VERGE: VERIFIED EARLY-EXIT REASONING WITH GROUNDED EVIDENCE IN MEDICAL MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Long medical reasoning traces can contain a correct intermediate prediction followed by redundant or harmful continuation. We introduce **VERGE** (Verified Early-exit Reasoning with Grounded Evidence), a GRPO-based framework that uses intermediate verification to supervise credit assignment and reasoning termination. A reasoning continuation probe forces candidate prefixes to transition to answering and checks the resulting predictions against task-specific verification criteria. Accepted boundaries guide advantage redistribution over the original rollouts and supervise the reasoning-termination token; probe completions are used only for verification. VERGE handles discrete answer correctness and continuous localization quality, preserving each grounding rollout's mean GRPO advantage. Starting from Qwen3-VL-4B-Thinking, VERGE improves average accuracy across four medical QA benchmarks from 54.17% to 57.54%. Under a matched 4,096-token protocol, it outperforms Vanilla GRPO by 4.95 percentage points in MedMCQA accuracy and 5.61 points in grounding mIoU on the evaluated subset, while reducing mean output tokens by 77.2% and 69.6%, respectively. These results support learning to terminate reasoning from verified intermediate predictions.

1 INTRODUCTION

Medical multimodal large language models (MLLMs) have advanced question answering (QA) and evidence-grounded prediction (Li et al., 2023; Moor et al., 2023; Tu et al., 2024). Recent models such as MedVLM-R1, MedMO, and MedReasoner use reinforcement learning (RL) with verifiable rewards to advance medical reasoning and localization (Pan et al., 2025; Deria et al., 2026; Yan et al., 2026).

Yet longer reasoning does not necessarily yield better predictions (Chen et al., 2024b; Ballon et al., 2026). Repeated coordinate checking can coexist with inaccurate localization, as illustrated in Figure 1. Recent work addresses reasoning efficiency through dynamic truncation and verified rollout editing (Chen et al., 2026; Wei et al., 2026), motivating policies that learn not only *how to reason*, but also *when to stop*.

Reliable early exit in medical QA and grounding presents three challenges. **(1)** An intermediate answer or box may be provisional or inaccurate, requiring verification of its quality and consistency when transitioning to answering. **(2)** Outcome-supervised Group Relative Policy Optimization (GRPO) broadcasts one advantage across response tokens (Shao et al., 2024), potentially rewarding redundant continuation or penalizing useful prefixes according to the final outcome. **(3)** QA provides discrete correctness, whereas grounding provides continuous intersection-over-union (IoU) feedback; a uniform binary credit rule would discard relative localization quality.

To address these challenges, we introduce **VERGE**, *Verified Early-exit Reasoning with Grounded Evidence*, a GRPO-based framework for efficient medical multimodal reasoning (Figure 2). A *Reasoning Continuation Probe* preserves sampled prefixes and forces a transition to answering, selecting the earliest boundary that passes task-specific verification. Accepted boundaries guide *task-aware credit redistribution* (CR) across the original prefix, continuation, and final answer, and supervise the reasoning-termination token for *early-exit learning*. For grounding, redistribution preserves

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.

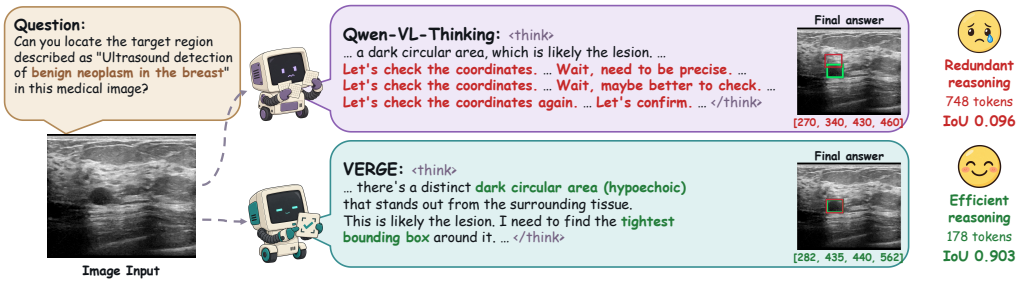

Figure 1: Medical grounding example. Repeated coordinate checks yield an inaccurate box, whereas VERGE uses a shorter trace and achieves higher IoU. Green: ground truth; red: predictions.

continuous quality information and each rollout's mean GRPO advantage. Probe completions are verification-only: they neither enter the rollout group nor become answer-generation targets. Without a verified boundary, standard GRPO is retained without exit supervision.

Starting from Qwen3-VL-4B-Thinking (Bai et al., 2025), VERGE raises average accuracy across four QA benchmarks from 54.17% to 57.54% and improves grounding across all eight evaluated modality groups. Under a matched 4,096-token protocol, it improves over Vanilla GRPO by 4.95 percentage points on MedMCQA and 5.61 mIoU points on the grounding subset, while reducing mean output tokens by 77.2% and 69.6%, respectively.

Our main contributions are summarized as follows:

- We formulate verified intermediate predictions as shared supervision for reasoning termination and within-trajectory credit assignment in medical QA and grounding.
- We develop VERGE, combining verification-only probing, task-aware credit redistribution over original rollouts, and explicit early-exit learning.
- We evaluate prediction quality and output length, with matched component ablations quantifying the incremental benefits of early-exit learning and credit redistribution.

2 Related Work

2.1 Medical MLLMs

Medical MLLMs develop biomedical vision–language capabilities through domain-specific training (Li et al., 2023; Moor et al., 2023; Chen et al., 2024a), supported by increasingly diverse VQA resources (Zhang et al., 2023; Hu et al., 2024). Generalist models further broaden medical task and modality coverage (Tu et al., 2024; Shu et al., 2025; Xu et al., 2026), while recent RL-based approaches advance medical reasoning and spatial grounding through outcome supervision, segmentation experts, and adaptive rewards (Pan et al., 2025; Deria et al., 2026; Yan et al., 2026; Yang et al., 2026). VERGE complements these advances by targeting reasoning efficiency and within-trajectory credit assignment across medical QA and grounding.

2.2 Efficient Reasoning and Early Exit

Studies of overthinking, trajectory probing, and post-conclusion continuation show that longer reasoning is not uniformly useful (Chen et al., 2024b; Ballon et al., 2026; He et al., 2026). Existing approaches explore length-constrained RL (Aggarwal & Welleck, 2025; Hou et al., 2025), adaptive perception–reasoning allocation (Diao et al., 2026), and dynamic early-exit training (Chen et al., 2026). In multimodal models, claims of visual re-examination need not reflect renewed sensitivity to visual evidence (Shi et al., 2026), motivating verification of intermediate predictions. Closely related Dynamic Rollout Editing (DRE) augments RL groups with verified shorter trajectories, favors accepted edits, and masks shared prefixes on original trajectories (Wei et al., 2026). In contrast, VERGE uses probe completions only for verification, preserving the original rollout group and

outcome-reward statistics. Verified boundaries guide task-aware credit redistribution on original trajectories and explicit termination supervision, with distinct treatment of discrete QA correctness and continuous grounding rewards.

3 DATA

3.1 DATA COLLECTION

We collect text QA examples from MedMCQA Pal et al. (2022), visual QA examples from PMC-VQA Zhang et al. (2023), and grounding examples from BiomedParse-Data Zhao et al. (2025) and U-MRG-14K Yan et al. (2026). Table 1 reports the training and validation set sizes. Appendix A provides the source composition and modality coverage.

Table 1: Statistics of the training and validation data.

Task	Train	Validation
QA	6k	600
Grounding	6k	300
Total	12k	900

3.2 DATA PROCESSING

QA Filtering. For each MedMCQA or PMC-VQA training candidate, Qwen3-VL-4B-Thinking(hereafter Qwen3-VL-4B)(Bai et al., 2025) samples eight responses with temperature 1.0, top-$p = 1.0$, top-$k = -1$, and a 6,144-token limit. A response is correct if its parsed option matches the reference answer. We retain a candidate only when its response group contains both correct and incorrect answers, removing groups with zero correctness variance and preserving an informative group-relative training signal.

Grounding Filtering. We remove empty masks and reject fragmented targets before converting masks into bounding boxes. For a nonempty mask, let $\{C_i\}_{i=1}^{m}$ denote the connected components sorted by decreasing area. Let B_{all} enclose the complete foreground and B_1 enclose the largest component. A mask is discarded if

$$\frac{\sum_{i=2}^{m} |C_i| \, \mathbf{1}[|C_i| \geq 16]}{\sum_{i=1}^{m} |C_i|} \geq 0.005 \quad \text{or} \quad \frac{|B_{\text{all}}|}{|B_1|} > 1.25. \tag{1}$$

The two criteria remove masks with substantial secondary components or spatially distant fragments that inflate the resulting box.

Grounding Standardization. MedMCQA and PMC-VQA retain their original questions and answer options. For BiomedParseData, we convert each segmentation target phrase into a standardized grounding prompt:

Can you locate the target region described as "{phrase}" in this medical image?

U-MRG-14K retains its original open-ended questions; only the questions and mask-derived boxes are used as supervision, excluding rationales and point annotations. For each retained mask, we extract the tight foreground box $B = (x_1, y_1, x_2, y_2)$ and normalize its coordinates to $[0, 1000]$:

$$\widetilde{B} = \text{round}\Big(1000 \left[\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}\right]\Big), \tag{2}$$

where W and H are the image width and height.

4 METHOD

VERGE uses verified intermediate predictions to learn both how to assign credit within a reasoning trajectory and when to terminate reasoning. As illustrated in Figure 2, we first sample a group of original responses for each input. We then identify candidate prediction boundaries and verify their correctness and consistency under a forced transition to answering (Section 4.1). The selected boundary guides task-aware credit redistribution over the original prefix, reasoning continuation, and final answer (Section 4.2). An early-exit objective additionally teaches the policy to generate the reasoning-termination token after verified prefixes (Section 4.3). Finally, we combine the resulting token-level policy objective with exit supervision and KL regularization (Section 4.4).

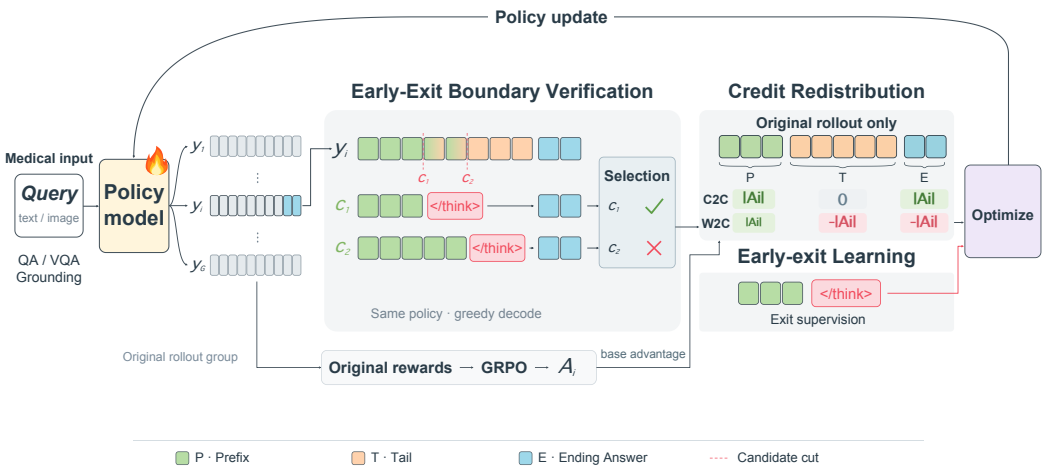

Figure 2: Overview of VERGE. Verification-only probes identify early-exit boundaries that guide credit redistribution over the original prefix (P), the tail (T), and original ending answer region (E), and supervise the reasoning-termination token.

4.1 Early-Exit Boundary Verification

For each input x, the rollout policy $\pi_{\theta_{\text{old}}}$ samples G responses $\{y_i\}_{i=1}^{G}$. We identify intermediate prefixes whose predictions remain correct under a forced transition to answering.

Candidate Boundaries. We extract candidate positions $\mathcal{C}_i$ at which a rollout explicitly commits to a task prediction, such as an answer option or a complete bounding box. Candidates are indexed directly in the original token sequence, with $P_i^{(c)}$ denoting the reasoning prefix ending at $c \in \mathcal{C}_i$.

Reasoning Continuation Probe. The Reasoning Continuation Probe (RCP) preserves each candidate prefix and forces a transition to the answer channel:

$$\widetilde{y}_i^{(c)} = \text{Decode}\Big(\pi_{\theta_{\text{old}}} \,\Big|\, x \oplus P_i^{(c)} \oplus \texttt{</think>}\Big). \tag{3}$$

We evaluate each candidate and the original full prediction using task-quality scores $q_i^{(c)}$ and q_i^{full}, respectively. These scores measure exact option correctness for QA and continuous IoU for grounding. Verification requires both the prefix and probe predictions to be correct, using IoU thresholds for grounding, and to satisfy the applicable consistency, confidence, and relative-quality checks (Appendix C). Candidates that pass all checks form $\mathcal{C}_i^{\text{ver}} \subseteq \mathcal{C}_i$. Probe completions provide verification evidence only and are excluded from the GRPO rollout group and policy surrogate.

Boundary Selection. When $\mathcal{C}_i^{\text{ver}}$ is nonempty, we select the earliest boundary, yielding the shortest verified reasoning prefix within the original rollout:

$$c_i^\star = \min \mathcal{C}_i^{\text{ver}}, \qquad q_i^{\text{cut}} := q_i^{(c_i^\star)}. \tag{4}$$

If no candidate is verified, the rollout retains its original GRPO advantages and receives no early-exit target.

4.2 Token-Level Credit Redistribution

A verified early-exit boundary distinguishes a prediction-supporting prefix from its subsequent continuation.We use task-aware credit redistribution (CR) to refine the sequence-level credit assigned by GRPO. Given the outcome rewards $\{\mathcal{R}_i\}_{i=1}^{G}$ of the original responses sampled for the same input, the base advantage is

$$A_i = \frac{\mathcal{R}_i - \mu_{\mathcal{R}}}{\sigma_{\mathcal{R}} + \epsilon}, \tag{5}$$

where $\mu_{\mathcal{R}}$ and $\sigma_{\mathcal{R}}$ are the within-group reward mean and standard deviation, and ϵ is a numerical stabilizer. Standard GRPO broadcasts A_i to every valid response token. VERGE instead assigns token-level advantages according to the verified boundary and the outcome of the original response.

Trajectory regions and outcomes. For a selected boundary $c_i^\star$, we decompose the original rollout as

$$y_i = P_i^{(c_i^\star)} \oplus T_i^{(c_i^\star)} \oplus E_i, \tag{6}$$

where $P_i^{(c_i^\star)}$ is the verified prefix, $T_i^{(c_i^\star)}$ is the subsequent reasoning continuation, and E_i contains the original reasoning-termination token and final answer. Separating the continuation from the answer allows us to retain supervision for a correct final answer without rewarding every preceding reasoning token. All three regions belong to the original rollout, not the probe completion.

We distinguish two cases using *full-to-cut* correctness: **C2C** denotes a correct full response with a verified correct prediction at the cut; **W2C** denotes an incorrect full response despite a verified correct prediction at the cut. Rollouts without a verified boundary retain $A_{i,t}^{\mathrm{VERGE}} = A_i$ at every valid response token.

Discrete-answer tasks. For QA, we set the region-specific advantages using the absolute trajectory-level advantage $|A_i|$ (Table 2).

For C2C trajectories, we assign $|A_i|$ to the verified prefix and the correct final answer, and zero advantage to the intervening continuation. This removes direct outcome-based reinforcement of that continuation without penalizing the correct answer. For W2C trajectories, the verified prefix receives $|A_i|$, whereas the continuation and original incorrect answer receive $-|A_i|$. Thus, credit follows the verified intermediate prediction rather than being determined uniformly by the final outcome.

Table 2: QA token-region advantages. P, T, and E denote the prefix, continuation, and original answer region, respectively; E includes the original reasoning-termination token.

Case	P	T	E
C2C	$\lvert A_i\rvert$	0	$\lvert A_i\rvert$
W2C	$\lvert A_i\rvert$	$-\lvert A_i\rvert$	$-\lvert A_i\rvert$
No verified cut	A_i	A_i	A_i

Continuous grounding rewards. Grounding rewards are continuous IoU scores that encode relative localization quality within a rollout group. Replacing the base advantage with only positive or negative region labels would discard this information. We therefore retain A_i and add a centered token-level contrast:

$$A_{i,t}^{\mathrm{VERGE}} = A_i + \delta_{i,t}. \tag{7}$$

Let $z_{i,t}$ be a bounded, region-dependent contrast before centering that favors the verified prefix over redundant or harmful continuation. For C2C trajectories, its magnitude depends on the quality of the verified localization. For W2C trajectories, it additionally reflects the improvement $q_i^{\mathrm{cut}} - q_i^{\mathrm{full}}$, evaluated at the selected boundary. Let $n_i = |y_i|$ denote the number of tokens in the original response, including reasoning and the final answer but excluding input and padding tokens. This response-length notation applies to both QA and grounding. We center the contrast over the n_i response tokens:

$$\delta_{i,t} = z_{i,t} - \frac{1}{n_i}\sum_{s=1}^{n_i} z_{i,s}, \qquad t = 1, \ldots, n_i. \tag{8}$$

Subtracting the within-response mean gives $\sum_{t=1}^{n_i} \delta_{i,t} = 0$. Consequently,

$$\frac{1}{n_i}\sum_{t=1}^{n_i} A_{i,t}^{\mathrm{VERGE}} = A_i. \tag{9}$$

This construction redistributes credit within each grounding trajectory while preserving its mean GRPO advantage. The contrast changes the relative credit of trajectory regions; it does not necessarily make the resulting advantage positive for every prefix token.

4.3 EARLY-EXIT LEARNING

Credit redistribution determines how to train the sampled trajectory, but it does not explicitly specify the action to take at a verified boundary. For example, a zero continuation advantage in C2C trajectories removes direct outcome-based reinforcement without providing a target for termination. We therefore introduce an auxiliary objective that supervises `</think>` after verified reasoning prefixes.

For task family $k \in \{\mathrm{QA}, \mathrm{G}\}$, let $\mathcal{V}_k$ denote the collection of verified input–prefix pairs (x, P) used for exit supervision. The exit loss is

$$\mathcal{L}_{\text{exit}}^{k} = -\frac{1}{|\mathcal{V}_k|} \sum_{(x,P)\in\mathcal{V}_k} \log \pi_\theta(\texttt{</think>} \mid x, P)\,. \tag{7}$$

QA and grounding losses are normalized separately so that their relative weighting is not determined solely by the number of accepted prefixes. The supervision target is the termination action, not the probe-generated answer. Credit redistribution and early-exit learning therefore provide complementary signals: the former differentiates regions of the original trajectory, while the latter teaches the policy to stop reasoning at verified prefixes.

4.4 POLICY OPTIMIZATION

We combine the task-aware token advantages with early-exit supervision to optimize the policy. For a rollout group, the clipped policy loss (Schulman et al., 2017) is

$$\mathcal{L}_{\text{policy}} = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{n_i}\sum_{t=1}^{n_i}\min\!\big(r_{i,t}A_{i,t}^{\text{VERGE}}, \text{clip}(r_{i,t}, 1-\varepsilon, 1+\varepsilon)A_{i,t}^{\text{VERGE}}\big)\,, \tag{8}$$

where $r_{i,t} = r_{i,t}(\theta)$ and ε is the clipping parameter. The inner average is over the n_i original response tokens, and the outer average is over the G sampled responses. The same policy loss is used for QA and grounding; their task-specific credit redistribution is reflected in $A_{i,t}^{\text{VERGE}}$. Without a verified boundary, $A_{i,t}^{\text{VERGE}} = A_i$, and the trajectory retains the standard GRPO surrogate. The complete training objective is

$$\mathcal{L}_{\text{VERGE}} = \mathcal{L}_{\text{policy}} + \beta\mathcal{L}_{\text{KL}} + \lambda_{\text{QA}}\mathcal{L}_{\text{exit}}^{\text{QA}} + \lambda_{\text{G}}\mathcal{L}_{\text{exit}}^{\text{G}}, \tag{9}$$

where $\mathcal{L}_{\text{KL}}$ regularizes the updated policy against the reference policy, and β, λ_{QA}, and λ_{G} control the corresponding regularization and supervision strengths.

Each training iteration samples and scores original rollouts, verifies candidate boundaries, and constructs the associated token advantages and exit targets before updating the policy. The probe affects training through verification-derived supervision, while outcome rewards, group statistics, and importance ratios remain tied to the original sampled trajectories.

5 EXPERIMENTS

VERGE was trained on the data summarized in Table 1 using **four NVIDIA GeForce RTX 4090 GPUs** (**48 GB each**) for approximately **71 hours** of active runtime, including validation, following the training method illustrated in Figure 2. We implemented training in EasyR1 Yaowei Zheng (2025), initializing the policy from Qwen3-VL-4B (Base) and training for two epochs. Each rollout batch contains 64 prompts with eight sampled responses per prompt, using temperature 1.0, top-p 1.0, and maximum prompt and response lengths of 2,048 and 6,144 tokens, respectively. We use AdamW(Loshchilov & Hutter, 2019) with a learning rate of 10^{-6}, weight decay of 10^{-2}, and an actor update batch size of 32. Both the vision encoder and language model are updated in bfloat16 precision with fully sharded data parallelism. The KL regularization coefficient is 10^{-3}, and the QA and grounding exit-loss coefficients are each 2.5×10^{-3}. Validation is performed every ten steps with temperature 0.6 and top-p 0.95.

Table 3: Comparison on medical multimodal and text benchmarks. Bold and underlined values indicate the best and second-best results among open-source models and our evaluations. Δ denotes absolute score-point gains over Qwen3VL-4B, not relative percentages.

Models	MMMU-Med	MMLU-Med	MedMCQA	SGPQA	Avg.
GPT-4.1	75.2	89.6	77.7	49.9	73.1
Claude Sonnet 4	74.6	91.3	79.3	56.3	75.4
Gemini-2.5-Flash	76.9	84.2	73.6	53.3	72.0
BiomedGPT	24.9	—	—	—	—
Med-R1-2B	34.8	51.5	39.1	17.9	35.8
MedVLM-R1-2B	35.2	51.8	39.7	19.1	36.5
MedGemma-4B-IT	43.7	66.7	52.2	21.6	46.1
LLaVA-Med-7B	29.3	50.6	39.4	16.1	33.9
HuatuoGPT-V-7B	47.3	69.3	51.2	21.9	47.4
BioMediX2-8B	39.8	68.6	52.9	25.2	46.6
Qwen2.5VL-7B	50.6	73.4	52.6	26.3	50.7
InternVL2.5-8B	53.5	74.2	52.4	26.1	51.6
InternVL3-8B(Zhu et al., 2025)	59.2	<u>77.5</u>	57.7	<u>31.2</u>	<u>56.4</u>
Lingshu-7B (Xu et al., 2026)	54.0	69.6	56.3	27.5	51.9
Fleming-VL-8B(Shu et al., 2025)	**63.3**	71.8	51.8	24.9	53.0
MedMO-4B(Deria et al., 2026)	54.6	75.7	58.0	29.4	54.4
MedMO-4B-Next	58.7	74.8	<u>58.1</u>	29.5	55.3
Qwen3VL-4B	55.48	76.54	54.84	29.80	54.17
VERGE (ours)	<u>59.30</u>	**78.83**	**58.98**	**33.03**	**57.54**
Δ	+3.82	+2.29	+4.14	+3.23	+3.37

Table 4: Comparison on medical visual grounding benchmarks across different imaging modalities.

Models	CT	MRI	US	Endo.	Fundus	Dermo.	Photo.	Mixed
MedMO-4B	7.15	6.67	18.08	18.21	4.25	46.60	29.71	6.84
Lingshu-7B	8.91	10.08	29.04	22.85	11.19	49.58	23.67	10.98
Fleming-VL-8B	9.32	12.96	20.79	30.46	22.05	46.36	34.88	15.13
InternVL3-8B	12.92	13.19	30.92	43.53	15.31	63.69	57.12	18.40
Qwen3VL-4B	5.92	5.83	40.33	47.52	27.15	73.04	78.18	12.56
VERGE (ours)	**22.31**	**28.33**	**63.87**	**58.05**	**54.26**	**78.40**	**81.87**	**26.00**
Δ	+16.39	+22.50	+23.54	+10.53	+27.11	+5.36	+3.69	+13.44

5.1 Overall comparison

We evaluate VERGE along two complementary dimensions, medical question answering (QA) and visual grounding, as summarized in Tables 3 and 4. QA evaluation covers MMMU-Med Yue et al. (2024), MMLU-Med Hendrycks et al. (2021), MedMCQA Pal et al. (2022), and the medical subset of SuperGPQA (SGPQA) M-A-P Team et al. (2025), encompassing both visual and text-only questions. Grounding evaluation comprises 44,839 annotations from 14 sources across eight modality groups, including the mixed-modality U-MRG-14K benchmark Yan et al. (2026). We report exact-answer accuracy for QA and bounding-box mean intersection over union (mIoU, $\times 100$) for grounding. Detailed dataset coverage and evaluation protocols are provided in Appendix B.

Comparison on QA VERGE improves over Base on all four benchmarks, raising average accuracy from 54.17% to 57.54% (+3.37 percentage points). It achieves the highest reported average among the open-source models in Table 3, exceeding InternVL3-8B and MedMO-4B-Next by 1.14

Table 5: Component ablation under a matched step-370 checkpoint and a 4,096-token greedy decoding budget. $\mathcal{L}_{\text{exit}}$ and CR denote exit loss and credit redistribution, respectively.

	MedMCQA		Grounding	
Model	Acc. (%) ↑	Tok. ↓	mIoU (%) ↑	Tok. ↓
Qwen3VL-4B (baseline)	51.04	2582.4	20.71	2506.1
+ Vanilla GRPO	55.27	2229.5	36.84	1019.6
+ $\mathcal{L}_{\text{exit}}$	59.81	522.0	41.75	353.6
+ $\mathcal{L}_{\text{exit}}$ + CR	**60.22** +9.18	**507.8** -80.3%	**42.45** +21.74	**310.0** -87.6%

and 2.24 points, respectively. On text-only QA, VERGE reaches 78.83% on MMLU-Med, 58.98% on MedMCQA, and 33.03% on SGPQA, with gains of 2.29, 4.14, and 3.23 points over Base. Visual QA accuracy on MMMU-Med also increases from 55.48% to 59.30%, ranking second among the listed open-source models. Base and VERGE share the same evaluation protocol; other QA scores are reported from MedMO Deria et al. (2026).

Comparison on Grounding VERGE achieves the highest mIoU across all eight modality groups among the models listed in Table 4, consistently outperforming Base. The largest absolute gains are observed in fundus imaging (+27.11 points), ultrasound (+23.54), and MRI (+22.50), where VERGE reaches 54.26, 63.87, and 28.33 mIoU, respectively. Substantial improvements are also observed in CT, mixed-modality imaging, and endoscopy, with gains of 16.39, 13.44, and 10.53 points, respectively. For dermoscopy and photography, where Base already achieves relatively high scores of 73.04 and 78.18, VERGE further improves mIoU to 78.40 and 81.87. These results demonstrate consistent grounding improvements over Base across diverse medical imaging modalities.

5.2 Ablation Study

We assess the incremental contributions of exit loss ($\mathcal{L}_{\text{exit}}$) and credit redistribution (CR) by adding them successively to Vanilla GRPO (Table 5). All trained variants use step-370 checkpoints and greedy decoding with a maximum budget of 4,096 tokens. Evaluation uses 4,183 MedMCQA validation examples and a shared grounding subset of 10,520 examples from 11 datasets. The QA prompt and format checks, as well as the grounding cohort, differ from those used in the main comparison; details are provided in Appendix B.

Effect of Early-exit Learning. Adding $\mathcal{L}_{\text{exit}}$ to Vanilla GRPO produces the larger incremental improvement in both task scores and output length among the two component additions. MedMCQA accuracy increases from 55.27% to 59.81%, while mean output length decreases from 2,229.5 to 522.0 tokens, a 76.6% reduction. Grounding mIoU similarly increases from 36.84 to 41.75, with mean output length decreasing from 1,019.6 to 353.6 tokens, a 65.3% reduction. These results show that supervising `</think>` at verified reasoning prefixes substantially shortens generation while improving prediction quality on both tasks.

Effect of credit redistribution. With exit loss enabled, adding CR further increases MedMCQA accuracy from 59.81% to 60.22% and grounding mIoU from 41.75 to 42.45. Mean output length decreases from 522.0 to 507.8 tokens on MedMCQA (-2.7%) and from 353.6 to 310.0 tokens on grounding (-12.3%). These results support the additional benefit of task-aware credit redistribution when combined with exit supervision, with a larger proportional reduction in output length on grounding. Overall, the combined model achieves the highest task scores and the lowest token counts among the evaluated variants. Relative to Base, it improves MedMCQA accuracy and grounding mIoU by 9.18 and 21.74 points while reducing output length by 80.3% and 87.6%, respectively. This staged ablation demonstrates the incremental benefit of CR with exit loss enabled.

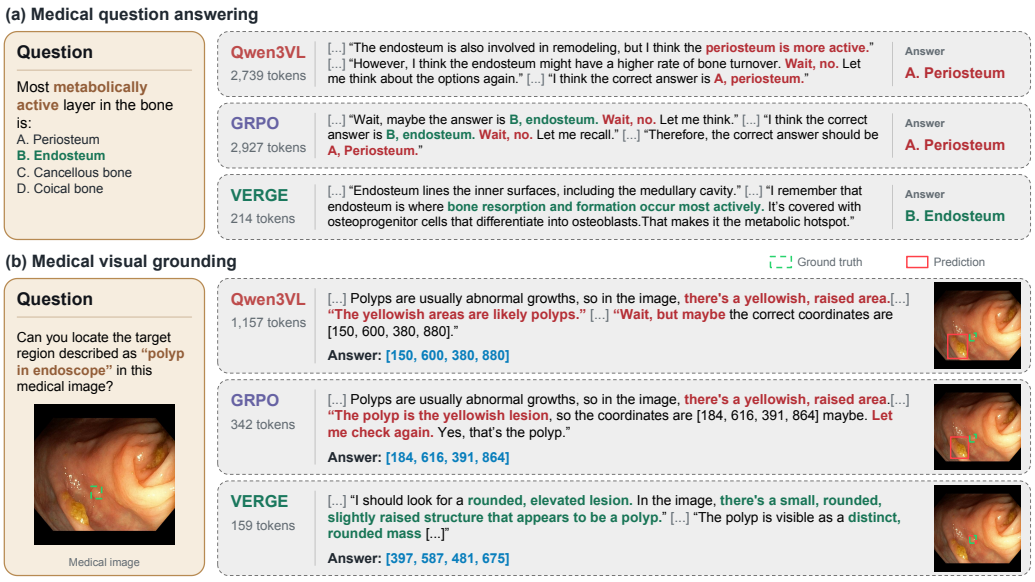

Figure 3: Qualitative comparison between the base model and VERGE on visual grounding. VERGE reduces redundant reasoning while preserving task-relevant visual evidence, producing a more accurate bounding box with substantially fewer output tokens.

5.3 Qualitative Study

Figure 3 compares the reasoning traces and predictions of Base, Vanilla GRPO, and VERGE on medical QA and visual grounding. In the QA example (a), Base and Vanilla GRPO repeatedly reconsider the options, generating 2,739 and 2,927 tokens, respectively, yet both select the incorrect periosteum answer. GRPO even identifies the correct endosteum option before reversing its decision. VERGE focuses on the relevant bone-remodeling functions and returns the correct answer in 214 tokens. In the grounding example (b), Base and GRPO focus on yellowish image regions and place their boxes away from the annotated polyp. Although GRPO shortens the response from 1,157 to 342 tokens, the localization error remains. VERGE instead attends to the small, rounded, slightly raised lesion and produces a box more closely aligned with the ground truth in 159 tokens. These examples illustrate that shorter generation alone does not ensure a better prediction: VERGE combines concise reasoning with attention to task-relevant evidence, avoiding the answer reversal and misplaced localization observed in the baseline traces.

6 Conclusion

In this work, we introduced VERGE, a reinforcement-learning framework for efficient medical multimodal reasoning. Our approach uses verified intermediate predictions to guide task-aware credit redistribution and explicit early-exit learning, distinguishing useful reasoning prefixes from redundant or harmful continuations. Experiments on medical question answering and visual grounding demonstrate improved answer accuracy and localization quality with substantially fewer output tokens. These results highlight the importance of learning not only how to reason, but also when to stop, supporting more accurate and efficient medical multimodal models.

AI Use Statement

We used AI to assist with drafting and refining sections of this manuscript. The authors have reviewed and thoroughly read all AI-assisted content within the paper and take full responsibility for the final content, including all AI-aided text and claims.

REFERENCES

Pranjal Aggarwal and Sean Welleck. L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*, 2025.

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Sharib Ali, Debesh Jha, Noha Ghatwary, Stefano Realdon, Renato Cannizzaro, Osama E Salem, Dominique Lamarque, Christian Daul, Michael A Riegler, Kim V Anonsen, et al. A multi-centre polyp detection and segmentation dataset for generalisability assessment. *Scientific Data*, 10(1): 75, 2023.

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*, 2025.

Muhammad Naseer Bajwa, Gur Amrit Pal Singh, Wolfgang Neumeier, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. G1020: A benchmark retinal fundus image dataset for computer-aided glaucoma detection. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2020.

Marthe Ballon, Brecht Verbeken, Vincent Ginis, and Andres Algaba. Probing the trajectories of reasoning traces in large language models. *arXiv preprint arXiv:2601.23163*, 2026.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Mateusz Buda, Ashirbani Saha, and Maciej A. Mazurowski. Association of Genomic Subtypes of Lower-Grade Gliomas with Shape Features Automatically Extracted by a Deep Learning Algorithm. *Computers in Biology and Medicine*, 109:218–225, 2019. doi: 10.1016/j.compbiomed.2019.05.002. URL https://arxiv.org/abs/1906.03720.

Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, et al. Multi-Centre, Multi-Vendor and Multi-Disease Cardiac Segmentation: The M&Ms Challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021. doi: 10.1109/TMI.2021.3090082. URL https://doi.org/10.1109/TMI.2021.3090082.

Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P. Musco, Rahul K. Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J. McDonald. Lung Segmentation in Chest Radiographs Using Anatomical Atlases with Nonrigid Registration. *IEEE Transactions on Medical Imaging*, 33(2):577–590, 2014. doi: 10.1109/TMI.2013.2290491. URL https://doi.org/10.1109/TMI.2013.2290491.

Benteng Chen, Weida Wang, Shufei Zhang, Mingbao Lin, and Min Zhang. Step-grpo: Internalizing dynamic early exit for efficient reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21710–21724, 2026.

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, 2024a. URL https://arxiv.org/abs/2406.19280.

Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, and Dong Yu. Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs, 2024b. URL https://arxiv.org/abs/2412.21187.

Muhammad E. H. Chowdhury, Tawsifur Rahman, Amith Khandakar, et al. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access*, 8:132665–132676, 2020. doi: 10.1109/ACCESS.2020.3010287. URL https://doi.org/10.1109/ACCESS.2020.3010287.

Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 168–172. IEEE, 2018. URL https://ieeexplore.ieee.org/document/8363547/.

Ankan Deria, Komal Kumar, Adinath Madhavrao Dukre, Eran Segal, Salman Khan, and Imran Razzak. Medmo: Grounding and understanding multimodal large language model for medical images. *arXiv preprint arXiv:2602.06965*, 2026.

Xingjian Diao, Zheyuan Liu, Chunhui Zhang, Weiyi Wu, Keyi Kong, Lin Shi, Kaize Ding, Soroush Vosoughi, and Jiang Gui. Addressing overthinking in large vision-language models via gated perception-reasoning optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4393–4410, 2026.

J. L. Glaister. Automatic Segmentation of Skin Lesions from Dermatological Photographs. Master's thesis, University of Waterloo, 2013. URL https://uwspace.uwaterloo.ca/items/f09aefc0-cc39-42e7-bf21-20e37ef0efd2.

Chen He, Yuhao Wu, Lei Wang, Wenxuan Zhang, and Fumin Shen. Diagnosing harmful continuation in answer-correct long-cot training traces. *arXiv preprint arXiv:2605.29288*, 2026.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Bairu Hou, Yang Zhang, Jiabao Ji, Yujian Liu, Kaizhi Qian, Jacob Andreas, and Shiyu Chang. ThinkPrune: Pruning long chain-of-thought of LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2504.01296.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K. Singh, Sameer Antani, George Thoma, Yi-Xiang Wang, Pu-Xuan Lu, and Clement J. McDonald. Automatic Tuberculosis Screening Using Chest Radiographs. *IEEE Transactions on Medical Imaging*, 33(2):233–245, 2014. doi: 10.1109/TMI.2013.2284099. URL https://doi.org/10.1109/TMI.2013.2284099.

Yuanfeng Ji, Haotian Bai, Jie Yang, Chongjian Ge, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhang, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A Large-Scale Abdominal Multi-Organ Benchmark for Versatile Medical Image Segmentation, 2022. URL https://arxiv.org/abs/2206.08023.

KiTS23. KiTS23: The 2023 Kidney and Kidney Tumor Segmentation Challenge. Official challenge website, 2023. URL https://kits-challenge.org/kits23/.

Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. NeoUNet: Towards Accurate Colon Polyp Segmentation and Neoplasm Detection, 2021. URL https://arxiv.org/abs/2107.05023.

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. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

M-A-P Team, Xinrun Du, Yifan Yao, Kaijing Ma, Bingli Wang, Tianyu Zheng, et al. SuperGPQA: Scaling LLM evaluation across 285 graduate disciplines. *arXiv preprint arXiv:2502.14739*, 2025. URL https://arxiv.org/abs/2502.14739.

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)*, pp. 353–367. PMLR, 2023.

Jose Ignacio Orlando, Huazhu Fu, Joao Barbossa Breda, Karel van Keer, Deepti R. Bathula, et al. REFUGE Challenge: A Unified Framework for Evaluating Automated Methods for Glaucoma Assessment from Fundus Photographs, 2019. URL https://arxiv.org/abs/1910.03667.

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*, pp. 248–260. 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*, pp. 337–347. Springer, 2025.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319, 2021.

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.

Chufan Shi, Cheng Yang, Yaokang Wu, Linghao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are vlms seeing or just saying? uncovering the illusion of visual re-examination. *arXiv preprint arXiv:2605.15864*, 2026.

Yan Shu, Chi Liu, Robin Chen, Derek Li, and Bryan Dai. Fleming-vl: Towards universal medical visual reasoning with multimodal llms. *arXiv preprint arXiv:2511.00916*, 2025.

Anas M Tahir, Muhammad EH Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M Sohel Rahman, Somaya Al-Maadeed, et al. Covid-19 infection localization and severity grading from chest x-ray images. *Computers in biology and medicine*, 139:105002, 2021.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

Vipin Venugopal, Justin Joseph, M Vipin Das, and Malaya Kumar Nath. Dtp-net: A convolutional neural network model to predict threshold for localizing the lesions on dermatological macro-images. *Computers in Biology and Medicine*, 148:105852, 2022.

Zihao Wei, Wenjie Shi, Liang Pang, et al. Dynamic rollout editing for reducing overthinking in RL-trained reasoning models, 2026. URL https://arxiv.org/abs/2606.17890.

Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, et al. Lingshu: Generalist foundation model for unified multimodal medical understanding and reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Zhonghao Yan, Muxi Diao, Yuxuan Yang, Ruoyan Jing, Jiayuan Xu, Kaizhou Zhang, Lele Yang, Yanxi Liu, Kongming Liang, and Zhanyu Ma. Medreasoner: Reinforcement learning drives reasoning grounding from clinical thought to pixel-level precision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11577–11585, 2026.

Guangjing Yang, Ziyuan Qin, Chaoran Zhang, Chenlin Du, Jinglin Wang, Wanran Sun, Zhenyu Zhang, Bing Ji, and Qicheng Lao. Medloc-r1: Performance-aware curriculum reward scheduling for grpo-based medical visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21036–21045, 2026.

Shenzhi Wang Zhangchi Feng Dongdong Kuang Yuwen Xiong Richong Zhang Yaowei Zheng, Junting Lu. Easyr1: An efficient, scalable, multi-modality rl training framework. `https://github.com/hiyouga/EasyR1`, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 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.

Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature methods*, 22(1): 166–176, 2025.

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.

A Data Collection Details

Data sources. We construct the training set using four data sources: MedMCQA Pal et al. (2022), PMC-VQA Zhang et al. (2023), BiomedParseData Zhao et al. (2025), and U-MRG-14K Yan et al. (2026). Together, these sources cover text-based medical question answering, image-conditioned medical question answering, and spatial grounding. Table 6 summarizes the included sources and their task and modality groups.

Medical question answering. MedMCQA supplies text QA examples, whereas PMC-VQA supplies visual QA examples that pair medical images with questions and answers. We retain this distinction when describing the data sources but combine both types under the QA task in Table 1.

Spatial grounding. We collect grounding examples from BiomedParseData and U-MRG-14K. BiomedParseData contributes biomedical segmentation annotations paired with textual descriptions, providing spatial supervision for anatomical structures and biomedical objects. The retained subdatasets cover diverse imaging modalities, such as MRI and CT. U-MRG-14K complements these modality-specific sources with clinically oriented queries paired with pixel-level localization annotations. We list it separately as the open-ended grounding group in Table 6.

B Evaluation Datasets and Protocols

B.1 Dataset Coverage

Medical QA. MMMU-Med comprises the Health and Medicine subjects of MMMU, and MMLU-Med covers nine subjects from MMLU. MedMCQA uses its labeled validation split. The QA average gives equal weight to each benchmark, regardless of its number of examples.

Table 6: Task groups and data sources used to construct train data. The seven modality-specific grounding groups are drawn from BiomedParseData; U-MRG-14K is listed separately as the open-ended grounding source.

Group	Task	Included sources
Medical QA	Text QA	MedMCQA
Medical VQA	Visual QA	PMC-VQA
MRI	Grounding	ACDC(Bernard et al., 2018), LGG(Buda et al., 2019), MMs(Campello et al., 2021), AMOS-MRI(Ji et al., 2022)
CT	Grounding	AMOS-CT(Ji et al., 2022), KiTS23(KiTS23)
X-ray	Grounding	COVID-QU-Ex(Tahir et al., 2021), CXR Masks(Jaeger et al., 2014; Candemir et al., 2014), Radiography(Chowdhury et al., 2020; Rahman et al., 2021)
Ultrasound	Grounding	BreastUS(Al-Dhabyani et al., 2020)
Fundus	Grounding	G1020(Bajwa et al., 2020), REFUGE(Orlando et al., 2019)
Skin imaging	Grounding	ISIC(Codella et al., 2018), UWaterloo Skin Cancer(Glaister, 2013; Venugopal et al., 2022)
Endoscopy	Grounding	NeoPolyp(Lan et al., 2021), PolypGen(Ali et al., 2023)
Open-ended	Grounding	U-MRG-14K

Medical grounding. The main comparison covers 44,839 annotations from 14 sources, excluding chest X-ray datasets. Results are grouped into computed tomography (CT), magnetic resonance imaging (MRI), ultrasound (US), endoscopy, fundus imaging, dermoscopy, skin photography (UWaterlooSkinCancer), and mixed-modality images (U-MRG-14K). Each group mean weights its annotations equally.

B.2 Evaluation Protocols

Baseline and checkpoints. Vanilla GRPO uses the same training data and configuration as VERGE, with sequence-level advantages and no exit loss or credit redistribution. The VERGE checkpoint specified in Section 5.2 is also used for both main comparisons. Vanilla GRPO is evaluated at step 374 in the main grounding comparison.

Decoding and scoring. Our evaluations generate one response per input using greedy decoding, with temperature 0, top-p 1.0, and seed 42. The maximum output length is 6,144 tokens for the main QA comparison and 4,096 tokens for grounding and component ablations. Base and VERGE share the same prompts, image preprocessing, and answer extraction within each comparison. Outputs are scored as generated, without forced answer completion or format repair: incomplete or unparseable QA answers are counted as incorrect, and invalid grounding predictions receive zero IoU. Token counts include both reasoning and the final answer.

B.3 Component Ablation Protocol

The QA ablation uses the training evaluation template and strict output-format checks, whereas the main QA comparison uses benchmark prompts and permits explanatory text outside the final answer block. MedMCQA scores in the two tables therefore reflect different prompting and scoring settings, in addition to the output limits specified above.

The grounding subset comprises ACDC, BreastUS, G1020, ISIC, LGG, MMs, NeoPolyp, PolypGen, REFUGE, UWaterlooSkinCancer, and U-MRG-14K. Its aggregate is weighted by dataset size after rounding per-dataset mIoU to two decimal places and mean token counts to one decimal place. These results describe the ablation subset rather than the full grounding evaluation cohort. The deltas in the final row of Table 5 are relative to Base; incremental component effects are discussed in the main text.

C CANDIDATE EXTRACTION AND BOUNDARY VERIFICATION

This appendix details candidate extraction and the scoring and verification performed by RCP before boundary selection in Section 4.1.

C.1 CANDIDATE EXTRACTION

For each original rollout y_i, we extract a small set $\mathcal{C}_i$ of token-exact semantic commitment points. For QA, a candidate corresponds to an explicit commitment to an answer option; for grounding, it corresponds to a complete, parsable bounding-box prediction. Each candidate c identifies an original reasoning prefix $P_i^{(c)}$. Candidate positions are located directly in the sampled token sequence, avoiding decode–truncate–retokenize operations that could shift the boundary. An intermediate prediction alone does not justify early exit: each candidate must also pass RCP verification.

C.2 RCP SCORING AND VERIFICATION

At each candidate c, RCP generates a completion $\widetilde{y}_i^{(c)}$ using Equation 3. We denote candidate and original full-prediction quality by $q_i^{(c)}$ and q_i^{full}, respectively. These scores measure task quality; verification additionally checks agreement between the prefix and probe predictions and applies the task-specific acceptance conditions below. Probe completions remain verification-only and do not enter the GRPO rollout group or policy surrogate.

QA. Let $a_i^{(c)}$ be the option committed to by the original prefix, a^* the reference answer, $\widetilde{a}_{i,1}^{(c)}$ the first option token generated by the probe, and $\widetilde{a}_{i,f}^{(c)}$ its parsed final answer. Quality is measured by binary option correctness: $q_i^{(c)} = \mathbf{1}[a_i^{(c)} = a^*]$ and $q_i^{\text{full}} = \mathbf{1}[a_i^{\text{full}} = a^*]$, where a_i^{full} is the parsed answer of the original full rollout. A candidate is accepted only when

$$a_i^{(c)} = a^* = \widetilde{a}_{i,1}^{(c)} = \widetilde{a}_{i,f}^{(c)}, \qquad p_{i,1}^{(c)} \geq \tau_{\text{QA}}, \tag{10}$$

where $p_{i,1}^{(c)}$ is the probability assigned to the first probe answer token. Thus, a correct prefix option is insufficient on its own: the probe must agree with it and meet the minimum confidence requirement.

Grounding. Let $b_i^{(c)}$, $\widetilde{b}_i^{(c)}$, b_i^{full}, and b^* denote the prefix box, probe box, original full-prediction box, and ground-truth box, respectively. The candidate quality $q_i^{(c)}$ is its continuous localization IoU, and $q_i^{\text{full}} = \text{IoU}(b_i^{\text{full}}, b^*)$. IoU thresholds determine acceptance without binarizing these quality scores. A candidate must satisfy

$$\begin{aligned} \text{IoU}(b_i^{(c)}, b^*) &\geq \tau_{\text{G}}, \\ \text{IoU}(\widetilde{b}_i^{(c)}, b^*) &\geq \tau_{\text{G}}, \\ \text{IoU}(b_i^{(c)}, \widetilde{b}_i^{(c)}) &\geq \tau_{\text{cons}}. \end{aligned} \tag{11}$$

The first two constraints require both predictions to meet the localization quality threshold; the third checks spatial consistency under the probe intervention.

The relative-quality check uses the same candidate score $q_i^{(c)}$ and the original full-prediction score q_i^{full}. For an originally correct rollout, early exit must preserve localization quality up to a tolerance δ:

$$q_i^{(c)} \geq q_i^{\text{full}} - \delta. \tag{12}$$

For an originally incorrect rollout, the candidate must instead provide a minimum improvement Δ:

$$q_i^{(c)} - q_i^{\text{full}} \geq \Delta. \tag{13}$$

A grounding candidate is accepted only if it satisfies both Equation 11 and the applicable relative-quality constraint.

Only candidates passing all applicable checks enter $\mathcal{C}_i^{\text{ver}}$. Boundary selection then follows Equation 4, with $q_i^{\text{cut}} = q_i^{(c_i^\star)}$ denoting the selected candidate's score. Thus, the grounding quality difference $q_i^{\text{cut}} - q_i^{\text{full}}$ in Section 4.2 remains a continuous IoU difference. No selected-cut score is used when no candidate is verified.