# STAC: EVALUATOR-FREE MEDICAL MULTIMODAL RLVR VIA STABILITY-AWARE CREDIT ASSIGNMENT

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Although reinforcement learning with verifiable rewards (RLVR) has shown promise in medical multimodal large language models, binary outcome supervision cannot distinguish robust reasoning from fragile successes that arrive at the correct answer by chance. While process reward models offer finer-grained guidance, they introduce heavy annotation or external evaluator overhead. In this work, we reveal that the rollout policy's own intermediate states inherently reflect trajectory stability: early prefix support for the eventual answer reliably predicts continuation stability under resampling, serving as a lightweight process signal orthogonal to final correctness. Grounded in this insight, we propose **STAC** (**ST**ability-**A**ware **C**redit Assignment), an evaluator-free framework that refines credit assignment using internal policy dynamics. STAC introduces **Prefix Support Probing** to evaluate early intermediate states against competing hypotheses, and couples it with terminal verification via **Outcome-Conditioned Reward Modulation** to reward stably supported correct reasoning while suppressing credit for stubborn errors. Crucially, STAC requires no external judges, step-level annotations, or rollout branching during training. Across six diverse medical visual question answering benchmarks, STAC consistently outperforms strong RLVR baselines, significantly improving both diagnostic accuracy and reasoning stability.

## 1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) has emerged as a promising paradigm for enhancing medical multimodal reasoning without costly expert demonstrations (Lai et al., 2025; Pan et al., 2025; Jiang et al., 2026b; Liu et al., 2025). However, the vanilla outcome-only paradigm introduces a fundamental limitation in credit assignment by evaluating solely terminal answers and treating all successful trajectories identically. In multimodal clinical reasoning, models frequently exploit language shortcuts or misread complex visual findings, arriving at correct conclusions for the wrong reasons (Wu et al., 2026; Sadanandan & Behzadan, 2026; Liu et al., 2026; Zafar et al., 2026). Consequently, standard RLVR equally rewards fragile successes and truly correct reasoning, weakening the supervisory signal and hindering reliable reasoning.

Supervising intermediate reasoning beyond the final answer is therefore promising, yet existing approaches face practical challenges in medical scenarios. Process reward models (PRMs) and rubric-based evaluators provide step-level feedback (Lightman et al., 2024; Yun et al., 2025; Fan et al., 2026), but they often require costly dense annotations or specialized adaptation across diverse medical modalities. Alternatively, recent efforts leverage the model's self-confidence to avoid external judges (Du et al., 2026), but such signals primarily assess the completed output and leave the dynamics of intermediate reasoning unexamined. Fundamentally, these attempts share the common objective of seeking supervision beyond final outcome correctness to foster more stable and reliable reasoning. However, lightweight attempts still lack awareness of the intermediate reasoning process, while process-aware evaluators introduce heavy external overhead. This dilemma motivates a central question: ***Can the rollout policy's own intermediate states inherently reflect trajectory stability, thereby providing a lightweight process signal for RLVR without relying on external annotation?***

To explore this question, we firstly conduct a pilot study on multimodal medical visual question answering. By probing the policy's relative support for its eventual answer from early reasoning prefixes, we discover that this intermediate signal could reflect the continuation stability of the



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.

reasoning trajectory, which in turn captures the intrinsic soundness of the intermediate reasoning process. Moreover, our analysis reveals that this internal stability represents an orthogonal dimension to terminal correctness. While final-answer verification assesses diagnostic validity, early prefix support captures the model's internal commitment to its reasoning path. This orthogonal perspective naturally enriches RLVR, helping prioritize genuinely robust reasoning over fragile successes among correct rollouts, while discouraging stubborn commitment to erroneous clinical deductions.

Grounded in these insights, we propose **STAC** (**ST**ability-**A**ware **C**redit Assignment), an evaluator-free RLVR framework that refines credit assignment directly from internal policy dynamics. Specifically, **Prefix Support Probing** examines early intermediate states to evaluate the policy's relative support for its eventual answer against competing hypotheses, capturing trajectory stability without costly resampling. Additionally, **Outcome-Conditioned Reward Modulation** anchors this support signal to terminal verification, allocating greater credit to stably derived correct reasoning while suppressing stubborn adherence to wrong deductions. Consequently, STAC enriches standard RLVR with lightweight process awareness, guiding the policy toward trajectories that are not merely correct, but inherently well-supported and clinically reliable.

We evaluate STAC across six medical vision-language benchmarks covering diverse clinical modalities and task formats. STAC consistently outperforms both standard RLVR and confidence-aware baselines in diagnostic accuracy, demonstrating strong generalization across datasets without increasing model size. Furthermore, ablation studies confirm that early prefix support probing and outcome-conditioned modulation are both essential to its effectiveness, while in-depth trajectory analyses reveal that STAC substantially improves continuation stability.

Our main contributions are summarized as follows: **(1)** We identify early prefix answer support as an intrinsic proxy for trajectory stability and show that this internal stability is orthogonal to final-answer correctness in medical reasoning. **(2)** We propose STAC, an evaluator-free RLVR framework that couples early prefix support probing with terminal verification to refine credit assignment without external process reward models. **(3)** Extensive experiments across six medical multimodal benchmarks demonstrate that STAC consistently outperforms strong baselines, improving both accuracy and reasoning stability across diverse clinical settings.

## 2 Related Work

### 2.1 Medical Multimodal Large Language Models (MLLMs)

Recent advances in medical multimodal large language models (MLLMs) have demonstrated substantial promise in integrating clinical domain knowledge with multi-step diagnostic reasoning (Sellergren et al., 2025; Xu et al., 2025; Shi et al., 2026a). Despite this progress, supervised fine-tuning alone often struggles to instill reliable deductive rigor across complex clinical scenarios, driving increasing interest in post-training optimization. Within this landscape, reinforcement learning with verifiable rewards (RLVR) has emerged as an effective paradigm to incentivize complex clinical deduction without requiring dense expert demonstrations (Lai et al., 2025; Su et al., 2025; Jiang et al., 2026b). However, standard medical RLVR frameworks remain predominantly outcome-oriented, awarding identical credit to all trajectories that reach the correct terminal answer regardless of their underlying reasoning stability. To enrich this coarse terminal feedback, process reward models (PRMs) and rubric-based evaluators introduce step-level supervision (Yun et al., 2025; Fan et al., 2026; Zhang et al., 2025; Rezaei et al., 2026; Han et al., 2026), yet they typically demand costly expert-designed rubrics, extra step annotations, or external judge models that struggle to generalize across heterogeneous medical modalities. Other lines of work incorporate diagnostic confidence or cross-modal consistency (Du et al., 2026; Jiang et al., 2026a), but these signals primarily characterize the completed prediction rather than the internal progression of the intermediate reasoning states. In contrast, STAC extracts a lightweight, policy-derived commitment signal directly from intermediate prefixes and modulates it via terminal correctness, distinguishing fragile successes from robust reasoning without relying on external process evaluators or dense annotations.

### 2.2 CoT Failure in Medical MLLMs

While chain-of-thought (CoT) prompting frequently benefits complex reasoning, longer rationales do not consistently translate to reliable multimodal clinical diagnosis. Empirical studies reveal that

extended reasoning traces can degrade both in-domain and out-of-domain generalization or trigger erratic responses under format shifts (Lai et al., 2025; Sadanandan & Behzadan, 2026). A fundamental source of this degradation is the perception-reasoning conflict in multimodal clinical settings, where verbose textual reasoning can detach from visual evidence and become dominated by strong linguistic priors, preserving nominal answers even under counterfactual visual perturbations (Shi et al., 2026a; Zafar et al., 2026). Consequently, extended traces may project an illusion of deductive depth while masking unsupported commitments and visual neglect, which terminal correctness alone fails to expose. These vulnerabilities highlight the necessity of decoupling intermediate trajectory reliability from final-answer correctness. Our work builds on this insight by probing the policy's early prefix support. More specifically, STAC leverages this intrinsic signal to reinforce stably supported correct trajectories while penalizing stubborn adherence to erroneous conclusions, thereby curbing spurious reasoning without sacrificing the scalability of outcome-based RLVR.

# 3 PRELIMINARY

## 3.1 PROBLEM SETUP

When applying RLVR to medical MLLMs, the model is encouraged to explore its own reasoning trajectories. Formally, consider training examples $(I, Q, y) \sim \mathcal{D}$, where $I$ contains one or multiple medical images, $Q$ is the associated query, and $y$ is the ground-truth answer. Group Relative Policy Optimization (GRPO) (DeepSeek-AI et al., 2025), a representative RLVR algorithm, samples a group of $G$ rollouts $\{o_i = (z_i, a_i)\}_{i=1}^{G}$, where $z_i$ and $a_i$ denote the reasoning trajectory and final answer of rollout $o_i$, respectively. Each sampled rollout is then evaluated by an outcome verifier, yielding a scalar reward $R_i = R(o_i, y)$, typically based on final-answer correctness. GRPO normalizes the rewards within the sampled group to obtain the rollout-level advantage $\hat{A}_i = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G}) + \epsilon_A}$, and optimizes the policy with the following clipped surrogate objective:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left(\min\left[\gamma_{i,t}(\theta)\hat{A}_i, \text{clip}\left(\gamma_{i,t}(\theta), 1-\epsilon, 1+\epsilon\right)\hat{A}_i\right] - \beta D_{\text{KL}}\left(\pi_\theta \| \pi_{\text{ref}}\right)\right)\right], \quad (1)$$

where, $\gamma_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}|I,Q,o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}|I,Q,o_{i,<t})}$ denotes the token-level importance ratio between the current and old policies, and the KL term regularizes $\pi_\theta$ toward the reference policy $\pi_{\text{ref}}$, with $\beta$ controlling its strength. This final-answer-oriented training paradigm has been shown to improve the performance of medical MLLMs (Su et al., 2025; Pan et al., 2025; Shi et al., 2026b).

## 3.2 PILOT STUDY: EARLY PREFIX SUPPORT REVEALS TRAJECTORY STABILITY

Despite its effectiveness, final-answer-oriented GRPO assigns identical outcome credit to all rollouts reaching the correct answer, overlooking differences in the stability of their underlying reasoning. Such distinctions are especially critical in multimodal medical scenarios, where chain-of-thought reasoning may be brittle or misleading (Zafar et al., 2026; Sadanandan & Behzadan, 2026). To characterize whether an intermediate reasoning state stably supports its eventual outcome rather than hitting it by chance, we introduce two complementary metrics.

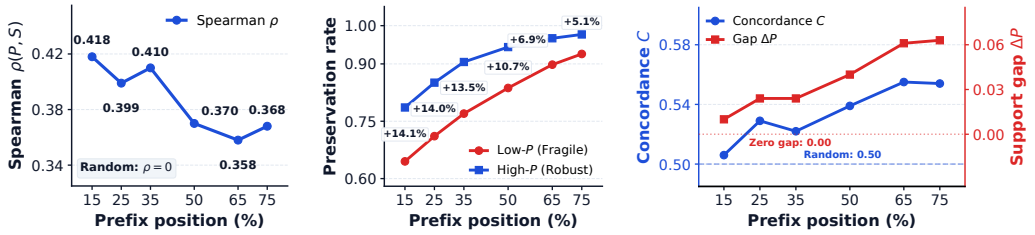
**Definition 1. (Prefix-Conditioned Answer Support, $P$).** Given a rollout $o_i = (z_i, a_i)$ in a rollout group, we denote its intermediate reasoning state at position $t$ as $s_{i,t} = (I, Q, z_{i,\leq t})$. For each rollout group $g$, we construct a candidate answer set containing the unique rollout answers together with the ground-truth answer $y$:

$$\mathcal{C}_g = \text{Unique}\big(\{a_j\}_{j=1}^{G} \cup \{y\}\big) \quad (2)$$

For each candidate $a = (a_1, \ldots, a_{|a|}) \in \mathcal{C}_g$, formatted within <answer> and </answer> tags, we score its support under the current reasoning state by the length-normalized mean token log-probability and softmax-normalize these scores over $\mathcal{C}_g$ to obtain the prefix answer support of the rollout's original answer $a_i$:

$$\ell_\theta(a \mid s_{i,t}) = \frac{1}{|a|}\sum_{m=1}^{|a|}\log \pi_\theta\left(a_m \mid s_{i,t}, a_{<m}\right), \quad P_{i,t} = \frac{\exp(\ell_\theta(a_i \mid s_{i,t}))}{\sum_{a\in\mathcal{C}_g}\exp(\ell_\theta(a \mid s_{i,t}))}. \quad (3)$$

Accordingly, a larger $P_{i,t}$ indicates stronger relative support for $a_i$ at prefix $s_{i,t}$.




(a) Prefix support tracks continuation stability.

(b) Prefix support distinguishes stability of correct traces.

(c) Prefix support weakly separates correct from incorrect rollouts.

Figure 1: **Prefix answer support reveals stability differences hidden by final-answer correctness.** Trajectories are truncated at 15%, 25%, 35%, 50%, 65%, and 75% of their generated length.

**Definition 2. (Continuation Stability, $S$).** Given an intermediate reasoning state $s_{i,t}$, we independently sample $K$ continuations $\{\tilde{o}_{i,t}^{(k)}\}_{k=1}^{K}$ from the rollout policy while keeping the prefix fixed. Let $\tilde{a}_{i,t}^{(k)}$ denote the final answer of the $k$-th continuation. We define the continuation stability $S_{i,t}$ as the fraction of continuations that preserve the original rollout answer $a_i$:

$$S_{i,t} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\left[\tilde{a}_{i,t}^{(k)} = a_i\right], \qquad \tilde{o}_{i,t}^{(k)} \sim \pi_\theta(\cdot \mid s_{i,t}). \tag{4}$$

**Association Metrics.** Within each query, we quantify the monotonic association between prefix answer support $P$ and continuation stability $S$ using the Spearman rank correlation $\rho = \mathrm{Spearman}(P, S)$, where $\rho = 0$ under random ranking.

To investigate whether intermediate reasoning states carry stability signals beyond terminal correctness, we conduct a pilot study on Qwen3-VL-4B-Instruct (Bai et al., 2025a) using 1,000 medical visual questions from MedXpertQA-MM (Zuo et al., 2025). For each question, we sample 8 rollouts, probe six prefix depths along each trajectory, and draw 8 independent continuations per prefix to compute continuation stability $S$. We summarize our primary observations as follows.

As shown in Fig. 1a, we quantify the relationship between prefix answer support $P$ and continuation stability $S$ using the within-question $\rho(P, S)$, which measures whether trajectories with higher prefix support more consistently preserve their original answers. Across all probed depths, $\rho(P, S)$ remains consistently positive and exhibits an overall downward trend as reasoning progresses. This pattern indicates that early intermediate states already retain rich information about subsequent trajectory stability before trajectories gradually stabilize toward their eventual outcomes.

To examine how effectively $P$ discriminates reasoning stability among trajectories that reach identical outcomes, we divide rollouts within each question into high-$P$ and low-$P$ groups using the within-question median. As shown in Fig. 1b, even among rollouts that achieve the correct final answer, this stability signal remains highly discriminative. High-$P$ correct rollouts are substantially more likely to preserve the correct answer under resampled continuations than their low-$P$ counterparts, and this discriminative gap is particularly pronounced at early reasoning depths where the separation between robust and fragile trajectories is most substantial. Consequently, standard outcome-based RLVR inherently risks reinforcing fragile, coincidental successes, underscoring the necessity of using intermediate stability to discern genuinely robust derivations among correct rollouts.

Crucially, however, this stability signal does not simply equate to answer correctness. We evaluate how well $P$ separates correct from incorrect rollouts via pairwise concordance $C = \Pr(P_{\text{correct}} > P_{\text{wrong}})$ and the mean support gap $\Delta P = \mathbb{E}[P_{\text{correct}} - P_{\text{wrong}}]$. As shown in Fig. 1c, both metrics remain close to their random baselines across all reasoning stages, hovering around $C \approx 0.5$ and $\Delta P \approx 0$. Thus, while $P$ reflects how stably a trajectory supports its eventual outcome, it provides negligible evidence about whether that outcome is correct. This indicates that internal stability and factual correctness are fundamentally orthogonal, because high prefix support reflects the model's internal commitment to a reasoning path rather than its objective validity, which cannot be directly rewarded without grounding in terminal verification.

While $P$ reliably captures trajectory stability, it naturally degenerates when all rollouts produce identical answers, reducing $\mathcal{C}_g$ to a singleton and yielding $P = 1$. Crucially, these consensus groups

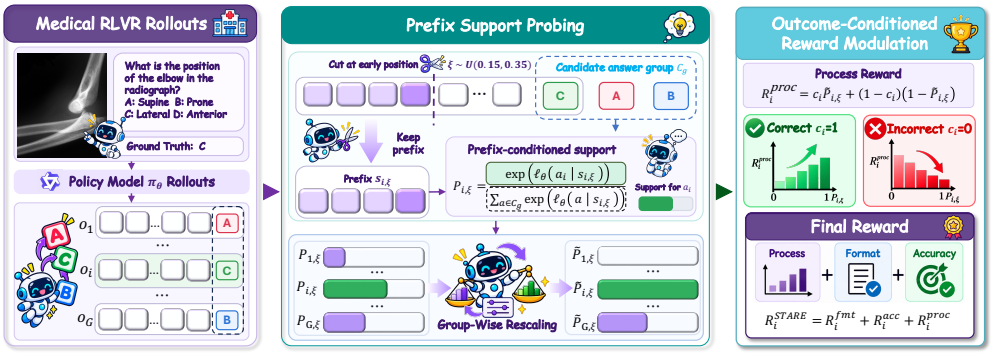



Figure 2: Overview of **STAC** (**ST**ability-**A**ware **C**redit Assignment). Prefix Support Probing extracts early prefixes to estimate reasoning stability via group-rescaled answer support. Outcome-Conditioned Reward Modulation couples this support with terminal correctness, providing finer-grained supervision beyond final-answer correctness for policy optimization.

are already highly robust, with continuation correctness exceeding 96% from early prefixes. Detailed pilot study settings and results are presented in Appendix B.

**Takeaway**

While naive RLVR treats the rollouts with the same answer equally, intermediate reasoning states exhibit substantial differences in trajectory stability. Early prefix support strongly predicts continuation stability under resampling, yet remains orthogonal to final-answer correctness. Crucially, this signal distinguishes robust correct reasoning from fragile success well before completion, offering a lightweight, evaluator-free basis for trajectory-level credit assignment.

## 4 METHODOLOGY

Motivated by the above findings, we propose **STAC** (**ST**ability-**A**ware **C**redit Assignment), as illustrated in Fig. 2. STAC enhances policy optimization through two complementary components. **Prefix Support Probing** probes an intermediate reasoning prefix and measures the policy's relative support for the rollout's eventual answer against competing candidates. **Outcome-Conditioned Reward Modulation** weights this support signal by terminal correctness, assigning more credit to trajectories with robust derivations and penalizing stubbornly incorrect reasoning.

**Prefix Support Probing.** Although continuation stability directly reflects the consistency of the reasoning process and its eventual answer, estimating it requires costly repeated continuation sampling. Since the pilot study shows that prefix answer support $P$ closely tracks continuation stability, with the strongest association occurring at early reasoning stages, we use $P$ as a lightweight proxy and focus our probing on the early region. To capture this informative early region without relying on a rigid reasoning depth, for each rollout group $\mathcal{G} = \{o_i\}_{i=1}^{G}$, we sample a relative prefix position $\xi \sim \mathcal{U}(0.15, 0.35)$ and determine the truncation step $t$ by locating the valid reasoning boundary nearest to $t = \xi \cdot |z_i|$, yielding the intermediate state $s_{i,t}$.

Given the intermediate state $s_{i,t}$, we compute the prefix answer support $P_{i,t}$ over the candidate set $\mathcal{C}_g$ defined in Eq. equation 2 following the scoring function in Eq. equation 3. The resulting $P_{i,t}$ quantifies how strongly the prefix supports the rollout's eventual answer relative to competing candidates, rather than whether the answer is correct. To account for cross-question variation in the scale of $P_{i,t}$ while preventing negligible within-group differences from inducing spurious rankings, we apply the following group-wise rescaling:

$$\widetilde{P}_{i,t} = \begin{cases} \dfrac{P_{i,t} - P_{\min}}{P_{\max} - P_{\min}}, & P_{\max} - P_{\min} > \epsilon, \\ 0.5, & \text{otherwise}, \end{cases} \tag{5}$$

where $P_{\min} = \min_{j \in \{1,\dots,G\}} P_{j,t}$ and $P_{\max} = \max_{j \in \{1,\dots,G\}} P_{j,t}$. When within-group differences vanish, particularly in the unanimous consensus boundary case where all rollouts produce identical

correct answers, $\widetilde{P}_{i,t}$ smoothly falls back to a neutral value of $0.5$, avoiding spurious credit differentiation on fully mastered questions. A higher $\widetilde{P}_{i,t}$ indicates stronger relative support for the eventual answer within the group, whereas a lower value indicates weaker support.

**Outcome-Conditioned Reward Modulation.** Although $\widetilde{P}_{i,t}$ captures the relative strength of a rollout's support for its eventual answer, it does not indicate whether such support should be reinforced or suppressed. We therefore condition the support signal on terminal correctness. Let $c_i \in \{0, 1\}$ indicate whether the final answer of rollout $i$ is correct, and define the process score as:

$$R_i^{\text{proc}} = c_i \, \widetilde{P}_{i,t} + (1 - c_i)(1 - \widetilde{P}_{i,t}). \tag{6}$$

For correct rollouts with $c_i = 1$, a higher $\widetilde{P}_{i,t}$ yields a larger process reward, assigning more credit to trajectories that stably derive the verified answer while discounting fragile successes that reach the correct outcome by chance. Conversely, for incorrect rollouts with $c_i = 0$, the relation is inverted: trajectories that stubbornly adhere to erroneous clinical deductions receive lower credit, whereas exploratory trajectories with lower prefix support retain higher scores. In this way, $R_i^{\text{proc}}$ complements terminal correctness with intermediate stability information, guiding the policy toward robust reasoning while preserving exploratory flexibility.

**Overall Reward.** The final reward for policy optimization combines format compliance, answer correctness, and the process signal:

$$R_i^{\text{STAC}} = R_i^{\text{fmt}} + R_i^{\text{acc}} + R_i^{\text{proc}}, \tag{7}$$

where $R_i^{\text{fmt}}$, $R_i^{\text{acc}}$, and $R_i^{\text{proc}}$ account for response format compliance, final-answer correctness, and intermediate reasoning stability, respectively. This composite reward is directly used to compute the advantage estimate $\hat{A}_i$ within GRPO for policy optimization. Detailed specifications of the format verifiers and the exact weightings of each reward component are provided in Appendix C.

## 5 Experiments

In this section, we systematically evaluate **STAC** across multiple medical vision-language benchmarks, with Qwen3-VL-4B-Instruct (Bai et al., 2025a) as the primary backbone. Through rigorous comparative experiments, component ablations, and fine-grained trajectory analyses, we aim to address the following research questions:

- **RQ1:** Does STAC achieve superior diagnostic accuracy compared with existing open-source MLLMs and outcome-based RLVR baselines across diverse medical benchmarks?
- **RQ2:** How do the core designs contribute to the overall effectiveness of STAC?
- **RQ3:** Beyond final-answer accuracy, does STAC genuinely enhance intermediate reasoning stability while mitigating stubborn error commitment?
- **RQ4:** Does STAC consistently generalize across different model scales and different training recipes, such as the Zero-RL setting?

### 5.1 Experimental Setup

**Evaluation Benchmarks.** We evaluate STAC on six medical vision-language benchmarks: PMC-VQA (Zhang et al., 2023), OmniMedVQA (OMVQA) (Hu et al., 2024), MedXpertQA-MM (MedXQA) (Zuo et al., 2025), VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and PathVQA (He et al., 2020). These benchmarks span diverse multimodal medical reasoning alongside fine-grained, image-grounded diagnostic understanding across several clinical modalities.

**Baselines.** We compare STAC with general and medical MLLMs, plus two RLVR controls. General baselines include Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-4B-Instruct (Bai et al., 2025a), Gemma3-4B (Team et al., 2025), and InternVL3-8B (Zhu et al., 2025). Medical models include MedVLM-R1-2B (Pan et al., 2025), MedGemma-4B-IT (Sellergren et al., 2025), LLaVA-Med-7B (Li et al., 2023), HuatuoGPT-Vision (Chen et al., 2024), Lingshu-7B (Xu et al., 2025), and BioMedIX2-8B (Mullappilly et al., 2025). Standard GRPO (DeepSeek-AI et al., 2025) serves as the vanilla RLVR baseline, while CARE (Du et al., 2026) augments RLVR with confidence signals.

**Implementation Details.** STAC is initialized from Qwen3-VL-4B-Instruct (Bai et al., 2025a). We first perform a lightweight SFT warm-up to establish a consistent reasoning-answer format, then

Table 1: **Comparison of multimodal large language models on medical vision-language benchmarks.** The best results among models with 2–4B parameters are highlighted in **bold**. For our models, we addtionally report the standard deviation $\sigma$ over 4 runs.

| Model | Size | PMC-VQA | MedXQA | OMVQA | VQA-RAD | SLAKE | PathVQA | Avg. |
|---|---|---|---|---|---|---|---|---|
| *General Open-source Models* | | | | | | | | |
| Gemma3-4B-IT | 4B | 48.5 | 21.0 | 60.9 | 49.9 | 61.1 | 43.2 | 47.4 |
| Qwen3-VL-4B-Instruct | 4B | 53.4 | 25.4 | 77.9 | 55.8 | 69.5 | 42.1 | 54.0 |
| Qwen2.5VL-7B-Instruct | 7B | 51.4 | 21.7 | 64.5 | 64.7 | 67.3 | 44.5 | 52.4 |
| InternVL3-8B | 8B | 53.8 | 22.4 | 79.1 | 65.4 | 72.8 | 48.6 | 57.0 |
| *Medical Open-source Models* | | | | | | | | |
| MedVLM-R1-2B | 2B | 48.3 | 20.4 | 77.7 | 48.6 | 56.0 | 32.5 | 47.3 |
| MedGemma-4B-IT | 4B | 49.9 | 22.3 | 69.8 | **72.5** | 76.4 | 48.8 | 56.6 |
| LLaVA-Med-7B | 7B | 30.5 | 20.3 | 44.3 | 53.7 | 48.0 | 38.8 | 39.3 |
| HuatuoGPT-V-7B | 7B | 53.3 | 21.6 | 74.2 | 67.0 | 67.8 | 48.0 | 55.3 |
| Lingshu-7B | 7B | 56.3 | 26.7 | 82.9 | 67.9 | 83.1 | 61.9 | 63.1 |
| BioMedxX2-8B | 8B | 44.8 | 21.8 | 63.3 | 49.2 | 57.7 | 37.0 | 45.6 |
| *RLVR Methods* | | | | | | | | |
| GRPO | 4B | 56.3 | 26.6 | 80.5 | 58.3 | 82.6 | 64.9 | 61.5 |
| CARE | 4B | 56.5 | 26.4 | 80.8 | 58.7 | 82.8 | 65.3 | 61.8 |
| **STAC** | **4B** | **57.0**$_{\pm 0.42}$ | **27.5**$_{\pm 0.61}$ | **81.2**$_{\pm 0.21}$ | 59.9$_{\pm 0.39}$ | **83.2**$_{\pm 0.27}$ | **65.6**$_{\pm 0.47}$ | **62.4**$_{\pm 0.40}$ |

optimize the policy with RLVR. Training data are drawn from PMC-VQA (Zhang et al., 2023), PathVQA (He et al., 2020), SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), MedQA-USMLE (Jin et al., 2020), and GMAI-Reasoning10K (Su et al., 2025). For datasets also used in evaluation, optimization uses only their training splits, and the SFT and RLVR training sets are disjoint. More details about the training setup and implementation are provided in Appendix D.

### 5.2 Results on Various Medical Benchmarks (RQ1)

Table 1 compares STAC with competing models and rlvr methods across six medical vision-language benchmarks, yielding three key observations.

**(1) Consistent Gains over Outcome-only RLVR.** STAC achieves the best performance among Qwen3-VL-4B-based RLVR methods on all six benchmarks, outperforming both standard GRPO and CARE under the same backbone and training setup. This controlled comparison isolates the effect of reward design. Unlike outcome-only or terminal-confidence supervision, STAC uses prefix-conditioned answer support to measure how strongly the eventual answer is supported at intermediate states. Combined with terminal correctness, it reinforces stable support for correct outcomes while suppressing commitment to incorrect ones, providing fine-grained supervisory guidance well beyond coarse binary final-answer correctness and spurious deductions.

**(2) Robustness across Diverse Task Formats.** The evaluation suite spans multiple-choice reasoning on PMC-VQA, OMVQA, and MedXQA, alongside free-form answers on VQA-RAD, SLAKE, and PathVQA across radiology, pathology, and general medicine. STAC improves consistently across these heterogeneous settings, confirming that its benefits are not tied to a specific response format. Because process rewards are conditioned on early prefixes, supervision directly targets the entire reasoning trajectory rather than superficial final response forms or dataset artifacts.

**(3) Competitive Performance at a Smaller Scale.** STAC improves over the original Qwen3-VL-4B-Instruct model across all benchmarks. With only 4B parameters, it surpasses InternVL3-8B in average performance, approaches Lingshu-7B (Xu et al., 2025), and outperforms several larger medical MLLMs. These results demonstrate that trajectory-level credit assignment yields substantial diagnostic gains without scaling parameters or extensive pretraining, making a compact 4B model competitive with substantially larger generalist systems.

### 5.3 Ablation Study of STAC (RQ2)

To examine the design choices behind STAC, we study three aspects: prefix selection, the trajectory specificity of prefix support, and the role of terminal correctness in reward modulation. All variants start from the same Qwen3-VL-4B-SFT checkpoint and follow the same RLVR training setup.

**Prefix Selection.** We compare the default early-window strategy with a later window and fixed-prefix variants. As shown in Table 2, earlier prefixes outperform later ones in both settings: 62.4 vs. 61.9

Table 2: **Ablation study of STAC.** The best results are in **bold**, and second best are underlined. Default prefix selection is $\xi \sim \mathcal{U}(0.15, 0.35)$ of the reasoning trajectory.

| Variant | Reward Design | | | Accuracy | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Prefix | Support | Outcome-Aware | PMC-VQA | MedXQA | OMVQA | VQA-RAD | SLAKE | PathVQA | Avg. |
| Qwen3VL-4B-SFT | - | - | - | 55.6 | 25.8 | 79.1 | 54.54 | 80.2 | 58.3 | 58.9 |
| Late Prefix Sampling | 65%–85% | Aligned | ✓ | 56.6 | 26.5 | 80.9 | 59.2 | 83.1 | 65.3 | 61.9 |
| Fixed Prefix | 25% | Aligned | ✓ | 56.7 | 26.3 | 80.5 | 59.0 | **83.2** | 65.3 | 61.8 |
| Fixed Prefix | 75% | Aligned | ✓ | 56.2 | 26.2 | 80.2 | 59.0 | 83.0 | 65.1 | 61.6 |
| Shuffled Support | Default | Shuffled | ✓ | 55.9 | 25.4 | 79.7 | 58.6 | 83.0 | 65.2 | 61.3 |
| Direct Prefix Support | Default | Aligned | × | 55.5 | 25.3 | 80.1 | 57.9 | 82.7 | 65.1 | 61.1 |
| **STAC** | **Default** | **Aligned** | ✓ | **57.0** | **27.5** | **81.2** | **59.9** | **83.2** | **65.6** | **62.4** |

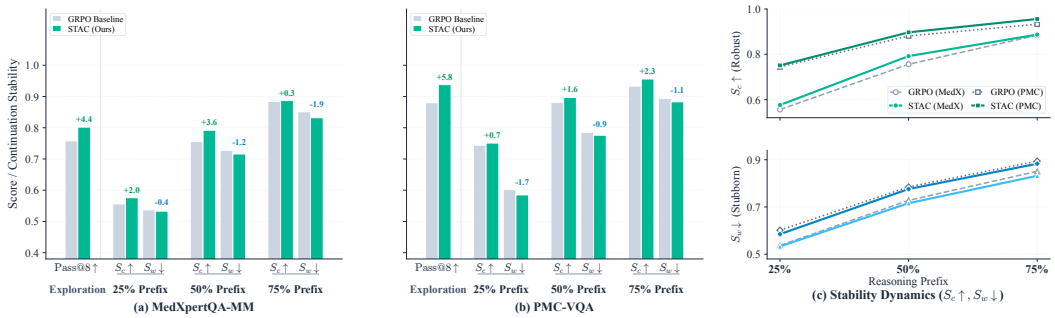



Figure 3: **Reasoning stability across prefix depths.** Fig.(a) and (b) show exploration ceiling Pass@8 alongside stage-wise stability $S_c$ and $S_w$. Fig.(c) shows dynamic trajectory curves of $S_c$ and $S_w$ across reasoning stages, where $c$ and $w$ denote correct and incorrect rollouts, respectively.

for dynamic sampling and 61.8 vs. 61.6 for fixed prefixes at 25% and 75%, respectively. This agrees with our pilot study, where reliability differences emerge before continuation behavior stabilizes, indicating that early states contain more information about future trajectory stability. Dynamic sampling also outperforms the corresponding fixed positions by 0.6 and 0.3 points, respectively. These results confirm that dynamic early-window sampling provides the most informative stability signal by capturing early reasoning divergence while accommodating varying trajectory lengths.

**Trajectory-Specific Support.** We investigate whether STAC's benefits arise from fine-grained credit assignment rather than a non-specific reward shift by permuting support scores across rollouts, thereby preserving their marginal distribution while severing the link to each originating trajectory. Table 2 shows that disrupting this alignment degrades average accuracy from 62.4 to 61.3, with the most pronounced drop occurring on the reasoning-intensive MedXQA from 27.5 to 25.4. This substantial gap demonstrates that process supervision cannot function as a generic regularizer; instead, prefix support operates as an effective trajectory-level signal intrinsically anchored in each rollout's actual step-by-step reasoning dynamics and clinical deductions.

**Outcome-Aware Modulation.** While prefix support captures continuation stability, it does not inherently guarantee diagnostic correctness. Setting $c_i \equiv 1$ for all rollouts, corresponding to Direct Prefix Support in Table 2, removes outcome modulation and lowers average accuracy from 62.4 to 61.1, marked by substantial drops from 57.0 to 55.5 on PMC-VQA and from 27.5 to 25.3 on MedXQA. This degradation directly reinforces our pilot finding that internal reasoning stability and factual correctness are fundamentally orthogonal: high prefix support reflects continuation commitment rather than clinical veracity. Consequently, unconditioned stability cannot serve as an autonomous optimization target; anchoring this internal stability to verified terminal outcomes is essential to prevent the policy from entrenching into confident yet erroneous clinical derivations.

### 5.4 In-depth Analysis of STAC (RQ3)

A central premise of STAC is that effective RLVR should not merely reinforce terminal outcomes, but actively align internal reasoning stability with diagnostic validity, promoting robust derivations while discouraging stubborn commitment to erroneous pathways. To evaluate whether STAC reshapes reasoning dynamics beyond surface accuracy, we conduct continuation probing across reasoning stages on held-out test splits from MedXpertQA-MM and PMC-VQA. As visualized in Figure 3, we

Table 3: Performance of STAC versus baselines across model scales in the zero-RL setting.

| Model | Method | PMC-VQA | MedXQA | OMVQA | Avg. |
|---|---|---|---|---|---|
| Qwen3-VL-2B-Instruct | Zero-shot | 48.9 | 21.3 | 68.8 | 46.3 |
| | GRPO | 49.3 | 22.2 | 69.1 | 46.9 |
| | **STAC** | **49.6** | **23.2** | **69.9** | **47.5** |
| Qwen3-VL-4B-Instruct | Zero-shot | 53.4 | 25.4 | 77.9 | 52.2 |
| | GRPO | 53.6 | 25.8 | 78.1 | 52.5 |
| | **STAC** | **54.1** | **26.1** | **78.7** | **53.0** |

evaluate continuation stability on correct ($S_c$) and incorrect ($S_w$) trajectories alongside exploration capacity $\mathrm{Pass@8}$, deferring full experimental protocols to Appendix E.

**Reinforcing Robust Successes over Fragile Guesses.** Across all probed depths on both benchmarks, STAC consistently yields higher continuation stability on correct rollouts $S_c$ than standard GRPO. This margin emerges as early as the 25% prefix and persists throughout the reasoning trajectory, confirming that correct predictions are grounded in reliable intermediate states anticipating the target answer rather than fragile stochastic fluctuations. Penalizing low-support correct rollouts during training suppresses chance successes, establishing robust reasoning pathways that reliably preserve their diagnostic validity and consistency under repeated continuation resampling.

**Mitigating Stubborn Errors and Overconfidence.** A critical requirement for clinical reasoning is that intermediate stability must not translate into indiscriminate stubbornness on flawed paths. Evaluating trajectory stability on incorrect rollouts $S_w$ confirms this distinction, as $S_w$ consistently decreases under STAC across both datasets and all probed prefixes. The resulting divergence between $S_c$ and $S_w$ in Figure 3 directly reflects the foundational insight from our pilot study that persistence does not equal correctness. While standard GRPO penalizes incorrect rollouts uniformly regardless of commitment, STAC imposes an oriented penalty on trajectories that stubbornly adhere to false hypotheses. This encourages the model to retain exploratory flexibility under diagnostic uncertainty rather than entrenching flawed deductions into premature commitment.

**Enhanced Exploratory Diversity and Solution Ceiling.** Beyond intermediate trajectory stability, STAC expands the policy exploration frontier, yielding notable gains in $\mathrm{Pass@8}$ that markedly elevate the multi-sample solution ceiling. By mitigating premature commitment to erroneous trajectories and refining credit assignment among correct rollouts, STAC preserves sampling diversity, enabling the model to discover valid clinical pathways that standard outcome-only optimization overlooks.

### 5.5 Generalization Across Model Scales and Training Settings (RQ4)

To assess whether STAC generalizes across model scales and training paradigms, we optimize Qwen3-VL-2B-Instruct and Qwen3-VL-4B-Instruct under the Zero-RL setting, optimizing instruct backbones directly via RL without prior SFT warm-up. Across both architectures under matched rollout budgets, STAC consistently surpasses Zero-shot baselines and standard GRPO on all evaluated benchmarks (Table 3). Crucially, the persistent performance margin across varying parameter scales highlights that trajectory-level credit assignment functions independently of model capacity. Even in the complete absence of supervised demonstrations to scaffold reasoning formats, the framework reliably outperforms outcome-only RLVR, achieving notable gains on complex clinical reasoning and out-of-domain evaluation. Prefix-conditioned stability thereby acts as an intrinsic supervisory signal, facilitating robust credit assignment and stable policy optimization in demonstration-free regimes.

## 6 Conclusion

In this work, we presented **STAC**, an evaluator-free framework that refines credit assignment in medical multimodal RLVR by leveraging the policy's intrinsic trajectory dynamics. By coupling Prefix Support Probing with Outcome-Conditioned Reward Modulation, STAC selectively rewards stably supported correct reasoning while penalizing stubborn commitment to erroneous clinical pathways, without external evaluators, dense annotations, or rollout branching. Across six medical vision-language benchmarks, STAC consistently outperforms standard RLVR baselines, achieving superior diagnostic accuracy, enhanced reasoning stability, and broader exploration capacity. These findings demonstrate that policy-derived stability offers a principled and lightweight supervisory signal, advancing reliable clinical deduction in multimodal foundation models.

### AI USE STATEMENT

In this work, we used generative AI tools for manuscript polishing, LaTeX formatting and typesetting assistance, debugging support during code implementation, and creating selected decorative icons used in the paper layout. All main experimental plots and quantitative result figures were generated manually with standard plotting tools (e.g., `matplotlib`) and were not produced by generative AI. We have not used generative AI tools for core research question formulation, research ideation, experimental design, experimental setup, result analysis, or generating the main scientific claims of this paper. Disclosure categories related to AI-generated research ideas, AI-designed experiments, and AI-authored analysis are therefore not applicable to this work. We reviewed all AI-assisted outputs through manual literature checks, author verification of technical claims, manual inspection and testing of any AI-suggested code changes, and author review of AI-generated icons before incorporation. We take responsibility for the final content of this work, including text, claims, and other artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work studies medical vision-language reasoning on publicly available benchmark datasets and does not involve newly collected human-subject studies or clinical trials conducted by the authors. All experiments are performed offline on existing datasets; we do not deploy the model in real clinical workflows or provide direct patient care. We acknowledge that medical AI systems can exhibit dataset bias, modality imbalance, and performance disparities across populations; results on benchmark splits should not be interpreted as clinical-grade diagnostic evidence. We use publicly released datasets under their original licenses and do not release additional personally identifiable patient information. Potential misuse includes over-reliance on model outputs for medical decision-making; we explicitly position STAC as a research method rather than a certified clinical product.

### REPRODUCIBILITY STATEMENT

We provide implementation and training details in Section 5 and Appendix C and Appendix D, including reward specifications (Appendix C.1), the STAC training algorithm (Appendix C.3), dataset splits (Appendix D.1), hyperparameters and hardware configuration (Appendix D.2), and evaluation protocols (Appendix D.3). All benchmarks used in this study are publicly available, and we follow each dataset's official train/test protocol with disjoint SFT and RLVR training splits where applicable.

## REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL `https://arxiv.org/abs/2511.21631`.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal llms at scale. *arXiv preprint arXiv:2406.19280*, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Yuetian Du, Yucheng Wang, Zhenyuan Chen, Luyuan Chen, Rongyu Zhang, Jinjian Zhang, Wei Zhou, Zhijie Xu, Ming Kong, Zhan Zhou, Jie Liu, and Qiang Zhu. Care: Confidence-aware reasoning for reliable medical vqa, 2026. URL https://arxiv.org/abs/2608.10964.

Kailong Fan, Anqi Pu, Yichen Wu, Wanhua Li, Yicong Li, Hanspeter Pfister, Huafeng Liu, Xiang Li, Quanzheng Li, and Ning Guo. Maple: Elevating medical reasoning from statistical consensus to process-led alignment, 2026. URL https://arxiv.org/abs/2603.08987.

Peixuan Han, Runhui Wang, Ketan Ramaneti, Jie Hao, Gerald Friedland, and Chris Kong. Cliff: Learning process rewards from the first mistake, 2026. URL https://arxiv.org/abs/2609.02817.

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.

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

Songtao Jiang, Yuan Wang, Ruizhe Chen, Yan Zhang, Ruilin Luo, Bohan Lei, Yeying Jin, Sibo Song, ZhiBo Yang, Jimeng Sun, Jian Wu, and Zuozhu Liu. Act as you think: Reinforcing consistent reasoning in medical visual question answering. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1788–1805, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.81. URL https://aclanthology.org/2026.acl-long.81/.

Zheng Jiang, Heng Guo, Chengyu Fang, Changchen Xiao, Xinyang Hu, Lifeng Sun, and Minfeng Xu. Medvr: Annotation-free medical visual reasoning via agentic reinforcement learning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*, 2020.

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, 2025. URL https://arxiv.org/abs/2503.13939.

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*, 5(1):1–10, 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. *arXiv preprint arXiv:2306.00890*, 2023.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

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 *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pp. 1650–1654. IEEE, 2021.

Guimeng Liu, Tianze Yu, Somayeh Ebrahimkhani, Lin Zhi Zheng Shawn, Kok Pin Ng, and Ngai-Man Cheung. How do medical MLLMs fail? a study on visual grounding in medical images. In *International Conference on Learning Representations (ICLR)*, 2026.

Zeyu Liu, Zhitian Hou, Guanghao Zhu, Zhijie Sang, Congkai Xie, and Hongxia Yang. Infimed: Low-resource medical mllms with advancing understanding and reasoning, 2025. URL https://arxiv.org/abs/2505.23867.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled Aldahmani, Fahad Khan, Rao Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. Bimedix2: Bio-medical expert lmm for diverse medical modalities, 2025. URL https://arxiv.org/abs/2412.07769.

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. *arXiv preprint arXiv:2502.19634*, 2025.

Mohammad Rezaei, Jens Lehmann, and Sahar Vahdati. Llm reasoning with process rewards for outcome-guided steps, 2026. URL https://arxiv.org/abs/2604.02341.

Binesh Sadanandan and Vahid Behzadan. When chain-of-thought backfires: Evaluating prompt sensitivity in medical language models, 2026. URL https://arxiv.org/abs/2603.25960.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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.

Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, Jinghao Lin, Kai Wu, Lin Yang, Shengsheng Yao, Tao Chen, Xiaojun Xiao, Xiaozhong Ji, Xu Wang, Yijun He, and Zhixiong Yang. MedXIAOHE: A comprehensive recipe for building medical MLLMs. *arXiv preprint arXiv:2602.12705*, 2026a. URL https://arxiv.org/abs/2602.12705. XIAOHE Medical AI / ByteDance; see paper for full author list.

Baorong Shi, Bo Cui, Boyuan Jiang, Deli Yu, Fang Qian, Haihua Yang, Huichao Wang, Jiale Chen, Jianfei Pan, Jieqiong Cao, Jinghao Lin, Kai Wu, Lin Yang, Shengsheng Yao, Tao Chen, Xiaojun Xiao, Xiaozhong Ji, Xu Wang, Yijun He, and Zhixiong Yang. Medxiaohe: A comprehensive recipe for building medical mllms, 2026b. URL https://arxiv.org/abs/2602.12705.

Yanzhou Su, Tianbin Li, Jiyao Liu, Chenglong Ma, Junzhi Ning, Cheng Tang, Sibo Ju, Jin Ye, Pengcheng Chen, Ming Hu, et al. Gmai-vl-r1: Harnessing reinforcement learning for multimodal medical reasoning. *arXiv preprint arXiv:2504.01886*, 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Yuan Wu, Zongxian Yang, Jiayu Qian, Songpan Gao, Guanxing Chen, Qiankun Li, Yu-An Huang, and Zhi-An Huang. Better eyes, better thoughts: Why vision chain-of-thought fails in medicine, 2026. URL https://arxiv.org/abs/2603.06665.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Jaehoon Yun, Jiwoong Sohn, Jungwoo Park, Hyunjae Kim, Xiangru Tang, Yanjun Shao, Yonghoe Koo, Minhyeok Ko, Qingyu Chen, Mark Gerstein, Michael Moor, and Jaewoo Kang. Med-prm: Medical reasoning models with stepwise, guideline-verified process rewards, 2025. URL https://arxiv.org/abs/2506.11474.

Anas Zafar, Leema Krishna Murali, Siddhant Bharadwaj, Ashish Vashist, and Jia Wu. Do medical vision language models actually see? a counterfactual grounding framework and hard-negative contrastive training for visually-reliant medical VLMs. *arXiv preprint arXiv:2607.03647*, 2026. URL https://arxiv.org/abs/2607.03647.

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.

Zheng Zhang, Ziwei Shan, Kaitao Song, Yexin Li, and Kan Ren. Linking process to outcome: Conditional reward modeling for llm reasoning, 2025. URL https://arxiv.org/abs/2509.26578.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 392–400, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhang-Ren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking expert-level medical reasoning and understanding. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80961–80990. PMLR, 2025. URL https://proceedings.mlr.press/v267/zuo25a.html.

# A CASE STUDY: LIMITATION OF OUTCOME-ONLY RLVR

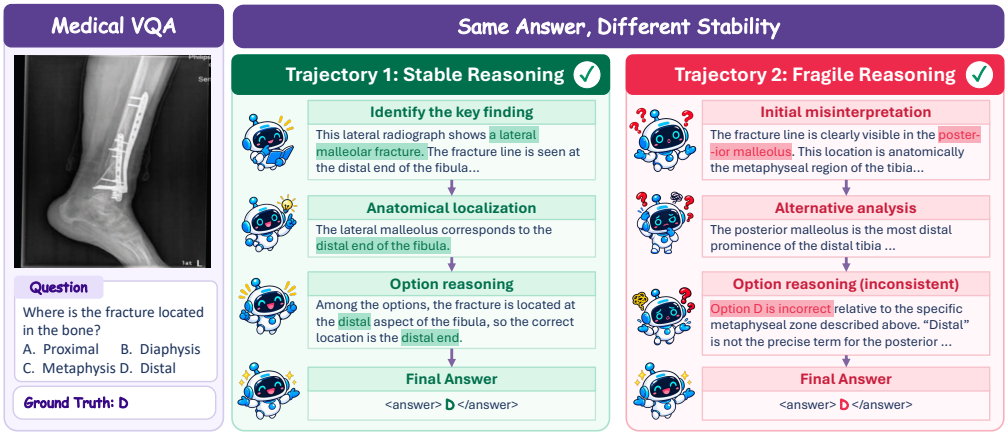



Figure 4: **Same Answer, Different Stability.** Two trajectories yield the same correct answer through either stable reasoning (left) or fragile guessing (right). Standard outcome-based RLVR treats them identically, mistakenly reinforcing lucky guesses.

To provide concrete clinical intuition for the limitations of outcome-only RLVR, Figure 4 illustrates two representative reasoning trajectories generated by the rollout policy for a radiographic VQA query. While both trajectories arrive at the identical ground-truth option (**D**), their internal reasoning paths exhibit fundamentally different clinical validity:

- **Trajectory 1 (Stable Reasoning):** The model correctly identifies the lateral malleolar fracture at the distal fibula and systematically grounds the anatomical localization in visual evidence. Intermediate states along this trajectory provide consistent, high prefix support for the eventual outcome.
- **Trajectory 2 (Fragile Guessing):** The model misinterprets the fracture location, presents internally contradictory deductions that explicitly argue against option D, yet fortuitously outputs the correct answer tag at the final step.

Under conventional RLVR, both rollouts receive identical positive reward simply because their final answers are correct. This unselective credit assignment inadvertently reinforces hallucinated visual features and language shortcuts. In contrast, STAC probes the intermediate prefix support to distinguish robust clinical deductions from fragile coincidences, guiding the policy toward trajectories that are both diagnostically correct and internally well-supported.

# B SUPPLEMENTARY PILOT DETAILS

## B.1 SAMPLING PROTOCOL AND IMPLEMENTATION DETAILS

Consistent with Sec. 3.2, the pilot evaluates Qwen3-VL-4B-Instruct (Bai et al., 2025a) on $1{,}000$ randomly sampled MedXpertQA-MM Zuo et al. (2025) questions. For each question we sample $G{=}8$ rollouts and probe six relative prefix positions

$$\alpha \in \{0.15, 0.25, 0.35, 0.50, 0.65, 0.75\}, \tag{8}$$

measured as a fraction of the generated reasoning token length. At each target $\alpha$, we truncate to the nearest valid reasoning span within an absolute position tolerance of $0.08$ (token-boundary alignment). From each retained prefix we draw $K{=}8$ independent continuations to estimate stability $S$.

**Decoding hyperparameters.** Unless otherwise noted, rollouts and continuations share the same decoding configuration: temperature $1.0$, top-$p{=}1.0$, and a maximum of $4{,}096$ newly generated tokens. Candidate answers are normalized before deduplication. Prefix support scores use the original token sequence of each prefix together with length-normalized mean token log-probability, followed by softmax normalization over the candidate set $\mathcal{C}_g$.

### B.2 Details of Pilot study

Table 4 reports the complete numerical evaluation across the six probed prefix positions, providing quantitative details that complement Figure 1. Beyond the within-query correlation $\rho_{\text{within}}$ and correctness concordance $C_{\text{correct vs wrong}}$ shown in the main text, the table presents the overall Spearman correlation $\rho_{\text{overall}}$ across all queries alongside the continuous within-query pairwise concordance $C_{\text{within}} = \Pr(P_i > P_j \mid S_i > S_j)$. Under random ranking, rank correlation is zero and concordance is $0.5$.

Table 4: **Prefix support metrics across reasoning positions.** Random concordance is $0.5$, whereas random rank correlation is zero. Overall associations incorporate cross-query difficulty variance, while within-query metrics isolate relative trajectory rankings for the identical prompt. Within-query rank correlations correspond to Figure 1a, and correct versus incorrect concordance aligns with Figure 1c.

| Position | $\rho_{\text{overall}}(P, S)$ | $\rho_{\text{within}}(P, S)$ | $C_{\text{within}}(P, S)$ | $C_{\text{correct vs wrong}}$ |
|---|---|---|---|---|
| 15% | 0.633 | 0.418 | 0.674 | 0.506 |
| 25% | 0.584 | 0.399 | 0.696 | 0.529 |
| 35% | 0.544 | 0.410 | 0.739 | 0.522 |
| 50% | 0.474 | 0.370 | 0.739 | 0.539 |
| 65% | 0.395 | 0.358 | 0.782 | 0.555 |
| 75% | 0.378 | 0.368 | 0.759 | 0.554 |

The tabular breakdown highlights two granular properties. First, the pronounced decline in $\rho_{\text{overall}}$ from $0.633$ to $0.378$ is primarily driven by intrinsic question difficulty, where challenging cases exhibit lower stability across all generated rollouts. Conditioning on trajectories for the same prompt removes this cross-query variance, keeping $\rho_{\text{within}}$ robustly bounded between $0.358$ and $0.418$ with the strongest association emerging at the earliest $15\%$ position. Second, pairwise concordance $C_{\text{within}}$ consistently spans $0.674$ to $0.782$, verifying that relative support reliably ranks trajectory stability. In contrast, $C_{\text{correct vs wrong}}$ remains between $0.506$ and $0.555$, demonstrating that early prefix support captures deductive commitment rather than factual correctness. These findings support sampling probing positions within the early reasoning window from $0.15$ to $0.35$, leading to the subsequent question of whether token-level uncertainty can further refine checkpoint placement.

### B.3 Evaluating Entropy-Guided Localization for Continuation Stability

In designing process-level signals, an intuitive alternative hypothesis is that local token-level uncertainty, such as aggregated token entropy or sharp entropy spikes, might pinpoint the critical branch points of clinical reasoning. Under this hypothesis, intermediate states exhibiting the highest predictive entropy would correspond to pivotal deduction forks where subsequent continuations are most susceptible to divergence, thereby yielding lower continuation stability. To empirically examine whether local entropy peaks can reliably localize critical reasoning junctures, we conduct a continuation-based localization experiment within valid reasoning spans.

Table 5: **Checkpoints evaluated in the continuation-based localization test.** Probed positions assess whether peak local uncertainty captures critical answer bifurcation points in reasoning trajectories.

| **Checkpoint** | **Construction** | **Role** |
|---|---|---|
| Highest-entropy | Peak aggregated entropy span within valid reasoning | Tests whether peak uncertainty marks decision forks |
| Matched low-entropy | Position- and length-matched span with lower entropy | Isolates local entropy under position and length controls |
| Random legal | Arbitrary non-overlapping valid reasoning span | Tests if entropy guidance beats arbitrary positions |

As summarized in Table 5, we construct three contrasting probing checkpoints for each reasoning trace. The highest-entropy checkpoint identifies the span exhibiting the maximal aggregated entropy score across the trajectory. To control for the confounding influence of position depth and span

Table 6: **Paired differences in continuation stability.** Point estimates and $95\%$ bootstrap confidence intervals demonstrate that local entropy peaks do not induce significant shifts in continuation stability relative to matched or random controls.

| Comparison | Stability Difference | 95% CI |
|---|---|---|
| High entropy vs matched low entropy | $-0.0047$ | $[-0.0238, 0.0134]$ |
| High entropy vs random legal | $0.0147$ | $[-0.0075, 0.0368]$ |

duration, the matched low-entropy checkpoint selects an alternative span of comparable length and relative position that exhibits lower entropy. Finally, the random legal checkpoint samples an arbitrary span from non-overlapping, non-adjacent valid reasoning segments, serving as an unguided control. From each checkpoint, we independently sample continuation rollouts and record continuation stability $S$. The treatment effect of local entropy on continuation stability is evaluated through paired differences against both control configurations across $N$ evaluated trajectories:

$$\Delta_{\text{high}-\text{control}} = \frac{1}{N} \sum_{i=1}^{N} \left( S_i^{\text{high}} - S_i^{\text{control}} \right). \tag{9}$$

Table 6 reports the paired stability differences along with their $95\%$ bootstrap confidence intervals. Comparing high-entropy spans against matched low-entropy counterparts yields a slight stability difference of $-0.0047$, with a confidence interval of $[-0.0238, 0.0134]$ spanning zero. Similarly, comparing high-entropy spans against random legal positions produces a stability difference of $0.0147$, with a confidence interval of $[-0.0075, 0.0368]$ that also overlaps zero.

These empirical measurements reveal that local token entropy spikes do not reliably delineate critical decision boundaries in clinical multimodal reasoning. Local token-level uncertainty often reflects surface lexical ambiguity or syntactic phrasing choices rather than genuine clinical divergence. Consequently, filtering or weighting intermediate states based purely on local entropy provides no measurable advantage over position-matched or random sampling. This negative finding motivates the design of STAC, which departs from brittle token-level entropy localization and instead probes global prefix answer support across a continuous early reasoning window.

### B.4 Boundary Cases: All-correct groups

Table 7: **Continuation correctness on all-correct groups.** Evaluated across all-correct groups where all $8$ original rollouts reach the correct answer and $P \equiv 1$. At each probed prefix depth, each of the $8$ rollouts is independently continued $8$ times, generating $64$ continuations per group.

| **Prefix** | 15% | 25% | 35% | 50% | 65% | 75% |
|---|---|---|---|---|---|---|
| **Continuation correctness** | 0.963 | 0.968 | 0.974 | 0.983 | 0.988 | 0.993 |

When every rollout for a question yields the same correct answer, the candidate set collapses to a singleton and $P \equiv 1$ for all trajectories by construction. Prefix support therefore cannot rank trajectories inside such groups. For all-correct groups where all $8$ initial rollouts produce the ground-truth outcome, we assess their reasoning stability by probing each trajectory across the six prefix checkpoints and independently sampling $8$ continuations from each prefix state, yielding $64$ resampled continuations per query group at each prefix position.

As reported in Table 7, resampled continuations remain correct with empirical probability $0.963$ even at the earliest $15\%$ prefix, rising monotonically to $0.993$ at the $75\%$ depth. Across all six checkpoints, $91\%$ of individual rollout prefix states preserve the correct answer across all $8$ resampled continuations. Because unanimous correct groups already possess high intrinsic stability, fine-grained within-group ranking is unnecessary. This observation directly supports the group-wise rescaling fallback in Eq. equation 5, where neutral support $0.5$ is assigned whenever $P_{\max} - P_{\min} \leq \epsilon$ to prevent inducing spurious rankings from negligible support differences.

# C Detailed Specifications of the Overall Reward

As formulated in Eq. equation 7 of Sec. 4, STAC guides RLVR policy optimization through an overall composite reward combining format compliance, terminal answer correctness, and intermediate derivation stability:

$$R_i^{\text{STAC}} = R_i^{\text{fmt}} + R_i^{\text{acc}} + R_i^{\text{proc}}, \tag{10}$$

which is directly used to compute the trajectory advantage $\hat{A}_i$ within the group relative policy optimization loop. This section provides the detailed verification rules, scoring protocols, and component calibrations for each term.

## C.1 Deterministic Verifiers and Reward Calibration

To prevent reward gaming and hallucinated credit assignment, reward evaluation in STAC relies strictly on programmatic, rule-based verifiers without auxiliary neural scoring models. Each reward component is governed by explicit verification specifications:

- **Format Verifier ($R_i^{\text{fmt}}$):** Evaluates structural presentation with a two-part reward summing to $0.2$. Specifically, a rollout receives $0.1$ if its final answer is properly enclosed within `<answer>...</answer>` tags, and receives an additional $0.1$ if a reasoning process precedes the answer. A rollout satisfying both criteria receives $R_i^{\text{fmt}} = 0.2$. If the final answer is not enclosed within `<answer>` tags, answer extraction fails, yielding $R_i^{\text{fmt}} = 0$ and automatically setting terminal correctness to zero.
- **Outcome Verifier ($R_i^{\text{acc}}$):** Evaluates final-answer correctness against ground truth. For multiple-choice questions, the verifier extracts candidate options within the answer block, stripping extraneous whitespace and punctuation to match option letters in $\{A, B, C, D, \dots\}$. For open-ended questions, the verifier computes word-level Jaccard similarity between the extracted answer and the reference text. Correct rollouts receive $R_i^{\text{acc}} = 0.8$, and incorrect ones receive $0$.
- **Evaluator-Free Process Reward ($R_i^{\text{proc}}$):** Incorporates intermediate stability through outcome-conditioned support modulation (Eq. 6), scaled by a coefficient of $0.2$:

$$R_i^{\text{proc}} = 0.2 \times [c_i \widetilde{P}_{i,t} + (1 - c_i)(1 - \widetilde{P}_{i,t})], \tag{11}$$

  where $c_i = \mathbf{1}[R_i^{\text{acc}} > 0]$ indicates terminal correctness, and $\widetilde{P}_{i,t} \in [0, 1]$ represents the group-rescaled prefix support defined in Eq. 5. When all rollouts reach the identical correct outcome or support differences fall below $\epsilon = 10^{-6}$, $\widetilde{P}_{i,t}$ defaults to neutral $0.5$, yielding a uniform process credit of $0.1$ across all rollouts and leaving relative advantage computation unperturbed.

Under this calibration, terminal accuracy provides the primary optimization gradient, while format compliance and process-level derivation stability supply complementary regularization and credit refinement without overtaking ground-truth accuracy.

## C.2 Prefix Support Computation at Training Time

Crucially, STAC does not branch continuation rollouts during training. Instead, prefix answer support is evaluated through a single parallel forward pass over truncated prefix states:

1. **Probing Position Selection:** For each rollout group, we sample a relative depth ratio $\xi \sim \mathcal{U}(0.15, 0.35)$ and truncate trajectory $o_i$ at the nearest syntactic token boundary to $\xi \cdot |z_i|$, forming prefix state $s_{i,t} = (I, Q, z_{i,\le t})$.
2. **Candidate Set Construction:** We form the candidate pool $\mathcal{C}_g = \text{Unique}(\{a_j\}_{j=1}^{G} \cup \{y\})$ by taking unique answers across group rollouts and the ground-truth label.
3. **Parallel Candidate Scoring:** For each candidate $a \in \mathcal{C}_g$, we evaluate its length-normalized log-probability $\ell_\theta(a \mid s_{i,t})$ via a standard policy forward pass, computing relative support $P_{i,t}$ through softmax normalization over $\mathcal{C}_g$.

**Computational Complexity Analysis.** In standard GRPO, the primary computational cost is dominated by the autoregressive rollout generation phase, which requires $\mathcal{O}(G \cdot L_{\text{gen}})$ sequential decoding steps for a group of $G$ rollouts with average sequence length $L_{\text{gen}}$, alongside parallel forward-backward passes during policy optimization. In contrast, STAC introduces no autoregressive

generation loops, continuation branching, or auxiliary scoring models. For each probed prefix state $s_{i,t}$, the candidate answers $a \in \mathcal{C}_g$ are concise target expressions (e.g., option letters or brief clinical terms) with sequence length $L_{\text{ans}} \ll L_{\text{gen}}$ (typically 1 to 5 tokens). Evaluating the length-normalized log-probability $\ell_\theta(a \mid s_{i,t})$ requires only a single additional forward pass to compute the log-likelihood of these subsequent candidate tokens under the current policy, requiring neither gradient tracking nor multi-step autoregressive decoding. As candidate scoring across $|\mathcal{C}_g| \leq G + 1$ concise targets can be evaluated in a single parallel batch of $\mathcal{O}(|\mathcal{C}_g| \cdot L_{\text{ans}})$ token logits, this additional forward pass incurs negligible overhead relative to the thousands of sequential decoding steps performed during rollout generation, while completely bypassing the training, storage, and inference footprint of separate process reward models.

### C.3 Algorithm of STAC

Algorithm 1 summarizes the complete training workflow of STAC within a single optimization step. By coupling parallel prefix support probing with outcome-conditioned reward modulation, STAC refines trajectory credit assignment directly from the policy's internal dynamics, updating parameters through standard group relative policy optimization without auxiliary evaluators or continuation branching.

**Algorithm 1** STAC Policy Optimization

**Require:** Clinical query $(I, Q, y)$, active policy $\pi_\theta$, group size $G$, probing window $[\xi_{\min}, \xi_{\max}]$
**Ensure:** Updated policy parameters $\theta$

Sample $G$ reasoning rollouts $\{o_i = (z_i, a_i)\}_{i=1}^{G} \sim \pi_\theta(\cdot \mid I, Q)$
Sample relative depth $\xi \sim \mathcal{U}(\xi_{\min}, \xi_{\max})$ and truncate each $z_i$ to prefix state $s_{i,t}$
Construct candidate answer set $\mathcal{C}_g \leftarrow \text{Unique}(\{a_j\}_{j=1}^{G} \cup \{y\})$
**for** $i = 1$ to $G$ **do**
    Score candidates $\{\ell_\theta(a \mid s_{i,t})\}_{a \in \mathcal{C}_g}$ via parallel teacher forcing
    Compute prefix answer support $P_{i,t} \leftarrow \frac{\exp(\ell_\theta(a_i|s_{i,t}))}{\sum_{a' \in \mathcal{C}_g} \exp(\ell_\theta(a'|s_{i,t}))}$
**end for**
Compute rescaled support $\widetilde{P}_{i,t} \leftarrow \frac{P_{i,t} - P_{\min}}{P_{\max} - P_{\min}}$ if $P_{\max} - P_{\min} > \epsilon$, else $0.5$
**for** $i = 1$ to $G$ **do**
    $c_i \leftarrow \mathbf{1}[R_i^{\text{acc}} > 0]$ ▷ Terminal correctness indicator
    $R_i^{\text{proc}} \leftarrow c_i \widetilde{P}_{i,t} + (1 - c_i)(1 - \widetilde{P}_{i,t})$ ▷ Outcome-conditioned process reward
    $R_i^{\text{STAC}} \leftarrow R_i^{\text{fmt}} + R_i^{\text{acc}} + R_i^{\text{proc}}$ ▷ Overall composite reward
**end for**
Compute normalized advantages $\{\hat{A}_i\}_{i=1}^{G}$ and update $\pi_\theta$ via GRPO (Eq. 1)

## D Details of the Experimental Setup

### D.1 Training Datasets

The training recipe of STAC adopts a two-stage sequential pipeline: Supervised Fine-Tuning (SFT) warm-up followed by Reinforcement Learning with Verifiable Rewards (RLVR), with the quantitative data composition summarized in Table 8. The primary goal of the SFT warm-up is format alignment and cold-start: base instruct models do not naturally isolate their reasoning from conclusions, so SFT guides the model to consistently place its final diagnosis within `<answer>...</answer>` tags preceded by step-by-step rationales, enabling deterministic verifiers to parse and evaluate outcomes reliably during RLVR. Following warm-up, the policy is optimized via STAC on verifiable clinical tasks using training samples that are strictly disjoint from the SFT split to prevent trajectory memorization and empirical reward hacking.

**Detailed Descriptions of Training Datasets.** A detailed description of each dataset utilized in our training pipeline is provided as follows:

- **PMC-VQA** (Zhang et al., 2023): PMC-VQA is a large-scale medical visual question answering dataset designed to facilitate research on multimodal understanding in the medical domain. It

Table 8: **Data composition across SFT warm-up and RLVR optimization stages.** All RLVR samples are strictly disjoint from the SFT training data, and all sources respect official train/test splits.

| Dataset | Modality / Domain | Question Format | SFT Samples | RLVR Samples |
|---|---|---|---|---|
| PMC-VQA (Zhang et al., 2023) | Biomedical Figures | Multiple-choice | 50,000† | 4,000 |
| PathVQA (He et al., 2020) | Pathology | Free-form / Yes-No | 15,000 | 3,000 |
| SLAKE (Liu et al., 2021) | Radiology (CT, MRI, X-ray) | Free-form / Closed | 5,000 | 2,000 |
| VQA-RAD (Lau et al., 2018) | Radiology | Free-form / Yes-No | 500 | 500 |
| GMAI-Reasoning10K (Su et al., 2025) | Multi-Specialty (12 Modalities) | Multiple-choice | 4,000 | 5,000 |
| MedQA-USMLE (Jin et al., 2020) | Textual Clinical Knowledge | Multiple-choice | – | 5,000 |
| **Total** | | | **74,500** | **19,500** |

†Chain-of-thought reasoning responses synthesized using Qwen3-VL-32B-Instruct (Bai et al., 2025a).

comprises 227K VQA pairs grounded in 149K medical images collected from PubMed Central Open Access papers, covering a wide range of imaging modalities, anatomical regions, and disease types.

- **PathVQA** (He et al., 2020): PathVQA is a large-scale dataset developed for medical visual question answering tasks in the domain of pathology. It comprises 32,799 expert-annotated question-answer pairs spanning seven question categories, grounded in 4,998 high-resolution pathology images curated from medical textbooks and the digital PEIR library, spanning both open-ended and binary diagnostic inquiries.
- **SLAKE** (Liu et al., 2021): SLAKE is a bilingual, semantically labeled, and knowledge-enhanced medical visual question answering dataset. It comprises 642 annotated radiological images spanning 39 anatomical structures and 12 disease categories, including conditions such as various cancers (e.g., brain, liver, kidney, lung) and thoracic diseases (e.g., atelectasis, pleural effusion, pulmonary masses, and pneumothorax).
- **VQA-RAD** (Lau et al., 2018): VQA-RAD is a medical visual question answering dataset constructed for assessing multimodal understanding of radiology. It consists of 315 radiological images paired with 3,515 manually curated question-answer pairs authored by clinical experts across head, chest, and abdominal imaging. The dataset includes both open-ended and binary (yes/no) questions.
- **GMAI-Reasoning10K** (Su et al., 2025): GMAI-Reasoning is a high-quality medical visual reasoning dataset comprising 10K curated multiple-choice questions constructed from 95 publicly available medical datasets spanning 12 imaging modalities (e.g., X-ray, CT, MRI, ultrasound, dermoscopy). Each question is paired with standardized visual inputs and clinical metadata, generated using GPT-based prompting following rigorous preprocessing and quality control procedures.
- **MedQA-USMLE** (Jin et al., 2020): MedQA-USMLE is a medical textual question answering benchmark compiled from the professional United States Medical Licensing Examination. It contains 12,723 four-option multiple-choice questions (10,178 in the official training split) demanding multi-step clinical reasoning across pathology, pharmacology, biochemistry, internal medicine, and patient management, providing robust clinical knowledge supervision that complements image-grounded VQA.

### D.2 Hyperparameter Configuration and Framework for Training

**SFT Warm-up Setup.** For the supervised warm-up stage, model training is conducted using the `LLaMA-Factory` framework (Zheng et al., 2024). The base model is fine-tuned for 2 epochs with a global batch size of 64 and a learning rate of $2.5 \times 10^{-6}$ scheduled by cosine decay. The maximum sequence length is set to 8,192 tokens, using the AdamW optimizer with a 10% linear warmup.

**RLVR Training Setup.** Subsequent RLVR policy optimization is conducted via `verl` (Sheng et al., 2024); Table 9 details the complete hyperparameters, execution backend, and hardware specifications.

### D.3 Evaluation Protocol and Benchmark Details

To ensure standardized and reproducible benchmarking aligned with prior work, all evaluations are conducted via MedEvalKit (Xu et al., 2025). We utilize its multimodal evaluation suite across diverse inquiry formats, coupling deterministic rule-based parsing for multiple-choice and closed-ended answers with an LLM-as-a-judge protocol for open-ended clinical rationales, where we utilize Qwen2.5-72B-Instruct as the judge model.

Table 9: **Hyperparameter configuration for RLVR training.** Configurations remain identical across baseline GRPO, CARE, and STAC runs.

| Hyperparameter | Value |
|---|---|
| Rollout group size ($G$) | 8 |
| Global batch size (trajectories) | 128 |
| Actor learning rate | $1 \times 10^{-6}$ |
| Learning rate schedule | Cosine with 10% linear warmup |
| KL regularization coefficient ($\beta$) | 0.01 |
| PPO clipping parameter ($\epsilon_{\text{clip}}$) | 0.2 |
| Advantage stability constant ($\epsilon_A$) | $1 \times 10^{-6}$ |
| Rollout temperature | 1.0 |
| Rollout top-$p$ | 1.0 |
| Maximum prompt length | 8,192 tokens |
| Maximum response length | 4,096 tokens |
| Prefix probing interval ($\xi$) | $[0.15, 0.35]$ |
| Group rescaling threshold ($\epsilon$) | $1 \times 10^{-6}$ |
| Training epochs | 3 |
| Distributed backend | verl + vLLM + FSDP2 |
| Hardware | 8 $\times$ NVIDIA H800 GPUs (80GB) |

**Detailed Descriptions of Evaluation Benchmarks.** To rigorously assess diagnostic reasoning across diverse clinical disciplines and question formulations, we benchmark **STAC** against six recognized medical vision-language suites:

- **PMC-VQA** (Zhang et al., 2023): Constructed from biomedical literature across PubMed Central, PMC-VQA provides extensive coverage of real-world clinical documentation spanning $149{,}000$ peer-reviewed medical figures and $227{,}000$ multiple-choice questions. In our evaluation, we utilize its cleaned and expert-reviewed subset, which systematically filters out automated extraction artifacts and ambiguous candidate options to guarantee ground-truth correctness and diagnostic reliability across diverse clinical pathologies.
- **MedXpertQA** (Zuo et al., 2025): Representing expert-level clinical decision-making, MedXpertQA benchmarks advanced diagnostic reasoning aligned with board certification examinations. The complete benchmark spans $4{,}460$ questions across 17 medical specialties and 11 organ systems; our experiments focus specifically on its multimodal partition, requiring intricate multi-step deductions from clinical case vignettes and matched radiological or pathological findings.
- **OmniMedVQA** (Hu et al., 2024): To test panoramic generalization across fragmented medical domains, OmniMedVQA integrates diverse clinical data from 73 independent biomedical repositories into a unified multiple-choice benchmark. It encompasses $118{,}010$ medical images spanning 12 distinct imaging modalities and over 20 anatomical regions, comprising $127{,}995$ question-answer pairs designed to expose out-of-distribution vulnerabilities in specialized clinical scenarios.
- **VQA-RAD** (Lau et al., 2018): Serving as a clinician-validated benchmark for radiology, VQA-RAD assesses whether multimodal models can reliably interpret authentic clinical scans across head, chest, and abdominal regions. Curated entirely by practicing clinicians to reflect genuine diagnostic inquiries, it links 315 radiological images with $2{,}248$ question-answer pairs, combining constrained yes/no inquiries with descriptive open-ended queries.
- **SLAKE** (Liu et al., 2021): SLAKE evaluates vision-language models under semantically grounded medical contexts, enriched with structured clinical knowledge bases and bilingual (English-Chinese) annotations. Grounded in $642$ radiological studies (CT, MRI, and chest radiographs) paired with $14{,}028$ multimodal instances, the benchmark probes both anatomical localization across 39 organ structures and disease characterization across 12 major pathological categories.
- **PathVQA** (He et al., 2020): Targeting microscopic histological comprehension, PathVQA benchmarks diagnostic deduction in pathology. Sourced from standard reference pathology textbooks and the digital PEIR archive, it compiles $4{,}998$ high-resolution microscopic images and $32{,}799$ corresponding question-answer pairs, challenging models to identify subtle cellular morphology, tissue alterations, and histological hallmarks through mixed closed- and open-ended queries.

Table 10: **Detailed numerical results for reasoning stability and continuation reliability (RQ3).** Model rollouts are probed across early (25%), middle (50%), and late (75%) reasoning prefixes, resampling 8 independent continuations per prefix to evaluate continuation persistence on correct ($S_c$) and incorrect ($S_w$) trajectories. STAC systematically reinforces robust correct derivations while mitigating stubborn wrong deductions.

| Metric | MedXpertQA-MM | | | PMC-VQA | | |
|---|---|---|---|---|---|---|
| | GRPO | STAC | Δ | GRPO | STAC | Δ |
| Pass@8 (↑) | 0.758 | **0.802** | +4.4pp | 0.880 | **0.938** | +5.8pp |
| 25% Prefix $S_c$ (↑) | 0.556 | **0.576** | **+2.1pp** | 0.744 | **0.751** | **+0.6pp** |
| 25% Prefix $S_w$ (↓) | 0.537 | **0.533** | **–0.4pp** | 0.602 | **0.585** | **–1.7pp** |
| 50% Prefix $S_c$ (↑) | 0.756 | **0.792** | **+3.5pp** | 0.881 | **0.897** | **+1.6pp** |
| 50% Prefix $S_w$ (↓) | 0.728 | **0.716** | **–1.2pp** | 0.785 | **0.776** | **–1.0pp** |
| 75% Prefix $S_c$ (↑) | 0.884 | **0.887** | **+0.3pp** | 0.933 | **0.956** | **+2.3pp** |
| 75% Prefix $S_w$ (↓) | 0.851 | **0.832** | **–1.8pp** | 0.894 | **0.883** | **–1.1pp** |

## E Supplementary Details of Reasoning Stability (RQ3)

To supplement the visual comparison presented in Figure 3 of the main text, this section provides the detailed experimental protocol for continuation probing alongside the exact numerical results.

**Continuation Probing Protocol and Sample Scale.** To rigorously capture the reasoning dynamics of policies trained by GRPO and STAC, continuation probing is conducted on held-out test sets from both complex multi-step reasoning (MedXpertQA-MM) and broad biomedical multimodal QA (PMC-VQA). For each benchmark, we randomly sample $N = 400$ test queries. For every query, the evaluated checkpoint generates $G = 8$ complete rollout trajectories under standard decoding ($T = 1.0$, top-$p = 1.0$), establishing a base evaluation cohort of 3,200 rollouts per benchmark. Across each rollout, we truncate the intermediate reasoning trace at early (25%), intermediate (50%), and late (75%) prefix lengths, measured proportionally against the total number of reasoning tokens generated prior to the final answer tag. Branching from each truncated prefix, the policy independently resamples $K = 8$ continuation trajectories under identical decoding temperatures, matching the generative exploration distribution of the policy.

**Statistical Computation of Evaluation Metrics.** Based on this multi-branch sampling cohort, the three primary indicators in Table 10 are statistically aggregated as follows:

- **Exploration Ceiling (Pass@8):** Computed directly over the $G = 8$ initial rollouts across all $N = 400$ test queries, measuring the proportion of queries where at least one rollout discovers the verified ground-truth target. This characterizes the empirical search frontier of the policy prior to prefix truncation.
- **Correct Continuation Stability ($S_c$):** Conditioned strictly on initial rollouts that were originally correct ($y_{\text{orig}} = y^*$). For each successful rollout at prefix depth $\alpha \in \{25\%, 50\%, 75\%\}$, we compute the empirical probability that its $K = 8$ resampled continuations retain the correct target answer, averaged over all successful original rollouts.
- **Incorrect Continuation Stability ($S_w$):** Conditioned strictly on initial rollouts that were originally incorrect ($y_{\text{orig}} \neq y^*$). For each failed rollout at prefix depth $\alpha$, we compute the empirical probability that its $K = 8$ resampled continuations stubbornly preserve the identical erroneous conclusion ($y_{\text{cont}} = y_{\text{orig}}$), averaged over all failed original rollouts. A lower $S_w$ reflects diagnostic flexibility and reduced entrenchment in false premises.

Table 10 compiles the comprehensive numerical results across all probed positions on MedXpertQA-MM and PMC-VQA. STAC systematically bolsters correct derivation robustness while suppressing stubborn incorrect persistence across all evaluated stages.