a noisy open-ended medical question answering training dataset based on existing public datasets, and propose a **N**oise-**R**obust **RLVR** (NR-RLVR) framework that jointly improves pseudo-label quality and dynamically refines supervision during training. First, we design a **R**eference-aware **P**seudo-label **G**eneration (RPG) mechanism. Specifically, pseudo-labels are constructed via majority voting over multiple rollouts of the base model. To further improve pseudo-label quality, we inject reference label information into the context of a reasoning trajectory, enabling reference-guided reasoning without compromising the diversity of rollouts. Second, we introduce a **C**onfidence-gated **S**upervision **R**efinement (CSR) strategy. When pseudo-labels conflict with reference annotations, we reformulate the original multimodal question as a binary selection problem between the two candidate labels. The model produces a structured output consisting of reasoning trajectories, a final answer, and a confidence score, which is used to determine whether the pseudo-label should replace the reference label during reward computation. In this way, NR-RLVR can leverage useful reference information and refine corrupted labels throughout the training.

We evaluate NR-RLVR on multiple medical benchmarks covering both text-only and vision-language tasks under varying levels of label noise. As shown in Figure 1(c), our method consistently improves reasoning accuracy and robustness compared with standard RLVR training, and the gains are especially evident under higher noise ratios. Our contributions are summarized as follows:

- We investigate RLVR under noisy supervision for open-ended medical multimodal reasoning, highlighting a practical and challenging setting where reference labels are partially corrupted, causing misleading rewards and biased policy updates during optimization.
- We propose RPG, a reference-aware pseudo-label generation mechanism that leverages majority voting over model rollouts and reference-guided reasoning to construct more reliable supervision candidates.

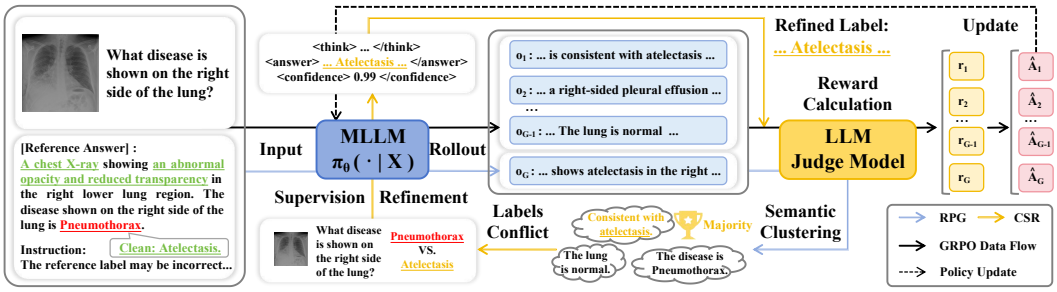

Figure 2: Illustration of the NR-RLVR framework. Given a medical multimodal question, the policy model generates multiple rollouts and one reference-guided trajectory. The Reference-aware Pseudo-label Generation (RPG) module leverages majority voting to obtain reliable pseudo-labels. When conflicts occur between pseudo-labels and reference annotations, the Confidence-gated Supervision Refinement (CSR) module reformulates the task as a binary preference decision and dynamically selects the more reliable supervision signal based on model confidence for reward computation.

- We propose CSR, a confidence-gated supervision refinement strategy that resolves conflicts between reference labels and pseudo-labels by selecting more reliable supervision signals for reward computation during policy optimization.
- We conduct extensive experiments on diverse medical text-only and vision-language benchmarks with different base models and RLVR algorithms, demonstrating that NR-RLVR consistently improves robustness and performance under various noise conditions.

2 Related Work

Reinforcement Learning with Verifiable Rewards. RLVR has emerged as an effective paradigm for enhancing the reasoning capabilities of MLLMs Yang et al. (2025b); Yao et al. (2025); Yu et al. (2026a). Taking GRPO Guo et al. (2025a); Shao et al. (2024) as a representative approach, it computes relative advantages based on verifiable rewards within each group, eliminating the need for an additional critic model Schulman et al. (2017). Subsequent studies have explored improving RLVR from different perspectives. DAPO Yu et al. (2026c) introduces several training strategies, including adaptive clipping, dynamic sampling, token-level policy optimization, and length-aware reward shaping, to improve training efficiency and stability. Dr.GRPO Liu et al. (2025) analyzes the limitations of standard GRPO and proposes a modified optimization objective by removing length normalization and variance normalization, alleviating biased optimization behaviors caused by response length and reward variance. Despite the remarkable success, existing approaches generally assume that the supervision used for reward computation is reliable Zheng et al. (2025b). However, in real-world scenarios, labels may inevitably contain noise, resulting in misleading rewards and suboptimal optimization. In this work, we investigate the underexplored problem of RLVR under noisy supervision and propose a framework that dynamically refines supervision during training.

Learning with noisy supervision. Existing learning with noisy supervision approaches have explored various strategies, including robust loss functions, co-training, and sample selection, to alleviate the negative impact of corrupted labels Reed et al. (2014); Han et al. (2018); Zhang et al. (2016). However, these methods are mainly designed for discriminative tasks with predefined output spaces, making them difficult to directly apply to generative reasoning scenarios. Recent studies have also explored RLVR under noisy supervision. However, existing approaches mainly focus on identifying unreliable supervision through statistical signals Yang et al. (2026) or optimizing heuristic uncertainty-based objectives Xu et al. (2026), while underutilizing the potentially valuable information contained in noisy reference labels. Moreover, these methods are primarily designed for closed-form or restricted-output tasks, such as multiple-choice or short-answer question answering, leaving robust RLVR for open-ended reasoning largely unexplored. In contrast, our framework exploits information from noisy references to construct higher-quality pseudo-labels and leverages model confidence to dynamically refine reward signals for open-ended medical reasoning.

Medical Multimodal Large Language Models. The rapid advancement of MLLMs has stimulated extensive research on their adaptation to medical scenarios Tian et al. (2024); AlSaad et al. (2024). Early medical MLLMs, such as LLaVA-Med Li et al. (2023), Med-Flamingo Moor et al. (2023), and BioMedGPT Luo et al. (2023), align medical visual encoders with large language models through lightweight projection modules, enabling multimodal medical understanding for tasks such as visual question answering and report generation Ning et al. (2025); Sun et al. (2026); Zhang et al. (2024). Subsequent studies have further improved medical MLLMs through larger-scale multimodal instruction datasets Hamamci et al. (2024); Ikezogwo et al. (2023); Li et al. (2026), advanced training strategies Wang et al. (2024); Nath et al. (2025), and parameter-efficient fine-tuning Lin et al. (2025), supporting diverse medical modalities and clinical applications. Recent efforts have explored RLVR strategies to enhance model reasoning capabilities Pan et al. (2025); Lai et al. (2026); Mullappilly et al. (2026). Despite these advances, existing RLVR methods largely rely on reliable reference supervision for reward computation, leaving their robustness under realistic noisy annotation scenarios insufficiently explored.

3 METHODOLOGY

3.1 PROBLEM DEFINITION AND PRELIMINARY

We consider the problem of RLVR under noisy supervision for open-ended medical multimodal reasoning. Given a training dataset $\mathcal{D} = \{(q_i, v_i, y_i)\}_{i=1}^{N}$, where (q_i, v_i) denotes a medical multimodal question and y_i denotes the reference label, we assume the existence of an ideal clean response y_i^*. Due to annotation noise, the observed label satisfy:

$$y_i = \begin{cases} y_i^*, & \text{with probability } 1-\rho, \\ \tilde{y}_i, & \text{with probability } \rho, \end{cases} \tag{1}$$

where ρ denotes the annotation noise ratio and $\tilde{y}_i$ represents a corrupted but semantically plausible response.

As a representative RLVR algorithm, GRPO Shao et al. (2024); Guo et al. (2025a) eliminates the need for an additional critic model by estimating response advantages from relative rewards across multiple responses sampled for the same query. Given a multimodal query $X = (q_i, v_i)$, GRPO samples a group of responses $\{o_g\}_{g=1}^{G}$ from the old policy $\pi_{\theta_{\text{old}}}$ and obtains corresponding rewards $\{r_g\}_{g=1}^{G}$. The policy optimization objective is as follows:

$$\begin{aligned} \mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{X,\{o_g\}_{g=1}^{G}\sim\pi_{\theta_{\text{old}}}} \Bigg[\frac{1}{G}\sum_{i=1}^{G} \min\Bigg(& \frac{\pi_\theta(o_g|X)}{\pi_{\theta_{\text{old}}}(o_g|X)}\hat{A}_g, \\ & \text{clip}\Big(\frac{\pi_\theta(o_g|X)}{\pi_{\theta_{\text{old}}}(o_g|X)}, 1-\epsilon, 1+\epsilon\Big)\hat{A}_g \Bigg) \Bigg], \end{aligned} \tag{2}$$

where ϵ denotes the clipping coefficient, and the relative advantage is computed as:

$$\hat{A}_g = \frac{r_g - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{3}$$

The reward is composed of an accuracy reward and a format compliance reward:

$$r = r_{\text{acc}} + r_{\text{format}}, \tag{4}$$

where r_{acc} measures semantic consistency between response and supervision, and r_{format} ensures that the reasoning process and final answer adhere to the prescribed output format, *i.e.* , enclosed within `<think></think>` and `<answer></answer>` tags, respectively.

3.2 NOISY MEDICAL DATASET CONSTRUCTION

To evaluate the robustness of RLVR under different levels of annotation noise, we construct noisy medical training data based on the publicly available MediX-RL dataset Mullappilly et al. (2026).

Specifically, we employ Qwen3-VL-30B-A3B Bai et al. (2025a) as a noise generation model, which takes the original multimodal question and answer as inputs and aims to simulate realistic annotation noise in medical scenarios. The complete noise generation prompt, including the instructions and constraints for constructing corrupted medical answers, is provided in the supplementary material.

Different from random perturbations that generate irrelevant or meaningless responses, realistic medical annotation noise should preserve partial semantic consistency while providing misleading supervision signals. Therefore, we generate noisy answers following three principles: **1) Structural consistency.** The generated answers preserve similar response structures, lengths, and linguistic styles as the original answers, preventing noise detection based on superficial patterns. **2) Medical plausibility.** The generated responses maintain clinically reasonable descriptions and reasoning processes while avoiding obvious medical contradictions. **3) Critical fact perturbation.** The noise is introduced by selectively modifying critical medical concepts, such as diagnostic conclusions, disease categories, or pathological interpretations, while preserving the original answer format and linguistic style. This enables the generated reference labels to provide plausible yet misleading supervision signals, better reflecting real-world annotation errors.

After noise generation, we conduct multiple rounds of automatic filtering and human verification to ensure that the generated noisy answers are semantically inconsistent with the original annotations while remaining medically plausible. By controlling the proportion of corrupted samples, we construct training sets with different noise ratios to evaluate the robustness of our framework as shown in Figure 2 under various noisy supervision scenarios.

3.3 Reference-aware Pseudo-label Generation

Given multimodal data (q_i, v_i, y_i), we first generate reliable pseudo-labels through self-consistent rollouts. Specifically, we sample $G-1$ reasoning trajectories $\{o_g\}_{g=1}^{G-1}$ from the base policy π_θ, where each trajectory contains a reasoning process and a final answer. Since open-ended medical answers may exhibit diverse linguistic expressions, directly matching generated strings is insufficient. Therefore, we reuse the LLM-based reward model introduced in accuracy reward computation to cluster generated answers according to semantic consistency. The detailed clustering prompt is provided in the supplementary material, and the clustering results are returned in a structured JSON format. Let $\mathcal{C}_i = \{C_i^1, \ldots, C_i^K\}$ denote the semantic clusters obtained from the rollouts. The dominant cluster is selected as:

$$k^* = \arg\max_k |C_i^k|, \tag{5}$$

and the corresponding normalized conclusion is used as the pseudo-label:

$$\hat{y}_i = \mathrm{Norm}(C_i^{k^*}), \tag{6}$$

where $\mathrm{Norm}(\cdot)$ generates the normalized medical conclusion of the selected cluster.

In realistic medical scenarios, noisy annotations are not purely random corruptions but often preserve partial clinical information, which can provide useful contextual cues to support the model medical reasoning process. However, incorporating reference labels into the majority of rollouts may bias the exploration process, reduce exploration diversity, and introduce a mismatch between training and inference scenarios. To balance reference utilization and exploration diversity, we introduce a Reference-aware Pseudo-label Generation (RPG) strategy. Specifically, we generate one additional reasoning trajectory conditioned on the reference label:

$$o_r \sim \pi_\theta(\cdot|q_i, v_i, y_i), \tag{7}$$

where the reference label is provided as an auxiliary reasoning clue with an uncertainty instruction. The reference-guided trajectory is combined with regular rollouts:

$$\mathcal{O}_i = \{o_1, \ldots, o_{G-1}, o_r\}, \tag{8}$$

and participates in clustering process. By introducing the reference-guided trajectory, RPG improves pseudo-label reliability while preserving sufficient exploration diversity.

3.4 Confidence-gated Supervision Refinement

Following RPG, each sample $X_i = (q_i, v_i)$ is associated with a reference label y_i and a pseudo-label $\hat{y}_i$. First, we determine whether the two candidate supervision signals are semantically consistent:

$$\delta_i = \mathrm{Sem}(y_i, \hat{y}_i), \tag{9}$$

Table 1: Results on Qwen2.5-VL-3B and Qwen3-VL-4B under different noise ratios. Each subscript indicates the improvement compared with naive GRPO. **Bold** indicates the best performance.

Method	Text-only benchmarks							Vision-language benchmarks					Avg.
	Anatomy	Genetics	MedMCQA	Bio	ProfMed	Clinical	Avg.	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.	
Qwen2.5-VL-3B	57.78	71.00	45.21	65.97	58.46	66.42	60.81	14.50	22.62	20.40	42.40	24.98	46.48
TTRL	58.52	67.00	46.86	69.44	55.15	62.26	59.87	9.60	36.78	22.40	38.40	26.80	46.64
$\text{GRPO}_{10\%}$	65.93	73.00	49.92	76.39	62.87	72.45	66.76	21.30	56.31	39.80	59.80	44.30	57.78
Two-Stage	67.41	73.00	50.59	71.53	60.66	70.57	65.63	21.80	55.43	40.40	59.80	44.36	57.12
NR-RLVR	66.67	78.00	48.60	75.69	64.71	71.32	**67.50**$_{+0.74}$	21.40	56.10	40.80	59.60	**44.48**$_{+0.18}$	**58.29**$_{+0.51}$
$\text{GRPO}_{30\%}$	63.70	75.00	49.46	79.86	58.46	69.06	65.92	21.30	52.99	38.40	57.40	42.52	56.56
Two-Stage	65.19	74.00	48.94	73.61	61.76	70.94	65.74	19.90	54.99	39.60	58.00	43.12	56.69
NR-RLVR	65.19	76.00	49.13	75.00	63.60	70.19	**66.52**$_{+0.60}$	20.00	54.77	40.20	59.00	**43.49**$_{+0.97}$	**57.31**$_{+0.75}$
$\text{GRPO}_{50\%}$	63.70	72.00	49.22	75.69	59.56	71.32	65.25	20.60	41.68	33.00	50.20	36.37	53.70
Two-Stage	62.96	73.00	48.14	75.00	60.29	69.06	64.74	19.70	41.46	33.40	51.60	36.54	53.46
NR-RLVR	64.44	76.00	48.89	76.39	62.13	70.94	**66.47**$_{+1.22}$	20.00	48.12	36.80	55.60	**40.13**$_{+3.76}$	**55.93**$_{+2.23}$
$\text{GRPO}_{70\%}$	62.96	73.00	48.55	73.61	58.82	69.06	64.33	20.20	39.69	31.80	40.40	33.02	51.81
Two-Stage	60.74	75.00	47.67	70.83	62.50	69.06	64.30	19.30	39.91	31.20	41.80	33.05	51.80
NR-RLVR	64.44	77.00	48.51	75.00	64.34	67.55	**66.14**$_{+1.81}$	20.30	44.35	34.20	48.20	**36.76**$_{+3.74}$	**54.39**$_{+2.58}$
$\text{GRPO}_{90\%}$	60.74	71.00	47.02	72.22	58.82	65.66	62.58	12.50	35.03	31.40	30.80	27.43	48.52
Two-Stage	56.30	68.00	47.62	68.06	54.78	62.64	59.57	11.70	35.48	31.60	30.80	27.40	46.70
NR-RLVR	61.48	75.00	48.63	74.31	61.03	68.68	**64.86**$_{+2.28}$	19.00	40.58	33.00	47.00	**34.90**$_{+7.47}$	**52.87**$_{+4.35}$
Qwen3-VL-4B	71.11	82.00	58.19	89.58	78.68	77.74	76.22	20.80	50.33	31.40	56.80	39.83	61.66
TTRL	65.19	78.00	57.59	87.50	76.47	73.21	72.99	21.10	46.12	24.80	53.00	36.26	58.30
$\text{GRPO}_{10\%}$	69.63	87.00	59.17	90.28	81.25	77.36	77.45	23.00	52.32	37.40	59.40	43.03	63.68
Two-Stage	69.63	82.00	55.61	85.42	76.84	77.74	74.54	23.40	52.11	37.60	62.00	43.78	62.24
NR-RLVR	71.11	87.00	59.34	91.67	80.15	79.25	**78.09**$_{+0.64}$	23.20	53.66	38.00	60.40	**43.82**$_{+0.79}$	**64.38**$_{+0.70}$
$\text{GRPO}_{30\%}$	68.89	84.00	58.54	90.97	80.88	78.11	76.90	22.50	52.11	37.20	59.00	42.70	63.22
Two-Stage	67.41	83.00	55.82	84.72	75.74	76.60	73.88	23.20	51.66	37.80	59.40	43.02	61.54
NR-RLVR	70.37	88.00	58.71	92.36	79.78	78.49	**77.95**$_{+1.05}$	22.60	53.44	38.60	59.20	**43.46**$_{+0.76}$	**64.16**$_{+0.94}$
$\text{GRPO}_{50\%}$	67.41	85.00	59.72	90.28	80.15	77.36	76.65	22.30	51.22	36.40	57.40	41.83	62.72
Two-Stage	67.41	83.00	59.07	88.19	77.94	76.98	75.43	21.70	52.11	35.60	57.80	41.80	61.98
NR-RLVR	71.85	86.00	59.67	89.58	80.51	79.25	**77.81**$_{+1.16}$	22.40	52.99	38.40	59.20	**43.25**$_{+1.42}$	**63.96**$_{+1.24}$
$\text{GRPO}_{70\%}$	69.63	84.00	58.93	89.58	80.51	75.47	76.35	22.20	51.44	34.40	54.20	40.56	62.04
Two-Stage	68.15	83.00	57.21	88.19	72.79	76.23	74.26	21.30	47.23	34.00	56.80	39.83	60.49
NR-RLVR	70.37	85.00	58.78	90.28	81.25	78.49	**77.36**$_{+1.01}$	22.80	51.88	36.40	58.40	**42.37**$_{+1.81}$	**63.37**$_{+1.33}$
$\text{GRPO}_{90\%}$	69.63	83.00	58.36	89.58	77.94	76.98	75.92	21.30	47.01	28.00	52.60	37.23	60.44
Two-Stage	66.67	81.00	56.68	87.50	67.28	74.34	72.25	8.90	32.59	30.60	25.00	24.27	53.06
NR-RLVR	71.11	85.00	58.21	92.36	78.68	77.74	**77.18**$_{+1.26}$	22.30	51.66	34.40	57.20	**41.39**$_{+4.16}$	**62.87**$_{+2.43}$

where $\mathrm{Sem}(\cdot,\cdot)$ measures semantic equivalence between two answers. $\delta_i = 0$ indicates a semantic conflict between the reference label and pseudo-label, while $\delta_i = 1$ indicates semantic consistency. When the two answers are semantically consistent, the reliability of the reference label is supported by the pseudo-label, and we retain the reference label for reward computation. When the two labels conflict, determining the correctness of the pseudo-label remains challenging. Unlike previous approaches that estimate label reliability based on statistical signals, such as entropy or sampling consistency, we leverage the model's intrinsic reasoning capability to evaluate the reliability of the two candidate labels in the given multimodal context.

To this end, we introduce a Confidence-gated Supervision Refinement (CSR) strategy, which dynamically determines the supervision signal used for reward computation. For conflicting samples ($\delta_i = 0$), we reformulate the open-ended reasoning task into a binary preference decision. Specifically, given the question and image, the model is asked to select the more reliable answer from two candidates, *i.e.* , Option A and Option B, which correspond to y_i and $\hat{y}_i$ with random ordering to avoid positional bias. The model is required to generate a structured response containing a reasoning process, the selected answer, and a confidence score, formatted as `<think>` reasoning `</think>`, `<answer>` a_i `</answer>`, and `<confidence>` c_i `</confidence>`. The selected answer and confidence score are denoted as:

$$(a_i, c_i) = f_\theta(q_i, v_i, y_i, \hat{y}_i), \tag{10}$$

where $a_i \in \{y_i, \hat{y}_i\}$ denotes the selected supervision signal, and the predicted confidence score $c_i \in [0, 1]$ serves as a reliability estimation of the selected answer. If the model selects the pseudo-label and the confidence exceeds a predefined threshold τ, we replace the reference label with the pseudo-label for reward computation. Otherwise, the original label is retained. Based on the selected answer and its confidence, the refined supervision is obtained as:

$$y_i^{\text{ref}} = \begin{cases} \hat{y}_i, & \delta_i = 0,\ a_i = \hat{y}_i,\ c_i > \tau, \\ y_i, & \text{otherwise.} \end{cases} \tag{11}$$

The refined supervision label y_i^{ref} is then used to calculate the accuracy reward. CSR enables the model to selectively correct unreliable annotations while preserving reference supervision when the model lacks sufficient confidence.

Table 2: Performance comparison of different baseline configurations using Qwen2.5-VL-3B under a 50% annotation noise ratio. **Bold** indicates the best performance.

ID	Ref. Guidance	Conf. Refinement	Text-only benchmarks							Vision-language benchmarks					Avg.
			Anatomy	Genetics	MedMCQA	Bio	ProfMed	Clinical	Avg.	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.	
-	***Qwen2.5-VL-3B***		57.78	71.00	45.21	65.97	58.46	66.42	60.81	14.50	22.62	20.40	42.40	24.98	46.48
-	$\text{GRPO}_{50\%}$		63.70	72.00	49.22	75.69	59.56	**71.32**	65.25	**20.60**	41.68	33.00	50.20	36.37	53.70
1	✗	✗	62.96	73.00	49.29	73.61	60.29	69.06	64.70	18.40	47.67	35.20	53.00	38.57	54.25
2	✓	✗	**65.19**	75.00	48.70	**77.08**	61.03	69.81	66.14	19.50	47.23	36.00	54.00	39.18	55.35
3	✗	✓	62.22	73.00	**49.63**	75.69	62.13	69.43	65.35	19.30	47.23	36.00	53.80	39.08	54.84
4	✓	✓	64.44	**76.00**	48.89	76.39	**62.13**	70.94	**66.47**	20.00	**48.12**	**36.80**	**55.60**	**40.13**	**55.93**

4 EXPERIMENTS

4.1 BENCHMARKS AND METRICS

To evaluate the robustness of our method under different levels of annotation noise, we construct training sets with varying noise ratios of 10%, 30%, 50%, 70%, and 90%. The models are evaluated on both text-only and vision-language medical benchmarks to comprehensively assess their capabilities across diverse clinical scenarios, including medical question answering, visual question answering, multiple-choice reasoning, and clinical image interpretation.

Text-only benchmarks. We adopt six widely used medical language benchmarks, including MedMCQA Pal et al. (2022), MMLU-Anatomy, MMLU-Bio, MMLU-Clinical, MMLU-Genetics, and MMLU-ProfMed Hendrycks et al. (2020). These benchmarks cover diverse medical knowledge and clinical reasoning tasks, enabling evaluation across multiple medical disciplines.

Vision-language benchmarks. We further evaluate our method on four medical multimodal benchmarks: PMC-VQA Zhang et al. (2023), RadVQA Lau et al. (2018), PathVQA He et al. (2020), and SLAKE-VQA Liu et al. (2021). These benchmarks cover diverse medical image understanding tasks across biomedical images, radiology, pathology, and clinical visual question answering.

For all benchmarks, we evaluate model responses using a combination of rule-based matching and LLM-as-a-judge evaluation to account for semantically equivalent answers. We report the accuracy to measure the reasoning capability of models under different levels of annotation noise.

4.2 IMPLEMENTATION DETAILS

Based on the MediX-RL dataset Mullappilly et al. (2026), we construct noisy training data following the procedure described in Section 3.2. We adopt Qwen2.5-VL-3B Bai et al. (2025b) and Qwen3-VL-4B Bai et al. (2025a) as the base policy models and train them with GRPO Shao et al. (2024) for 200 iterations on two H200 GPUs. An additional H200 GPU is used to deploy the reward model, Qwen3-32B-Instruct Yang et al. (2025a), via the vLLM inference engine. Each batch contains 256 samples, which are further divided into 2 PPO mini-batches Sheng et al. (2024). For each query, the policy model generates 8 sampled rollouts. The clipping parameter ϵ is set to 0.2, and the confidence threshold τ in CSR is set to 0.95. We use the AdamW optimizer Loshchilov & Hutter (2017) with an initial learning rate of 1×10^{-6}. The maximum response length is set to 2048 tokens.

4.3 MAIN RESULTS

Table 1 presents the performance comparison of different training strategies under various noise ratios using two base models, Qwen2.5-VL-3B and Qwen3-VL-4B, across both text-only and vision-language medical benchmarks. ***Qwen2.5-VL-3B*** and ***Qwen3-VL-4B*** denote the zero-shot performance of the corresponding base models. TTRL Zuo et al. (2026) denotes an unsupervised training strategy that utilizes majority-voting pseudo-labels for reward computation. $\text{GRPO}_{10\%}$ represents standard GRPO with 10% noisy annotations, and other noise settings follow the same notation. Two-Stage Xu et al. (2026) denotes a noise-robust RLVR baseline based on staged entropy optimization. NR-RLVR represents our proposed method under different noise ratios.

For both base models, standard GRPO consistently improves over the zero-shot baselines under most noise levels. Except for Qwen3-VL-4B under 90% noise, where a slight degradation is observed, GRPO maintains competitive performance across different noise settings. This indicates that GRPO itself exhibits certain robustness to noisy supervision, potentially benefiting from sparse reward sig-

Table 3: Ablation study on the number of reference-guided reasoning trajectories used for pseudo-label generation under a 50% noise ratio. **Bold** indicates the best performance.

#Ref.	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.
0	19.30	47.23	36.00	53.80	39.08
1	**20.00**	**48.12**	**36.80**	**55.60**	**40.13**
2	19.80	47.67	35.60	55.00	39.52
4	11.90	36.14	30.00	38.60	29.16

Table 4: Confidence thresholds ablation for supervision refinement at 50% annotation noise ratio. **Bold** indicates the best performance.

Threshold	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.
0	19.50	47.23	36.00	54.00	39.18
0.9	19.90	47.01	36.60	54.80	39.58
0.93	**20.50**	47.89	36.40	55.20	40.00
0.95	20.00	48.12	**36.80**	**55.60**	**40.13**
0.97	19.20	**48.56**	35.80	53.00	39.14
0.99	18.80	47.89	35.40	52.80	38.72

Table 5: Generalization to different RLVR algorithms under a 50% annotation noise. **Bold** indicates the best performance.

Method	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.
GRPO	**20.60**	41.68	33.00	50.20	36.37
w/ NR-RLVR	20.00	**48.12**	**36.80**	**55.60**	**40.13**
Dr.GRPO	20.00	42.13	32.80	51.60	36.63
w/ NR-RLVR	**20.50**	**47.01**	**34.60**	**54.00**	**39.03**
DAPO	22.30	44.57	34.20	52.40	38.37
w/ NR-RLVR	**18.90**	**49.89**	**36.80**	**56.60**	**41.70**

Table 6: Performance comparison between GRPO and NR-RLVR on four benchmarks under noise-free and fully noisy settings. **Bold** indicates the best performance.

Method	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.
$\text{GRPO}_{0\%}$	**21.70**	**57.21**	41.20	**61.40**	**45.38**
NR-RLVR	21.50	56.54	**41.40**	61.00	45.11
$\text{GRPO}_{100\%}$	8.80	34.37	31.60	24.40	24.79
NR-RLVR	**18.90**	**39.47**	**31.60**	**42.60**	**33.14**

nals. However, directly applying majority-voting-based pseudo-label supervision does not provide reliable improvements in open-ended medical question answering. Although TTRL has demonstrated effectiveness on tasks with constrained outputs, such as multiple-choice or short-answer question answering, its performance degrades in open-ended medical reasoning scenarios. Without explicit external supervision, pseudo-label-based optimization mainly reinforces the model's existing knowledge and generation preferences, making it difficult to correct factual errors or recover missing clinical information. Two-Stage improves robustness by adjusting the entropy optimization objective during training. However, it does not explicitly identify or refine corrupted supervision signals. Consequently, when the noise ratio increases, the optimization process may still be affected by unreliable annotations, limiting its effectiveness under severe noise conditions.

In contrast, NR-RLVR consistently achieves superior performance across different levels of annotation noise on both text-only and vision-language medical benchmarks. Moreover, the performance gain becomes more pronounced as the noise ratio increases, demonstrating the effectiveness of our framework in leveraging reliable supervision signals and dynamically correcting noisy reference annotations through confidence-gated supervision refinement during RLVR training.

4.4 Ablation and Analysis

In this section, we conduct a series of ablation studies based on Qwen2.5-VL-3B under the 50% noise to systematically examine the contributions of individual components and key design choices.

The Influence of Each Module. Table 2 presents the ablation results of different components in NR-RLVR. ***Qwen2.5-VL-3B*** and $\text{GRPO}_{50\%}$ follow the same definitions as those in Table 1. In #1, we first establish a baseline that generates pseudo-labels through majority voting and performs binary preference selection when semantic conflicts occur between the pseudo-label and reference answer. Compared with standard GRPO, this baseline improves performance on vision-language benchmarks, demonstrating the effectiveness of utilizing model-generated supervision signals. However, without exploiting reference information and confidence-based supervision refinement, its improvements on text-only benchmarks remain limited. In #2, adding reference-guided reasoning during pseudo-label generation further improves performance, indicating that the partial information preserved in noisy annotations can provide useful supervision signals for constructing more reliable pseudo-label candidates. In #3, introducing confidence-based supervision refinement further enhances robustness by dynamically selecting more reliable supervision signals during the computation of training rewards. In #4, combining the RPG and CSR components achieves the best overall performance across both text-only and vision-language medical benchmarks.

Effect of Reference-guided Trajectory Number. We investigate how the number of reference-guided reasoning trajectories affects pseudo-label generation. As shown in Table 3, introducing reference information into only one trajectory achieves the best performance. The setting with 0

reference-guided trajectories denotes the standard majority voting without reference information. When more trajectories are guided by reference answers, the generated responses become increasingly influenced by potentially noisy references. Particularly, when the diversity among rollouts is high, reference-guided trajectories may dominate majority voting process, reducing the independence of sampled responses and weakening the benefit of self-consistent reasoning.

Effect of the Confidence Threshold. We investigate the impact of the confidence threshold τ in the CSR module. As shown in Table 4, $\tau = 0.95$ achieves the best overall performance. When $\tau = 0$, the confidence gating mechanism is removed, and the model directly trusts the selected label whenever a conflict occurs, resulting in less reliable supervision refinement. Lower thresholds allow low-confidence predictions to be accepted as supervision signals, reducing reward reliability. Conversely, overly large thresholds impose excessively strict selection criteria, preventing useful pseudo-labels from contributing to supervision refinement. Across the other evaluated noise settings, $\tau = 0.95$ also consistently provides effective filtering of unreliable pseudo-labels, indicating its robustness to varying annotation noise levels. These results suggest that $\tau = 0.95$ provides a reliable balance between noise suppression and effective utilization of pseudo-labels.

Generalization across RLVR Algorithms. To demonstrate that NR-RLVR is not restricted to GRPO, we further evaluate its compatibility with different RLVR variants, including Dr.GRPO and DAPO. As shown in Table 5, incorporating NR-RLVR consistently improves the performance of different RLVR algorithms under noisy supervision. These results demonstrate that NR-RLVR is orthogonal to the underlying RLVR algorithms and effectively enhances robustness by refining supervision signals and reward reliability.

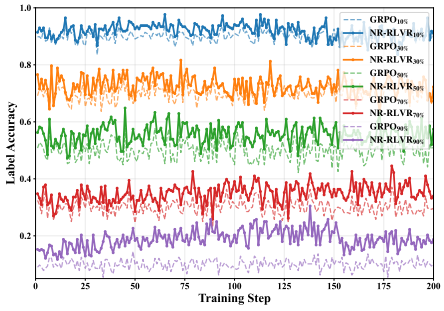

Figure 3: Label quality comparison before and after supervision refinement. Dashed and solid curves denote the original noisy annotations and refined labels, respectively.

Robustness under Extreme Noise Ratios. To evaluate the robustness of NR-RLVR, we investigate two extreme scenarios with 0% and 100% annotation noise. As shown in Table 6, NR-RLVR maintains comparable performance under the noise-free setting, indicating that our method can effectively preserve reliable supervision signals when reference annotations are accurate, while the confidence-based refinement mechanism avoids excessive intervention on correct annotations. More importantly, NR-RLVR still achieves substantial improvements over standard RLVR in the fully noisy setting, demonstrating its ability to refine unreliable supervision and maintain effective optimization even when all reference annotations are corrupted.

Label Refinement Analysis. We analyze the effectiveness of our supervision refinement strategy by comparing the quality of original noisy annotations and refined labels under different noise ratios. As shown in Figure 3, the dashed curves represent the accuracy of original noisy annotations, while the solid curves indicate the accuracy after refinement. Compared with the original annotations, the refined labels consistently achieve higher accuracy across all noise levels. demonstrating that NR-RLVR can effectively refine supervision even under severe annotation noise.

5 Conclusion

In this work, we study the problem of RLVR for multimodal medical reasoning under noisy supervision. Unlike existing RLVR methods that rely on accurate answers for reward computation, in real-world medical scenarios, annotations often contain structured noise, leading to unreliable rewards and biased policy optimization. To address this challenge, we propose NR-RLVR, a noise-robust RLVR framework that improves supervision through two key components: Reference-aware Pseudo-label Generation, which leverages model reasoning trajectories and reference information to construct reliable pseudo-labels, and Confidence-gated Supervision Refinement, which dynamically selects supervision based on model confidence. Extensive experiments on diverse text-only and vision-language medical benchmarks demonstrate that NR-RLVR consistently improves reasoning performance and robustness across different noise levels.

AI USE STATEMENT

We used generative AI models to construct synthetic noisy annotations and perform semantic evaluation, as described in the paper. Additionally, we used AI tools for translation, grammar checking, and language polishing. The human authors reviewed the AI-assisted outputs for accuracy and consistency and take responsibility for the final manuscript, including its technical content, experimental findings, and conclusions. The use of AI tools does not affect the authorship of this work.

ETHICS STATEMENT

This research follows the ICLR Code of Ethics. Our work focuses on noise-robust medical reasoning using publicly available datasets and pretrained models. Synthetic annotation noise is introduced for research purposes. The authors declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3.2 describes the construction of synthetic noisy annotations under different noise ratios. Sections 4.1 and 4.2 specify the evaluation benchmarks and metrics, policy and reward model configurations, training hyperparameters, and computational resources. Section A of the supplementary material provides the prompts for training and inference, reference-guided reasoning, semantic clustering, confidence-based preference selection, answer evaluation, and noise generation. No proprietary data or models are used in this work.

REFERENCES

Mohamed Abdalla and Benjamin Fine. Hurdles to artificial intelligence deployment: noise in schemas and "gold" labels. *Radiology: Artificial Intelligence*, 2023.

Rawan AlSaad, Alaa Abd-Alrazaq, Sabri Boughorbel, Arfan Ahmed, Max-Antoine Renault, et al. Multimodal large language models in health care: applications, challenges, and future outlook. *Journal of Medical Internet Research*, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated input-dependent label noise in large-scale image classification. In *Computer Vision and Pattern Recognition Conference*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *The Association for Computational Linguistics*, 2019.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025a.

Erjian Guo, Zhen Zhao, Zicheng Wang, Tong Chen, Yunyi Liu, and Luping Zhou. Din: Diffusion model for robust medical vqa with semantic noisy labels. In *Computer Vision and Pattern Recognition Conference*, 2025b.

Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Neural Information Processing Systems*, 2018.

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

Dan Hendrycks, Mantas Mazeika, et al. Using trusted data to train deep networks on labels corrupted by severe noise. *Neural Information Processing Systems*, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda Shapiro. Quilt-1m: One million image-text pairs for histopathology. *Neural Information Processing Systems*, 2023.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

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

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Neural Information Processing Systems*, 2023.

Mengting Li and Chuang Zhu. Noisy label processing for classification: A survey. *arXiv preprint arXiv:2404.04159*, 2024.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. In *AAAI Conference on Artificial Intelligence*, 2026.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, et al. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

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

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: Open multimodal generative pre-trained transformer for biomedicine. *arXiv preprint arXiv:2308.09442*, 2023.

Qinghe Ma, Zhen Zhao, Yiming Wu, Jian Zhang, Lei Bai, and Yinghuan Shi. Are tools always beneficial? learning to invoke tools adaptively for dual-mode multimodal llm reasoning. *arXiv preprint arXiv:2605.19852*, 2026.

Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature Communications*, 2018.

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

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Omair Mohamed, Mohamed Zidan, Fahad Khan, Salman Khan, Rao Anwer, and Hisham Cholakkal. Medix-r1: Open ended medical reinforcement learning. *arXiv preprint arXiv:2602.23363*, 2026.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Computer Vision and Pattern Recognition Conference*, 2025.

Junzhi Ning, Wei Li, Cheng Tang, Jiashi Lin, Chenglong Ma, Chaoyang Zhang, Jiyao Liu, Ying Chen, et al. Unimedvl: Unifying medical multimodal understanding and generation through observation-knowledge-analysis. *arXiv preprint arXiv:2510.15710*, 2025.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning*. PMLR, 2022.

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

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

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

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. *Neural Information Processing Systems*, 2026.

Dianzhe Tian, Shitao Jiang, Lei Zhang, Xin Lu, and Yiyao Xu. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 2024.

Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering*, 2024.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition Conference*, 2017.

Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 2004.

Yishu Wei, Yu Deng, Cong Sun, Mingquan Lin, Hongmei Jiang, and Yifan Peng. Deep learning with noisy labels in medical prediction problems: a scoping review. *Journal of the American Medical Informatics Association*, 2024.

Donglai Xu, Hongzheng Yang, Yuzhi Zhao, Pingping Zhang, Jinpeng Chen, Wenao Ma, Zhijian Hou, Mengyang Wu, Xiaolei Li, Senkang Hu, et al. From exploration to exploitation: A two-stage entropy rlvr approach for noise-tolerant mllm training. In *Computer Vision and Pattern Recognition Conference*, 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Shenzhi Yang, Guangcheng Zhu, Bowen Song, Sharon Li, Haobo Wang, Xing Zheng, Yingfan Ma, Zhongqi Chen, Weiqiang Wang, and Gang Chen. Can llms learn to reason robustly under noisy supervision? *arXiv preprint arXiv:2604.03993*, 2026.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *International Conference on Computer Vision*, 2025b.

Huanjin Yao, Qixiang Yin, Jingyi Zhang, Min Yang, Yibo Wang, Wenhao Wu, Fei Su, Li Shen, Minghui Qiu, Dacheng Tao, et al. R1-sharevl: Incentivizing reasoning capability of multimodal large language models via share-grpo. *arXiv preprint arXiv:2505.16673*, 2025.

En Yu, Kangheng Lin, Liang Zhao, Yana Wei, Yuang Peng, et al. Perception-r1: Pioneering perception policy with reinforcement learning. *Neural Information Processing Systems*, 2026a.

Jiachen Yu, Zhihao Xu, Junjie Wang, and Yujiu Yang. Think-with-rubrics: From external evaluator to internal reasoning guidance. *arXiv preprint arXiv:2605.07461*, 2026b.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Neural Information Processing Systems*, 2026c.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 2024.

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

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025a.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025b.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, Xinwei Long, et al. Ttrl: Test-time reinforcement learning. *Neural Information Processing Systems*, 2026.

A Prompt Design

Training and Inference Prompts. To ensure consistent interaction between the model and the environment, we carefully design the prompts for both training and inference stages. The prompts are designed to encourage the model to perform explicit reasoning before generating the final answer, enabling more reliable reasoning trajectories for reward computation and RLVR optimization.

<image>
Question:{question}

Your task: 1. Think through the question step by step, enclose your reasoning process in <think>...</think>tags. 2. Then provide only the final answer inside <answer>...</answer>tags. 3. No extra information or text outside of these tags.

For the Reference-aware Pseudo-label Generation (RPG) module, we additionally construct a reference-guided prompt by providing the reference label as an auxiliary reasoning clue. Importantly, the reference label is only introduced in the reference-guided rollout, while the remaining trajectories follow the standard prompt to preserve exploration diversity.

<image>
Question:{question}

Reference label: $\{y_i\}$
Instruction: The reference label may be incorrect. Use it as an auxiliary clue to determine the final answer.

Your task: 1. Think through the question step by step, enclose your reasoning process in <think>...</think>tags. 2. Then provide only the final answer inside <answer>...</answer>tags. 3. No extra information or text outside of these tags.

Semantic Clustering Prompt for Pseudo-label Generation. To obtain reliable pseudo-labels from multiple sampled reasoning trajectories, we reuse the LLM-based reward model Qwen3-32B-Instruct Yang et al. (2025a) to cluster rollouts according to their semantic consistency. The following prompt is used to identify semantically consistent answers and generate normalized medical conclusions.

You are a medical expert.

Given a clinical question and multiple candidate answers, your task is to group the answers into clusters based on semantic meaning.

Instructions:

1. Cluster the answers:
   - Answers that express the same clinical conclusion should be grouped together.
   - Answers with different clinical meanings MUST be placed into different clusters.
   - You MUST create multiple clusters if there are semantically different answers.
   - Consider:
     • disease/diagnosis

• anatomical location
• synonyms and abbreviations
• negation (e.g., ”no pneumonia” $\neq$ ”pneumonia”)

2. Each cluster must represent ONE normalized medical conclusion.

3. For each cluster:
- Provide a concise normalized_answer
- Include ALL original answers belonging to this cluster
- Provide the count

4. Constraints:
- Every answer must appear in exactly ONE cluster (no missing, no duplication)
- The sum of all cluster counts must equal the number of input answers
- Do NOT merge answers with different diseases, locations, or negation
- Do NOT output only one cluster unless ALL answers are semantically equivalent

Output format (strict JSON): {{
“clusters”: [
{{
“normalized_answer”: “...”,
“count”: X,
“members”: [“answer1”, “answer2”, ...]
}},
{{
“normalized_answer”: “...”,
“count”: Y,
“members”: [...]
}}
]
}}
Clinical Question:
{question}

Candidate Answers:
{answers}

Confidence-based Preference Selection Prompt. When the pseudo-label and the reference answer exhibit semantic inconsistency, we reformulate the original open-ended question as a binary preference selection problem. Specifically, the model is provided with the question, image, and two candidate answers with random ordering, and is instructed to select the more reliable answer while estimating its confidence. The following prompt is used for preference selection process.

<image>
Question:{question}
A){pseudo-label}, B){reference label}

Your task: 1. Think through the question step by step, enclose your reasoning process in <think>...</think>tags.

2. Then provide only the final answer inside <answer>...</answer>tags, and the answer must be only the option letter (A or B). Also output your confidence in <confidence>...</confidence>tags as a number between 0 and 1, where:
- 0.0-0.2: you have no idea which is correct
- 0.3-0.5: you lean toward one option but are very uncertain
- 0.5-0.7: you think one is more likely but could be wrong

- 0.7-0.9: you are fairly confident
- 0.9-1.0: you are absolutely certain. If you believe neither A nor B is correct, output 0 as the confidence and choose the one you think is closer to correct.

3. No extra information or text outside of these tags.

Accuracy Evaluation Prompt. To evaluate the correctness of model responses, in addition to rule-based string matching, we employ an LLM-based reward model Qwen3-32B-Instruct Yang et al. (2025a) to determine whether the generated answer is semantically consistent with the ground truth. The same evaluation criterion is used for both accuracy reward computation during training and final performance evaluation on medical benchmarks.

You are a strict medical VQA evaluator.

Task:
Determine whether the model answer correctly answers the question
and is semantically equivalent to the ground truth.

Key principles:
1. The answer must directly answer the question with a clear and specific conclusion.
2. The answer must match the meaning of the ground truth (synonyms or paraphrases are allowed).

Strict rule:
- Output 0 if the answer:
  (a) refuses to answer or claims insufficient information,
  (b) is vague or lists multiple possibilities without a clear conclusion,
  (c) does not directly answer the question,
  (d) is more general than the ground truth.

Question:
{question}

Ground truth:
{solution}

Model answer:
{completion}

Evaluation:
- Output "1" ONLY if the answer clearly matches the ground truth.
- Otherwise output "0".

Output ONLY a single digit: 1 or 0.

Prompt for Noisy Medical Dataset Construction To construct noisy medical training data, we employ a stronger MLLM, Qwen3-VL-30B-A3B Bai et al. (2025a), to perturb original clean answers. The generator is instructed to introduce medically plausible but incorrect modifications, simulating realistic annotation errors rather than random irrelevant responses. The following prompt is used to generate corrupted reference answers.

You are a medical expert creating adversarial training data.

Given a medical VQA question and its correct answer, your task is to generate a WRONG answer.

Table 7: Results on Qwen2.5-VL-7B and Qwen3-VL-8B under 50% noise ratios. Each subscript indicates the improvement compared with naive GRPO. **Bold** indicates the best performance.

Method	Text-only benchmarks							Vision-language benchmarks					Avg.
	Anatomy	Genetics	MedMCQA	Bio	ProfMed	Clinical	Avg.	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.	
Qwen2.5-VL-7B	66.67	74.00	53.14	81.94	67.28	77.74	70.13	16.7	46.56	20.40	45.20	32.22	54.96
$\text{GRPO}_{50\%}$	70.37	76.00	53.62	81.25	70.96	76.98	71.53	23.90	54.77	37.40	55.40	42.87	60.07
NR-RLVR	71.85	80.00	53.79	84.72	72.43	78.87	**73.61**$_{+2.08}$	24.20	56.54	41.20	55.20	**44.29**$_{+1.42}$	**61.88**$_{+1.81}$
Qwen3-VL-8B	72.59	84.00	64.28	89.58	80.51	83.02	79.00	22.40	54.77	32.40	55.00	41.14	63.86
$\text{GRPO}_{50\%}$	71.85	86.00	64.76	91.67	82.72	84.53	80.26	24.20	55.21	39.60	60.20	44.80	66.07
NR-RLVR	74.07	89.00	64.40	92.36	83.82	85.66	**81.55**$_{+1.29}$	24.30	57.87	41.80	59.40	**45.84**$_{+1.04}$	**67.27**$_{+1.20}$

Table 8: Performance comparison with different reward model scales under a 50% annotation noise. **Bold** indicates the best performance.

Method	PMC-VQA	RadVQA	PathVQA	SLAKE-VQA	Avg.
$\text{GRPO}_{Qwen3-8B}$	19.70	39.91	33.40	**49.40**	35.60
NR-RLVR	**20.00**	**46.56**	**34.60**	49.20	**37.59**
$\text{GRPO}_{Qwen3-14B}$	20.30	40.13	34.00	49.00	35.86
NR-RLVR	**20.80**	**46.78**	**36.20**	**49.60**	**38.35**
$\text{GRPO}_{Qwen3-32B}$	**20.60**	41.68	33.00	50.20	36.37
NR-RLVR	20.00	**48.12**	**36.80**	**55.60**	**40.13**

[Faithfulness]
- Keep the same format and style

[Local Perturbation]
- Only modify key medical facts

[Plausibility]
- Must sound medically reasonable

[Incorrectness]
- Must be clearly incorrect overall

Question:
{question}

Correct Answer:
{answer}

Wrong Answer:

B Further Analysis

Scalability to Larger MLLMs. To evaluate the scalability of NR-RLVR to larger MLLMs, we conduct additional experiments beyond the main experiments on Qwen2.5-VL-3B Bai et al. (2025b) and Qwen3-VL-4B Bai et al. (2025a). As shown in Table 7, NR-RLVR consistently improves the performance of these models, including Qwen2.5-VL-7B Bai et al. (2025b) and Qwen3-VL-8B Bai et al. (2025a), over GRPO across text-only and vision-language medical benchmarks. These results demonstrate that NR-RLVR remains effective when applied to MLLMs with different scales.

Impact of Reward Model Scale. We investigate the impact of reward model Scale on NR-RLVR. Specifically, we replace the default reward model Qwen3-32B with smaller variants, including Qwen3-8B and Qwen3-14B, while keeping other configurations unchanged. As shown in Table 8, stronger reward models consistently improve the final performance. First, a more capable reward model provides more accurate semantic evaluation for the accuracy reward computation. Second, unlike conventional RLVR methods, NR-RLVR additionally relies on the reward model for supervision refinement, including semantic clustering of rollouts during pseudo-label generation and semantic consistency evaluation between candidate answers. Therefore, improved reward model capability

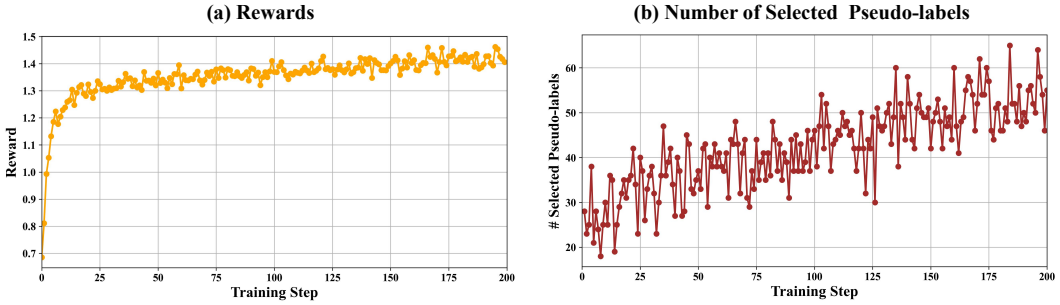

Figure 4: (a) Training reward curve of NR-RLVR under 50% annotation noise. (b) Evolution of selected pseudo-labels during training under 50% annotation noise.

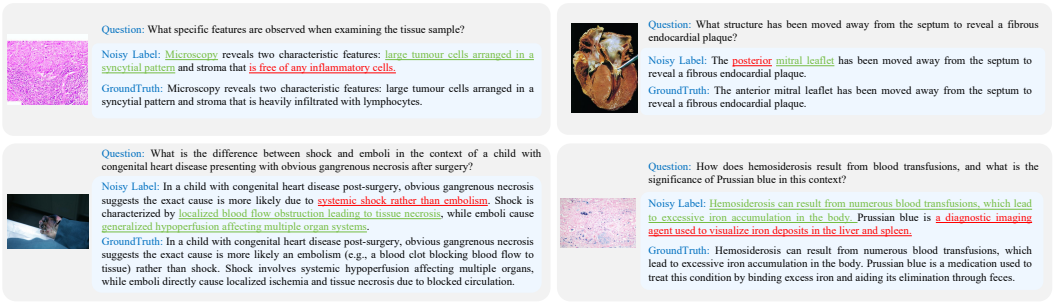

Figure 5: Examples of synthetic annotation noise generated by the proposed noise construction strategy. The corrupted answers preserve structural consistency and medical plausibility while introducing critical fact perturbations, where reliable information such as imaging modality and anatomical details is retained but key medical conclusions are altered.

leads to higher-quality pseudo-labels and more reliable supervision selection, further enhancing robustness under noisy annotations.

Training Dynamics Analysis. To understand the optimization behavior of NR-RLVR, we analyze both the training reward evolution and pseudo-label selection dynamics under the 50% annotation noise setting. As shown in Figure 4(a), the training reward steadily increases and gradually converges during optimization, indicating stable policy improvement despite noisy supervision. Figure 4(b) further illustrates the evolution of pseudo-label selection during training. The number of pseudo-labels selected for reward computation gradually increases as training progresses, suggesting that the model becomes increasingly capable of identifying reliable supervision signals through semantic conflict resolution and confidence estimation. These results demonstrate that NR-RLVR can progressively refine noisy supervision while maintaining stable RLVR optimization.

Training Efficiency Analysis. Under the 50% noise setting with Qwen2.5-VL-3B, standard GRPO requires 21.43 hours for training, while NR-RLVR takes 24.95 hours. The additional overhead mainly comes from the semantic clustering of rollout responses and the extra reasoning process for resolving semantic conflicts, which introduce an average cost of 56 seconds and 5 seconds per step, respectively. Although NR-RLVR introduces additional computation compared with standard GRPO, the overhead remains acceptable relative to the overall training time. This moderate cost enables more reliable supervision refinement and improves RLVR robustness under noisy medical annotations.

Examples of Structured Annotation Noise To better understand the characteristics of the constructed noisy data, we provide several examples of corrupted medical annotations in Figure 5. The generated noisy answers preserve structural consistency and medical plausibility while introducing critical fact perturbations that make them semantically inconsistent with the original ground truth. Notably, the corrupted annotations often retain reliable low-level clinical information, such as imaging modality and anatomical details, while altering key medical conclusions.

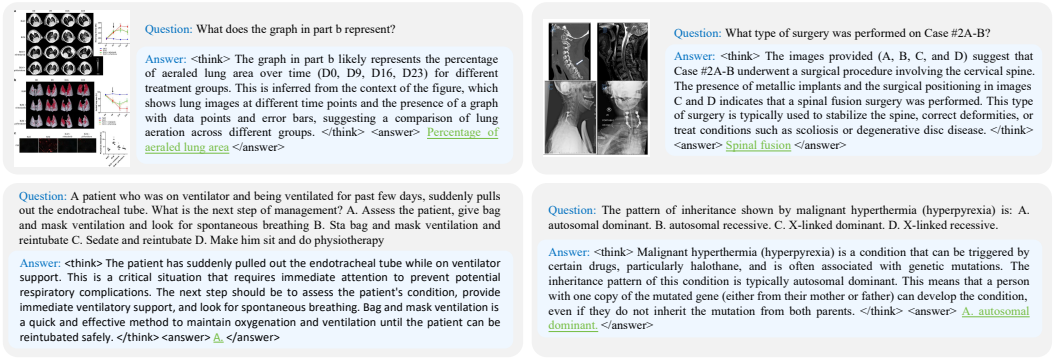

Figure 6: Qualitative examples of model responses generated by Qwen2.5-VL-3B trained with NR-RLVR under 50% annotation noise. The examples include both vision-language and text-only medical question answering tasks.

Qualitative Analysis of Synthetic Noise. Our noise construction aims to emulate plausible human annotation errors by modifying key medical facts while preserving the surrounding semantic context. As shown in Figure 7, the joint t-SNE visualization of BERT-extracted Devlin et al. (2019) features from 1,000 original and 500 noisy QA pairs shows substantial overlap between the two groups. This observation suggests that our semantic noise preserves the medical context of the original data. Compared with arbitrary, semantically unrelated answer corruption, our construction is designed to more closely resemble human annotation errors in real-world settings by introducing plausible factual mistakes within a coherent medical context.

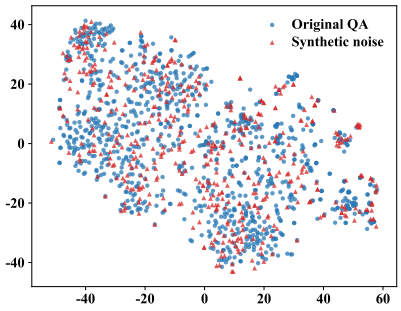

Figure 7: Joint t-SNE visualization of BERT representations of original and synthetically corrupted data.

Qualitative Examples. We provide qualitative examples of Qwen2.5-VL-3B trained with NR-RLVR under 50% annotation noise. As shown in Figure 6, the examples cover both vision-language and text-only medical question answering tasks, where the model produces reliable responses and demonstrates robust visual understanding, medical reasoning, and domain knowledge capabilities.

Relation to Medical-Specific Noisy-Label Learning. Medical annotation noise often involves semantically plausible errors, motivating domain-specific supervision refinement. For example, DiN Guo et al. (2025b) predicts answers within a predefined vocabulary and dynamically combines predicted class distributions with observed labels to supervise diffusion training, alongside a robust classification objective. NR-RLVR instead targets autoregressive reasoning, where refined textual supervision determines rewards for policy optimization. This setting introduces additional challenges: identifying semantic equivalence across diverse generated responses, retaining useful information in partially incorrect references, and preventing erroneous rewards from distorting group-relative advantages and reinforcing incorrect responses. NR-RLVR thus complements existing medical VQA denoising methods by addressing supervision reliability in generative policy optimization.

Limitations. Due to the limited availability of Med-VQA datasets with documented label noise and the difficulty of collecting such data, we construct noisy training annotations synthetically. Despite preserving medical plausibility and semantic structure, synthetic errors may differ from real annotation mistakes, diagnostic uncertainty, and inter-annotator disagreements. Future work could collect naturally noisy medical annotations and curate multi-annotator datasets to validate real-world robustness. Additionally, potential failures during training may primarily stem from incorrect pseudo-labels generated by RPG and erroneous supervision replacement by CSR. Developing more accurate and reliable mechanisms for pseudo-label generation and selection therefore remains an important research direction.