THINK AGAIN AND LOOK CLOSER: DUAL-EXPLORATION TEST-TIME REINFORCEMENT LEARNING FOR MEDICAL MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical Multimodal Large Language Models (MLLMs) have demonstrated remarkable capabilities across healthcare tasks, yet their post-training strategies still heavily depend on substantial annotated data, overlooking the potential of large-scale unlabeled clinical data. An alternative approach, test-time reinforcement learning, operates by sampling multiple trajectories for each test input and exploiting the consensus among their answers as a self-supervised signal. However, for medical visual question answering (VQA), consensus among reasoning answers alone does not constitute reliable evidence. From a *textual* perspective, over-reliance on answer consensus can amplify confidence in unsubstantiated conclusions and may even encourage shorter, less informative responses that shortcut reasoning in favor of textual agreement. *Visually*, when multiple sampled trajectories converge on the same answer, their consensus should be grounded in consistent visual evidence rather than textual consistency alone. *Clinically*, physicians focus intently on subtle abnormalities, whereas even with attention mechanisms, general-purpose models are distracted by large irrelevant areas in the whole image. These limitations call for test-time learning improvements that promote consensus without textual shortcuts, visual grounding consistency across trajectories, and localized exploration beyond holistic features. We therefore introduce TALC, a dual-exploration test-time reinforcement learning framework that asks the model to Think Again and Look Closer. Think Again first anchors on answer support, then ranks reasoning paths by semantic novelty within each support group while also measuring visual consistency across trajectories. Look Closer, in turn, through explicit and executable region-selection actions, assigns each region proposal a tool-token uncertainty score, and branches from that decision point to resample alternative local inspections. Experiments across multiple medical VQA benchmarks and MLLM backbones demonstrate the effectiveness and generalizability of TALC for test-time medical multimodal reasoning. Our code is available at: `https://anonymous.4open.science/r/TALC`.

1 INTRODUCTION

Medical multimodal large language models, referred to as MLLMs, have demonstrated strong potential across healthcare tasks (Ma et al., 2026), yet their adaptation for medical applications remains challenging. Direct inference often leaves substantial room for improvement, while supervised fine-tuning (Xu et al., 2025; Wang et al., 2025) and reinforcement learning (Zhu et al., 2025; Zhi et al., 2026) continue to rely on large volumes of annotated data. In medicine, this dependence is especially costly, since labeled examples are difficult to obtain, easily raise privacy concerns during collection, and are expensive to verify (Sood et al., 2025; Wac et al., 2025). Clinical cases also vary across institutions, scanners, and patient populations, which necessitates continued adaptation after deployment (Li et al., 2024; 2026). These constraints have driven test-time computation as a way to leverage unlabeled inputs for additional predictive gains. A concrete realization of this paradigm is test-time reinforcement learning, which samples multiple trajectories per test input and uses output consensus as a self-supervised signal (Zuo et al., 2025b).

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.

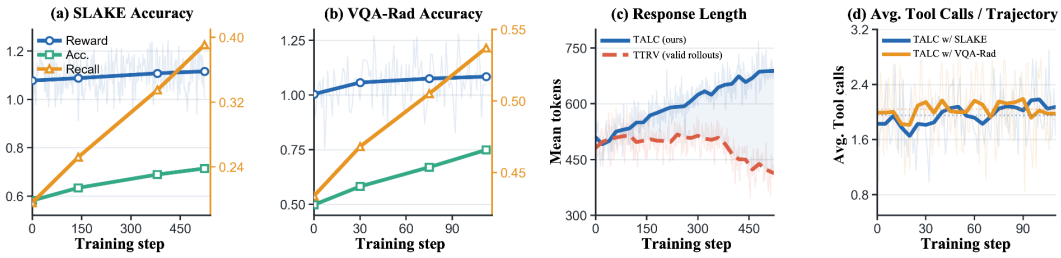


Figure 1: Training dynamics during test-time adaptation with Qwen2.5-VL-7B-Instruct as the backbone. (a,b) Reward, accuracy, and recall on SLAKE and VQA-RAD, respectively. (c) Mean response length for TALC and TTRV. (d) Average tool calls per trajectory for TALC.

The attraction of this paradigm is clear, but its central assumption that consensus among sampled outputs reflects correctness is fragile. Methods such as TTRL (Zuo et al., 2025b) and TTRV (Singh et al., 2026) leverage answer-frequency statistics to guide policy optimization at test time. This implicitly treats frequency as a proxy for correctness, yet the underlying premise that repeated samples provide independent evidence does not hold with all trajectories coming from the same model and therefore sharing its response biases, language priors, and visual blind spots (Manakul et al., 2023). This inevitably calls into question:

> *whether such consensus can serve as a reliable learning signal, particularly in medicine, where diagnostic decisions demand far more than output agreement.*

We examine this question in the context of medical visual question answering and identify three fundamental limitations of consensus-based reasoning.

First, since the learning signal depends solely on output consensus, once a particular response becomes the majority, the policy update reinforces it, creating an echo chamber in which repetition, not correctness, drives confidence (Zhou et al., 2025). Although TTRV attempts to mitigate this by adding entropy to every candidate, GRPO normalization eliminates this shared term and leaves their relative advantages unchanged (Shao et al., 2024). This, in turn, encourages shorter and more generic responses that are more likely to match other samples, favoring surface-level alignment over grounded reasoning (Zhou et al., 2025). Second, consensus alone provides no guarantee that the model has examined consistent visual evidence (Yu et al., 2026). Since standard test-time reinforcement learning verifies only the final answer agreement, it never checks whether the consensus is grounded in consistent visual evidence. Two trajectories may reach identical conclusions while attending to completely different image regions, or even relying on textual priors without ever consulting the visual input, making such consensus diagnostically vacuous. Third, in clinical practice, physicians direct their attention to localized regions, such as the site of a suspected lesion, an abnormal contour, or a subtle density shift, where the evidence needed for diagnosis lies. However, even with attention mechanisms, general-purpose MLLMs remain distracted by large irrelevant regions that dominate the visual field and suppress the signal from clinically meaningful areas (Nguyen et al., 2025; Jiang et al., 2026), while consensus-based reinforcement learning, by rewarding only final consensus, offers no incentive for the model to overcome this distraction.

Together, these observations reveal questions hidden behind consensus in three respects, textually, visually, and clinically: whether agreement truly reflects reasoning quality, whether textual consensus alone is sufficient for visual question answering, and whether such a method follows the diagnostic logic of physicians. Based on these observations, we introduce TALC, a dual-exploration test-time reinforcement learning framework that asks the model to *Think Again* and *Look Closer*. TALC adopts a structured generation interface that organizes model outputs into executable visual actions and standardized final responses. This structured output format provides the foundation for two complementary explorations, *Think Again* and *Look Closer*. *Think Again* first anchors on answer support and then ranks reasoning paths by semantic novelty within each support group. Through this novelty-aware selection, TALC mitigates the late-stage response shortening observed in TTRV, as shown in Fig. 1(c). Beyond semantic diversity, Think Again further measures visual consistency across trajectories, ensuring that their convergence is supported by shared visual evidence rather than textual repetition alone. *Look Closer*, through explicit and executable region-selection actions, assigns each region proposal an uncertainty score and branches from that decision point to resample alternative local inspections beyond holistic image features. This uncertainty-driven branching en-

ables local evidence acquisition, with Fig. 1(d) showing roughly two tool calls per trajectory. The two mechanisms form a closed loop: *Look Closer* supplies diverse region-level evidence for *Think Again*, while *Think Again*, through its reward signal, guides which region proposals are worth exploring for *Look Closer*, addressing the textual, visual, and clinical limitations of consensus-based learning without requiring any external annotations. The resulting adaptation dynamics are reflected in Fig. 1(a,b), where reward improvements on SLAKE and VQA-RAD are accompanied by gains in answer accuracy and recall. Our contributions are threefold:

- We identify three fundamental limitations, textual, visual, and clinical, that reveal why consensus is an unreliable supervisory signal for medical test-time reinforcement learning.
- We propose TALC, a dual-exploration test-time reinforcement learning framework for medical multimodal reasoning that couples Think Again and Look Closer to enable grounded reasoning and region-level visual exploration without external annotations.
- We evaluate TALC across multiple medical VQA benchmarks and MLLM backbones, demonstrating its effectiveness and generalizability for test-time medical multimodal reasoning.

2 Related Work

Unsupervised Reasoning at Test Time. Test-time computation improves reasoning without additional annotations by allocating extra inference-time computation to sampling, search, or adaptation. Self-consistency established a simple yet effective instance of this paradigm by sampling multiple reasoning paths and selecting the answer through majority voting (Wang et al., 2023), whereas TTRL converts that majority into pseudo-rewards for policy optimization on unlabeled inputs (Zuo et al., 2025b). This principle has been extended to multimodal reasoning by MM-UPT and TTRV (Wei et al., 2025; Singh et al., 2026). Beyond answer-level consensus, model-intrinsic uncertainty provides another source of guidance for reasoning and adaptation. EM-INF reduces next-token entropy through inference-time logit optimization without updating model weights (Agarwal et al., 2025), while Intuitor uses self-certainty as an intrinsic reward for GRPO without relying on answer voting (Zhao et al., 2026b). ECHO combines majority-vote rewards with entropy- and confidence-guided tree exploration, adaptive clipping, and advantage shaping to mitigate rollout collapse and premature convergence under noisy pseudo-labels (Zhao et al., 2026a). Unlike methods relying solely on answer consensus or model confidence, TALC couples semantic exploration with cross-trajectory visual grounding to discourage agreement unsupported by consistent visual evidence.

Visual Localization and Cross-Trajectory Grounding Consistency. Existing work on visual grounding in multimodal reasoning follows two complementary directions: strengthening dependence on visual evidence and improving the ability to locate task-relevant regions. Perception-R1 rewards reasoning that retains task-relevant visual facts (Xiao et al., 2026), PAPO contrasts behavior under intact and perturbed images (Wang et al., 2026), and CFPO estimates visual contribution through counterfactual interventions on cross-modal representations (Yu et al., 2026). Beyond encouraging dependence on visual input, another line of work explores how models locate and inspect task-relevant image regions to support reasoning. Localizing Before Answering studies whether medical multimodal models can identify task-relevant regions before answering (Nguyen et al., 2025). Building on the importance of such localization, MedVR formulates medical reasoning as an iterative process of selecting and inspecting image regions without requiring region annotations (Jiang et al., 2026). TALC lies at the intersection of these two lines of work: Look Closer branches at uncertain region actions to acquire alternative local observations, and Think Again encourages reliance on shared visual evidence through spatial agreement with a consensus mask.

3 Method

3.1 Overview

These textual, visual, and clinical limitations motivate three interrelated design requirements for test-time learning: encouraging localized visual inspection, promoting visual consistency, and broadening reward assignment beyond answer-level consensus. Together, these requirements call for test-time adaptation that goes beyond terminal-answer agreement. To this end, we introduce TALC, illustrated in Fig. 2, a method that couples two complementary mechanisms: Look Closer, which revisits

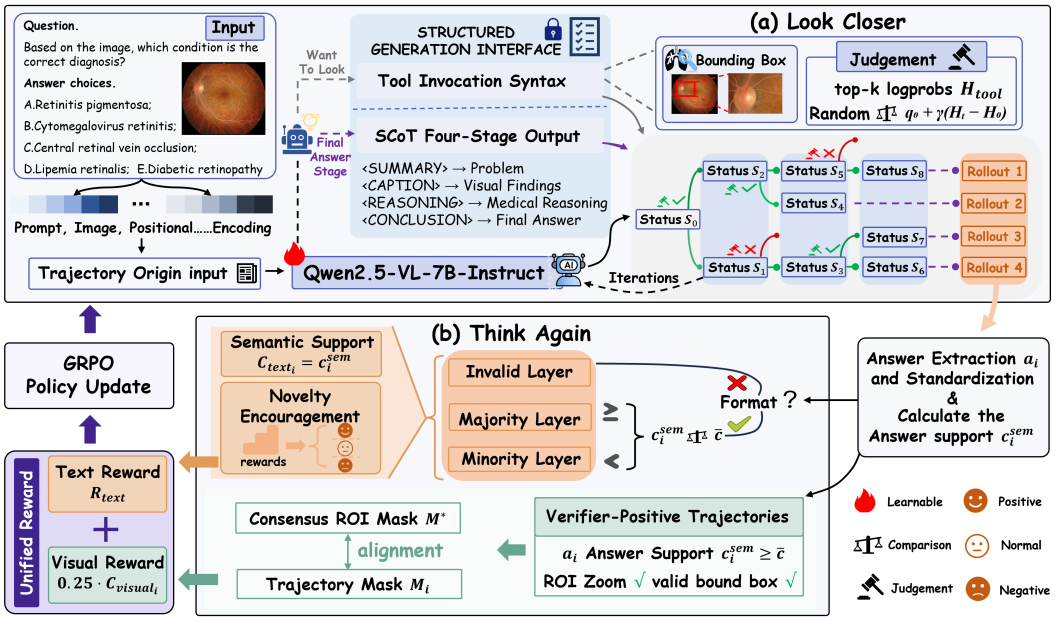


Figure 2: Overview of TALC. The Structured Generation Interface produces executable region requests or four-stage final responses. (a) Look Closer restores uncertain region-request states and samples alternative local observations under a fixed rollout budget. (b) Think Again groups valid trajectories by semantic support, rewards non-redundant reasoning, and compares each inspection mask with the supported consensus mask. The combined textual and visual reward drives GRPO.

uncertain region choices to gather alternative visual evidence, and Think Again, which reassesses completed trajectories by answer support, semantic novelty, and cross-trajectory spatial consistency. However, Look Closer requires the model to specify region locations, while Think Again requires final responses to follow a comparable form for cross-trajectory assessment. This motivates a shared structured generation interface. Accordingly, TALC constrains the model's outputs through predefined formatting rules, restricting decoding to structurally valid action and response formats. In the following, we first introduce the interface, then detail Look Closer, and finally describe Think Again.

3.2 Structured Generation Interface

As the subsequent operations in TALC rely on well-structured outputs, TALC follows XGrammar (Dong et al., 2025) and introduces a Structured Generation Interface to make model outputs controllable and well-structured. Specifically, this interface serves two purposes: it provides executable visual actions for Look Closer, and yields comparable final responses for Think Again.

To fulfill these purposes, model outputs are restricted to two types of forms. A region request takes the form $\textsc{RequestRegion}(m, \hat{b})$, where m selects an image and $\hat{b} = (l, t, r, b)$ is a normalized box. This form defines an executable action carried out by Look Closer on the immutable source image, with the resulting crop appended as a new observation to guide the model toward region-focused visual inspection. A finalization takes the form $\textsc{Finalize}(y)$, where the terminal response follows:

$$y = (\textsc{Summary}, \textsc{Caption}, \textsc{Reasoning}, \textsc{Conclusion}),$$

which provides deterministic boundaries for answer extraction and reasoning comparison, enabling Think Again to assess semantic novelty and cross-trajectory consistency.

To implement this constraint, TALC uses grammar-constrained decoding. For a decoding prefix h, let $\mathcal{A}_G(h)$ denote the set of next tokens that preserve at least one valid completion under grammar G. At each step, the policy is restricted to this set and normalized over the remaining tokens:

$$\pi_\theta^G(v \mid h, x) = \frac{\pi_\theta(v \mid h, x)\mathbf{1}[v \in \mathcal{A}_G(h)]}{\sum_{u \in \mathcal{A}_G(h)} \pi_\theta(u \mid h, x)}. \tag{1}$$

This projection excludes structurally invalid continuations while leaving reasoning and region choices unconstrained, enabling subsequent stages to reliably parse responses and execute actions.

3.3 LOOK CLOSER: UNCERTAINTY-DRIVEN LOCAL VISUAL EXPLORATION

Motivated by clinicians' focus on localized diagnostic evidence, Look Closer revisits uncertain region choices to acquire alternative local observations. Building on the structured interface, it estimates uncertainty over region-request decisions rather than the entire trajectory. This targeted estimation contrasts with naive whole-trajectory resampling, which wastes the limited rollout budget and conflates spatial ambiguity with linguistic ambiguity.

Concretely, Look Closer treats each region request as a decision point and quantifies how strongly the model hesitates over its spatial choice. For the t-th region request generated after prefix h_t, let $\mathcal{L}_t$ denote the token span of its action $\textsc{RequestRegion}(m, \hat{b})$. We compute the normalized predictive entropy over the retained top-k candidates:

$$H_t = -\frac{1}{\log |\mathcal{V}_{\text{tok}}|} \sum_{\ell \in \mathcal{L}_t} \sum_{j \in \text{Top } k(\ell)} p_{\ell j} \log p_{\ell j}, \tag{2}$$

where $p_{\ell j}$ is the probability of candidate j at position ℓ, and $\mathcal{V}_{\text{tok}}$ is the tokenizer vocabulary. H_t measures uncertainty over region-request tokens, excluding reasoning and answer tokens.

Absolute entropy alone is unreliable for deciding when to branch, because different images and questions induce different baseline levels of spatial uncertainty. A challenging image may yield high entropy from overall difficulty, while an easy image may yield low entropy despite an increase from its baseline. To account for these variations, the first valid region request in a trajectory provides a reference H_0. The deviation $H_t - H_0$ measures how much more uncertain the current region choice is than the first decision. This relative uncertainty controls the branching probability:

$$q_t = \text{clip}(q_0 + \gamma(H_t - H_0), 0, 1)\,, \tag{3}$$

where q_0 is a base exploration rate and γ controls how strongly uncertainty reallocates the remaining rollout budget. A larger positive deviation means the model is more hesitant about this region choice relative to its earlier decision, and thus more likely to trigger an alternative inspection.

Once a branch is triggered according to q_t, TALC keeps the original proposal and returns to the pre-action prefix h_t to generate an alternative region:

$$h_t \rightarrow \begin{cases} \textsc{RequestRegion}(m, \hat{b}_t), \\ \textsc{RequestRegion}(m', \hat{b}'_t). \end{cases} \tag{4}$$

The two descendants share all preceding text and observations and first diverge at the region-selection decision. Compared with independent rollouts, this shared-prefix construction allows alternative region choices to be examined from the same context, reducing variation introduced by different preceding reasoning paths. Although subsequent generation may also affect their outcomes, this control makes comparisons between alternative inspections more interpretable and helps clarify how regional observations inform the subsequent reasoning. Importantly, H_t guides region-level exploration, while Think Again evaluates the completed trajectories based on their resulting evidence and reasoning. Through this division of roles, Look Closer broadens region-level evidence acquisition without relying on external region annotations.

3.4 THINK AGAIN: SUPPORT-CONSTRAINED TRAJECTORY REASSESSMENT

After Look Closer produces alternative regional observations, TALC must select which completed trajectories should guide the policy update. A natural choice is answer consensus, but exact-answer voting treats semantically equivalent answers as different, over-rewards repeated reasoning, and never checks whether agreeing trajectories inspected consistent image regions. Think Again addresses these gaps at two levels. At the textual level, semantic support measures corroboration across trajectories while semantic novelty acts as a variation signal, ranking reasoning paths within each support group to distinguish equally supported trajectories. At the visual level, it derives a spatial hypothesis from high-support trajectories and evaluates each trajectory's alignment with it, reducing rewards for textually supported but visually inconsistent trajectories.

Joint semantic support and novelty. At the textual level, Think Again compares trajectories through their conclusions and reasoning in semantic space, measuring support from conclusions

and novelty from reasoning, rather than through surface-level agreement. Before any comparison, a trajectory must first be valid: it ends with a complete FINALIZE(y) and both its REASONING and CONCLUSION fields are non-empty. For each valid trajectory τ_i, let r_i and a_i denote its REASONING and CONCLUSION fields. We encode them separately with a frozen encoder f_ϕ (Zhang et al., 2025):

$$e_i^a = \frac{f_\phi(a_i)}{\|f_\phi(a_i)\|_2}, \qquad e_i^r = \frac{f_\phi(r_i)}{\|f_\phi(r_i)\|_2},$$

where e_i^a represents the conclusion and e_i^r represents the reasoning. The selection signal is computed over conclusions: pairwise conclusion affinity is $\kappa_{ij}^a = (1+e_i^{a\top} e_j^a)/2$, and the soft semantic support of trajectory i is the average conclusion affinity to other trajectories in the valid set $\mathcal{V}$,

$$c_i^{\text{sem}} = \frac{1}{|\mathcal{V}|-1} \sum_{j\in\mathcal{V},\, j\neq i} \kappa_{ij}^a,$$

with $c_i^{\text{sem}} = 0$ when $|\mathcal{V}| < 2$. This support replaces exact-match voting: semantically equivalent conclusions are recognized as mutually supporting even when their surface forms differ.

Support alone, however, provides only the selection signal. It does not distinguish trajectories that reach the same conclusion through different reasoning paths. To apply the variation signal, Think Again partitions valid trajectories by their support. With mean support $\bar{c}$, we define:

$$\mathcal{G}_{\text{high}} = \{i : c_i^{\text{sem}} \geq \bar{c}\}, \qquad \mathcal{G}_{\text{low}} = \mathcal{V} \setminus \mathcal{G}_{\text{high}}.$$

This adaptive split requires neither an exact majority answer nor a fixed threshold. Within each stratum, novelty is computed from reasoning embeddings rather than conclusions. We compute a local mean similarity to same-stratum peers and a global maximum similarity to any other trajectory:

$$\bar{\kappa}_i^g = \underset{j\neq i,\, g_j=g_i}{\text{mean}} \kappa_{ij}^r, \qquad \kappa_i^{\max} = \max_{j\neq i} \kappa_{ij}^r,$$

where $\kappa_{ij}^r = (1 + e_i^{r\top} e_j^r)/2$. The novelty score is then

$$n_i = 1 - \left[\alpha\bar{\kappa}_i^g + (1-\alpha)\kappa_i^{\max}\right],$$

with $\alpha = 0.5$. This score penalizes redundancy at two levels: a high $\bar{\kappa}_i^g$ marks a trajectory as merely typical of its group, while a high $\kappa_i^{\max}$ marks it as a copy of another trajectory. In both cases, the trajectory provides redundant reasoning that contributes no new evidence and wastes the rollout budget. By subtracting both terms, novelty is high only when the trajectory departs from its group norm and does not duplicate any peer. Within each stratum, novelty scores are normalized to $[0, 1]$, denoted $\bar{n}_i$ while singleton or constant strata receive zero. The variation reward then maps the two strata to non-overlapping bands:

$$E_i = \begin{cases} -1, & i \notin \mathcal{V}, \\ 0.5 + 0.5\bar{n}_i, & i \in \mathcal{G}_{\text{high}}, \\ -1 + 0.5\bar{n}_i, & i \in \mathcal{G}_{\text{low}}. \end{cases}$$

This reward guarantees any high-support trajectory, regardless of novelty, receives a higher reward than any low-support one. The final textual reward combines the selection and variation signals:

$$R_i^{\text{text}} = c_i^{\text{sem}} + \lambda E_i,$$

where λ balances the two terms. Consequently, a distinctive low-support trajectory cannot outrank a high-support one, while the variation reward within each stratum still encourages distinct reasoning paths, avoiding collapse toward a single explanation driven by semantic priors.

Cross-trajectory visual grounding consistency. Semantically supported reasoning may still be produced from incompatible regions or textual priors alone. To verify visual consistency, Think Again maps each executed region request back to source-image coordinates and merges them into an inspection mask M_i. A trajectory is marked $z_i = 1$ when $i \in \mathcal{G}_{\text{high}}$ and it contains at least one valid region request. This condition uses semantic support, rather than any external label, to decide which inspections can define a spatial hypothesis. The supported set $\mathcal{S} = \{i : z_i = 1\}$ defines:

$$M^\star = \mathbf{1}\left[\frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} M_i > \tau_c\right], \qquad \rho_i = \text{IoU}(M_i, M^\star),$$

where τ_c is a threshold, $M^\star$ retains regions shared by more than τ_c of the supported trajectories and thus summarizes their common spatial evidence, while ρ_i measures the IoU between trajectory i and this group-derived pattern. If no supported trajectory contains a valid region request, no spatial hypothesis can be constructed, and visual credit is omitted for that rollout group.

The visual credit assigns three levels based on semantic support and spatial alignment:

$$C_i^{\text{visual}} = \begin{cases} 1, & z_i = 1 \text{ and } \rho_i \geq \tau_a, \\ 0.5, & (z_i = 1 \text{ and } \rho_i < \tau_a) \text{ or } (z_i = 0 \text{ and } \rho_i \geq \tau_a), \\ 0, & \text{otherwise}, \end{cases}$$

where τ_a is the required alignment threshold. Together with the textual reward, this visual credit encourages trajectories to be both semantically supported and spatially consistent. In this way, Think Again closes the loop with Look Closer, ensuring that the visual evidence acquired by exploration is actually used to shape the final credit assignment.

3.5 Joint Test-Time Optimization

For each completed trajectory, Think Again produces two reward terms: R_i^{text} which combines semantic support and novelty in text space, and C_i^{visual} which measures spatial consistency with the group inspection pattern. TALC sums these two terms into a single reward:

$$R_i = R_i^{\text{text}} + \lambda_v C_i^{\text{visual}}, \qquad \lambda_v = 0.25.$$

This reward serves as the trajectory-level reward for a GRPO (Shao et al., 2024) update, where advantages are obtained by group-wise standardization within the rollout group. The policy is then optimized with clipped GRPO, KL regularization to the initial policy, and an entropy bonus:

$$\mathcal{L} = \mathcal{L}_{\text{GRPO}} + \beta_{\text{KL}} \mathcal{D}_{\text{KL}}(\pi_\theta \| \pi_{\theta_0}) - \beta_{\text{ent}} \mathcal{H}(\pi_\theta).$$

Here, the KL regularization and entropy bonus both prevent the policy from collapsing too early under self-generated rewards. The KL term keeps the policy close to its initialization, while the entropy bonus keeps the output distribution from becoming overly sharp.

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We evaluate on four public medical VQA benchmarks with their frozen test splits: SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), VQA-Med 2019 (Ben Abacha et al., 2019), and MedXpertQA-MM (Zuo et al., 2025a). Together, they cover general medical image understanding, radiology question answering, and expert-level multimodal questions; MedXpertQA-MM additionally contains examples with multiple images (one to six per question). For performance evaluation, we report accuracy for closed-ended and multiple-choice questions, and token-level Recall and ROUGE-1 (Lin, 2004) for open-ended questions. Since the released VQA-Med test set does not provide open-/closed-ended annotations, we classify a sample as closed-ended only when all of its normalized reference answers are exactly "yes" or "no"; all remaining samples are treated as open-ended. MedXpertQA-MM contains only closed-ended questions, so we report Acc alone on this benchmark.

Implementation Details. We study two backbone settings: Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and MedVLM-R1 (Pan et al., 2025). Test-time adaptation follows a transductive and label-free setting: for each dataset and random seed, adaptation starts from the corresponding unadapted checkpoint and uses only the test images and questions. Reference answers are excluded from reward computation, policy updates, and checkpoint selection, and are used only by the evaluator after adaptation. Both adaptation and the matched evaluation use a temperature of 0.6 and top-p of 0.95, and a response budget of 1,536 tokens to ensure stable and reproducible adaptation behavior. Policy updates use a batch size of 4 and an actor learning rate of 5×10^{-7}. Policy optimization is implemented with FSDP and vLLM 0.8.3 and conducted on four NVIDIA A100 GPUs, while a frozen Qwen3-Embedding-4B model provides sentence embeddings through a local service. To avoid confounding effects from output formatting and answer extraction, all methods in the main

Table 1: Performance comparison on four medical VQA datasets using Qwen2.5-VL-7B-Instruct as the base model. Bold indicates the best performance for each metric.

	SLAKE			VQA-RAD			VQA-Med			MedXpertQA-MM
Method	Acc	Recall	ROUGE-1	Acc	Recall	ROUGE-1	Acc	Recall	ROUGE-1	Acc
Base Model	58.13	19.56	24.05	49.80	43.35	20.68	62.50	24.10	15.39	23.80
EM-INF (Agarwal et al., 2025)	59.69	19.74	24.24	57.37	41.50	21.13	65.63	21.15	17.27	23.75
TTRL (Zuo et al., 2025b)	60.89	21.11	25.62	58.57	42.12	20.39	64.06	26.13	18.32	22.85
TTRV (official) (Singh et al., 2026)	55.38	17.42	23.81	51.79	38.64	19.46	39.06	23.78	16.42	14.20
TTRV (valid rollouts) (Singh et al., 2026)	59.69	18.55	25.36	54.18	40.37	20.19	42.19	25.58	17.75	15.65
Intuitor (Zhao et al., 2026b)	24.88	23.51	24.65	54.98	44.80	20.70	67.19	**27.24**	17.30	15.30
ECHO (Zhao et al., 2026a)	58.01	21.68	24.89	52.59	39.95	19.72	64.06	25.73	16.55	23.10
Ours	**71.41**	**39.08**	**37.70**	**74.90**	**53.70**	**49.38**	**68.75**	26.43	**22.76**	**26.25**

Table 2: Performance comparison on four medical VQA datasets using MedVLM-R1 as the base model.

	SLAKE			VQA-RAD			VQA-Med			MedXpertQA-MM
Method	Acc	Recall	ROUGE-1	Acc	Recall	ROUGE-1	Acc	Recall	ROUGE-1	Acc
Base Model	61.36	17.99	26.82	49.00	13.06	19.33	43.75	9.58	17.52	20.35
EM-INF (Agarwal et al., 2025)	68.78	25.21	33.50	58.57	17.22	24.58	46.88	10.18	18.54	20.75
TTRL (Zuo et al., 2025b)	71.29	26.20	36.72	57.37	17.09	25.02	39.06	6.59	12.43	**23.40**
TTRV (official) (Singh et al., 2026)	64.23	23.82	31.56	55.78	15.58	20.18	40.63	8.98	16.51	14.50
TTRV (valid rollouts) (Singh et al., 2026)	68.42	25.64	33.74	59.36	17.23	21.46	45.31	10.18	18.49	16.00
Intuitor (Zhao et al., 2026b)	28.23	21.04	24.12	55.38	16.39	20.82	**51.56**	**13.17**	**23.32**	13.80
ECHO (Zhao et al., 2026a)	63.40	20.15	27.65	53.78	14.18	19.33	45.31	9.88	17.98	19.65
Ours	**74.04**	**29.00**	**39.50**	**66.53**	**22.02**	**28.74**	48.44	10.78	19.52	20.20

experiments use the same structured format and the same CONCLUSION answer extractor; this unified interface is also kept active throughout all ablation studies. Region requests are enabled only when Look Closer is active, and structural tokens are excluded from uncertainty estimation. Thus, performance differences mainly reflect the learning and exploration mechanisms themselves.

4.2 Main Results

Qwen2.5-VL-7B results. Table 1 shows that TALC achieves the best performance on nine of the ten reported metrics, demonstrating consistently strong gains across datasets. Compared with the strongest non-TALC baseline, it improves accuracy by 10.52 points on SLAKE and 16.33 points on VQA-RAD, while increasing VQA-RAD ROUGE-1 by 28.25 points. TALC also improves VQA-Med accuracy from 62.50 to 68.75 and achieves the best ROUGE-1 score of 22.76, although its Recall of 26.43 is slightly below Intuitor's 27.24. Following TTRL's original setup, all methods discard invalid rollouts before reward computation. Because the official TTRV implementation retains such rollouts, we report TTRV (official) and TTRV (valid rollouts), which applies the filtering rule. The valid-rollout setting strengthens TTRV and serves as the protocol-matched comparison, while the official setting is included as an implementation-faithful reference.

MedVLM-R1 results. Table 2 shows that TALC achieves the best results on all SLAKE and VQA-RAD metrics. It improves the strongest alternatives by 2.75–2.80 points on SLAKE and 3.72–7.17 points on VQA-RAD. TALC also raises VQA-Med accuracy from 43.75 for the base model to 48.44, showing that the method transfers effectively to a medically specialized backbone.

Ablation analysis. Both components of TALC improve TTRL on two backbones in Table 3. Look Closer contributes a larger gain in accuracy, whereas Think Again yields a greater improvement in Recall. Combining two components delivers the strongest performance across all metrics, improving Qwen2.5-VL-7B over TTRL by 10.52 points in accuracy, 17.97 points in Recall, and 12.08 points in ROUGE-1, with corresponding gains of 2.75, 2.80, and 2.78 points for MedVLM-R1.

Entropy analysis. Fig. 3(a) shows that average token entropy decreases from 0.40 to 0.33 as region mIoU increases, linking spatially consistent inspection with more concentrated responses. In Fig. 3(b), TTRL entropy steadily declines to approximately 0.23, whereas Think Again and full TALC maintain stable policy entropy, with full TALC finishing near 0.39. Such sustained entropy decay suggests reduced exploration and an increasing risk of policy collapse. The combined method therefore preserves diverse policy behavior while producing increasingly consistent visual evidence.

Table 3: Ablation study of TALC on the SLAKE dataset using Qwen2.5-VL-7B-Instruct and MedVLM-R1 as backbone models.

Qwen2.5-VL-7B-Instruct			
Method	Acc	Recall	ROUGE-1
TTRL	60.89	21.11	25.62
TTRL + Think Again	61.36	26.84	27.95
TTRL + Look Closer	69.50	23.42	32.18
Full TALC	71.41	39.08	37.70

MedVLM-R1			
Method	Acc	Recall	ROUGE-1
TTRL	71.29	26.20	36.72
TTRL + Think Again	71.77	28.15	37.14
TTRL + Look Closer	73.56	27.04	38.63
Full TALC	74.04	29.00	39.50

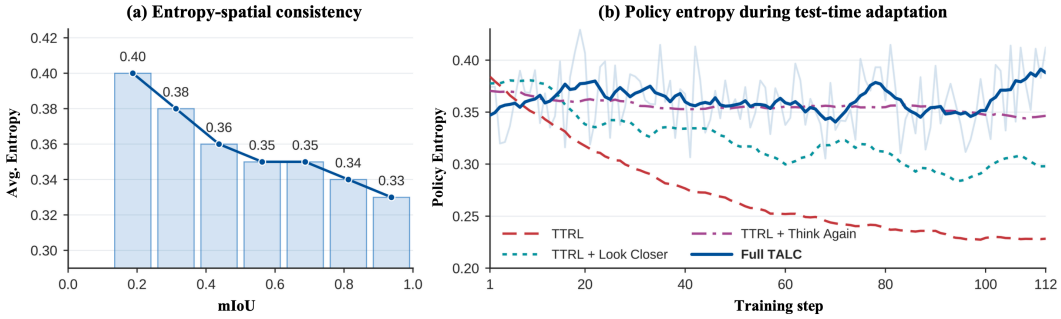


Figure 3: (a) Correlation between token entropy and mIoU. (b) Policy entropy over the first 112 test-time adaptation steps. Both are based on SLAKE (SLAKE provides bounding boxes).

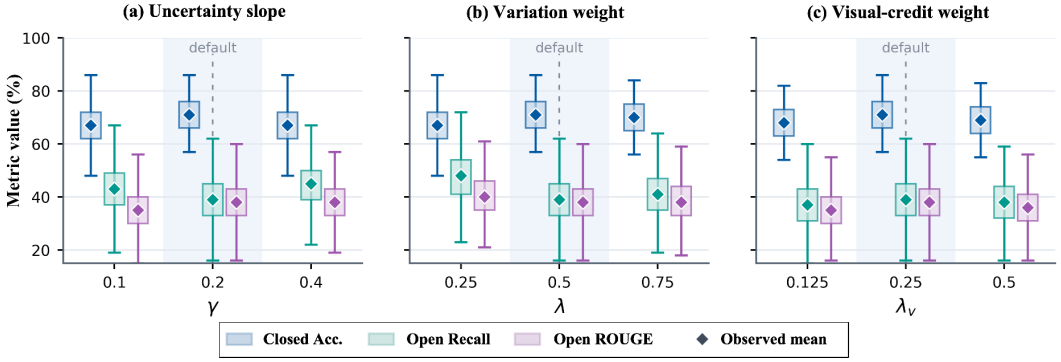


Figure 4: Hyperparameter sensitivity of TALC on SLAKE using Qwen2.5-VL-7B. Each panel varies one coefficient while holding the others at their default values, $\gamma = 0.2$, $\lambda = 0.5$, and $\lambda_v = 0.25$.

Hyperparameter sensitivity. Fig. 4 analyzes the sensitivity of TALC to the uncertainty slope γ, semantic variation weight λ, and visual-credit weight λ_v. Performance exhibits only limited fluctuations across the evaluated ranges, while the selected default values achieve strong or near-optimal overall performance. This suggests that TALC does not rely on narrowly tuned hyperparameters and maintains a stable balance among uncertainty-driven exploration, semantic variation, and visual consistency. We therefore use $\gamma = 0.2$, $\lambda = 0.5$, and $\lambda_v = 0.25$ throughout all reported experiments.

5 Conclusion

We presented TALC, a label-free test-time reinforcement learning framework for medical VQA that jointly improves visual exploration and reasoning trajectory selection. Look Closer acquires additional regional evidence through uncertainty-guided inspection, while Think Again reassesses trajectories using semantic support, reasoning novelty, and visual consistency. Experiments on four medical VQA benchmarks and two MLLM backbones consistently demonstrate their complementary benefits. These results suggest that reliable test-time adaptation for medical multimodal reasoning should not merely reinforce agreement among sampled answers, but should also improve the quality and diversity of the evidence and reasoning processes that produce the agreement.

AI USE STATEMENT

We used LLMs solely for language polishing, such as improving wording and checking for typographical and grammatical errors. We did not use generative AI tools for any research-related tasks, including hypothesis generation, methodology design, experimental implementation, data analysis, or result interpretation. All AI-assisted modifications were reviewed by the authors, and we take full responsibility for the final content of this work.

ETHICS STATEMENT

To the best of our knowledge, this work has no potential ethical issues to disclose.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of this work, all implementation details have been clearly described, and we have also released the code.

REFERENCES

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 107150–107180. Curran Associates, Inc., 2025. doi: 10.52202/085713-3573. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9a14cf2ee36cf2524f54d8f196a6734c-Paper-Conference.pdf.

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, 2025. URL https://arxiv.org/abs/2502.13923.

Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *Working Notes of CLEF 2019*, volume 2380 of *CEUR Workshop Proceedings*, Lugano, Switzerland, September 9-12 2019. CEUR-WS.org. URL https://ceur-ws.org/Vol-2380/paper_272.pdf.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Zhenyang Cai, Ke Ji, Xiang Wan, et al. Towards injecting medical visual knowledge into multimodal llms at scale. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 7346–7370, 2024.

Yixin Dong, Charlie F Ruan, Yaxing Cai, Ziyi Xu, Yilong Zhao, Ruihang Lai, and Tianqi Chen. Xgrammar: Flexible and efficient structured generation engine for large language models. *Proceedings of Machine Learning and Systems*, 7, 2025.

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. *arXiv preprint arXiv:2604.08203*, 2026.

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

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):180251, 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. *Advances in neural information processing systems*, 36:28541–28564, 2023.

Xiang Li, Huihui Fang, Changmiao Wang, Mingsi Liu, Lixin Duan, and Yanwu Xu. Cache-driven spatial test-time adaptation for cross-modality medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 146–156. Springer, 2024.

Xiang Li, Huihui Fang, Mingsi Liu, Jinghao Wang, Lixin Duan, Yuqi Fang, and Yanwu Xu. Beyond visual prompt: Unlocking parameter-efficient online adapter for continual test-time adaptation in medical image segmentation. *Information Fusion*, pp. 104256, 2026.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

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.

Siqi Ma, Jiajie Huang, Fan Zhang, Jinlin Wu, Yue Shen, Guohui Fan, Zhu Zhang, and Zelin Zang. Medla: A logic-driven multi-agent framework for complex medical reasoning with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 845–853, 2026.

Potsawee Manakul, Adian Liusie, and Mark Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 9004–9017, 2023.

Dung Nguyen, Minh Khoi Ho, Huy Ta, Thanh Tam Nguyen, Qi Chen, Kumar Rav, Quy Duong Dang, Satwik Ramchandre, Son Lam Phung, Zhibin Liao, et al. Localizing before answering: A benchmark for grounded medical visual question answering. In *Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)*. International Joint Conferences on Artificial Intelligence Organization, 2025.

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.

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

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Akshit Singh, Shyam Marjit, Wei Lin, Paul Gavrikov, Serena Yeung-Levy, Hilde Kuehne, Rogério Feris, Sivan Doveh, James Glass, and M Jehanzeb Mirza. Ttrv: Test-time reinforcement learning for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 33153–33163, 2026.

Anubhuti Sood, Tanurag Pattnaik, Rewa Malhotra, Vivek Nayyar, Bhaskar Narayan, Deepika Mishra, Varun Surya, et al. Medical imaging privacy: A systematic scoping review of key parameters in dataset construction and data protection. *Journal of Medical Imaging and Radiation Sciences*, 56(5):101914, 2025.

Marceli Wac, Raul Santos-Rodriguez, Chris McWilliams, and Christopher Bourdeaux. Capturing requirements for a data annotation tool for intensive care: Experimental user-centered design study. *JMIR Human Factors*, 12(1):e56880, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR 2023*, 2023. URL https://arxiv.org/abs/2203.11171.

Zhenhailong Wang, Xuehang Guo, Sofia Stoica, Haiyang Xu, Hongru Wang, Hyeonjeong Ha, Xiusi Chen, Yangyi Chen, Ming Yan, Fei Huang, et al. Perception-aware policy optimization for multimodal reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 80502–80530, 2026.

Zipei Wang, Mengjie Fang, Linglong Tang, Jie Tian, and Di Dong. Hicur-npc: Hierarchical feature fusion curriculum learning for multi-modal foundation model in nasopharyngeal carcinoma. *IEEE Transactions on Medical Imaging*, 44(10):3997–4009, 2025.

Lai Wei, Yuting Li, Chen Wang, Yue Wang, Linghe Kong, Weiran Huang, and Lichao Sun. First sft, second rl, third upt: Continual improving multi-modal llm reasoning via unsupervised post-training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 62293–62318. Curran Associates, Inc., 2025. doi: 10.52202/085713-2084. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/59ddfff7979b43f54690fa986c0e5138-Paper-Conference.pdf.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-r1: Advancing multimodal reasoning capabilities of mllms via visual perception reward. In *International Conference on Learning Representations*, volume 2026, pp. 26868–26898, 2026.

Dunyuan Xu, Xikai Yang, Yaoqian Li, Juzheng Miao, Jinpeng Li, and Pheng-Ann Heng. Perceive and calibrate: Analyzing and enhancing robustness of medical multi-modal large language models. *arXiv preprint arXiv:2512.21964*, 2025.

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.

Zhangyuan Yu, Wanran Sun, Guangjing Yang, Xiaohu Wu, and Qicheng Lao. Cfpo: Counterfactual policy optimization for multimodal reasoning. *arXiv preprint arXiv:2606.23206*, 2026.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, et al. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*, 2025.

Chu Zhao, Enneng Yang, Yuting Liu, Jianzhe Zhao, and Guibing Guo. Echo: Entropy-confidence hybrid optimization for test-time reinforcement learning. *arXiv preprint arXiv:2602.02150*, 2026a.

Xuandong Zhao, Zhewei Kang, Aosong Feng, Sergey Levine, and Dawn Song. Learning to reason without external rewards. In *International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=OU9nFEYR2M.

Weihai Zhi, Jiayan Guo, and Shangyang Li. Medgr2: Breaking the data barrier for medical reasoning via generative reward learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28901–28909, 2026.

Yujun Zhou, Zhenwen Liang, Haolin Liu, Wenhao Yu, Kishan Panaganti, Linfeng Song, Dian Yu, Xiangliang Zhang, Haitao Mi, and Dong Yu. Evolving language models without labels: Majority drives selection, novelty promotes variation. *arXiv preprint arXiv:2509.15194*, 2025.

Wenhui Zhu, Xuanzhao Dong, Xin Li, Peijie Qiu, Xiwen Chen, Abolfazl Razi, Aris Sotiras, Yi Su, and Yalin Wang. Toward effective reinforcement learning fine-tuning for medical vqa in vision-language models. *arXiv preprint arXiv:2505.13973*, 2025.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding. *arXiv preprint arXiv:2501.18362*, 2025a.

Yuxin Zuo, Kaiyan Zhang, Li Sheng, Shang Qu, Ganqu Cui, Xuekai Zhu, Haozhan Li, and yuchen zhang. Ttrl: Test-time reinforcement learning. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 131459–131483. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4376.

APPENDIX

This appendix provides further details and analyses supporting the main paper. We first carefully describe the medical VQA benchmarks, evaluation subsets, answer-processing rules, and metric definitions. We then present the compared methods and explain how their original learning signals are consistently integrated into the shared medical test-time adaptation workflow. Additional experiments systematically compare TALC with domain-specific medical vision-language models and examine whether correct answers are accessible within sampled rollout groups before they become the final consensus. Finally, we provide comprehensive theoretical analyses of the reward hierarchy, visual consensus, novelty-driven coverage, and uncertainty-guided branching, followed by qualitative examples illustrating TALC's structured region inspection and reasoning process.

A DATASETS AND EVALUATION METRICS

This section describes the four medical benchmarks and scoring rules used to evaluate TALC and the compared methods. The datasets span different imaging modalities, clinical settings, and answer formats, ranging from questions about medical images to expert multimodal reasoning tasks that may involve multiple figures. We first summarize each dataset and identify the evaluation subsets, including the numbers of closed-ended and open-ended questions used as metric denominators. We then describe answer extraction and normalization and define accuracy, token Recall, and ROUGE-1 F1 under a common question-level protocol. Together, these details provide a consistent and reproducible basis for evaluating the methods across datasets and answer types.

A.1 DATASETS

We evaluate TALC and the compared methods on SLAKE (Liu et al., 2021), VQA-RAD (Lau et al., 2018), VQA-Med 2019 (Ben Abacha et al., 2019), and MedXpertQA-MM (Zuo et al., 2025a). In the first three datasets, each question is paired with a single medical image, and the evaluation includes both closed-ended and open-ended answers. MedXpertQA-MM instead presents expert multiple-choice questions that may require information from several figures and the accompanying clinical context. Together, these benchmarks cover both conventional medical image question answering and more complex multimodal decisions. Table A.1 identifies the evaluation split for each dataset and reports its question counts, answer types, and number of images per question. Unless stated otherwise, dataset sizes in the descriptions below refer to the full original datasets rather than the subsets selected for evaluation.

SLAKE. SLAKE is a bilingual medical VQA dataset with semantic annotations created by physicians and an accompanying medical knowledge graph (Liu et al., 2021). The original dataset contains 642 images and 14,028 English and Chinese question–answer pairs. Its CT, MRI, and X-ray images cover the head, neck, chest, abdomen, and pelvis. Questions range from identifying visible organs, anatomical positions, and abnormalities to answering queries that require associated medical knowledge. SLAKE also provides segmentation masks and bounding boxes for image content, alongside annotations that distinguish visual questions from knowledge-based questions. The original split is made at the image level, keeping questions about the same image within one split. In our evaluation, we use the full test subset of 2,094 questions, comprising 836 closed-ended and 1,258 open-ended questions. We use the images and questions for adaptation but do not use the supplied masks, boxes, or knowledge graph as supervision.

VQA-RAD. VQA-RAD is a clinician-authored dataset of questions and reference answers about radiology images from MedPix (Lau et al., 2018). Its 315 images comprise 104 head CT or MRI scans, 107 chest radiographs, and 104 abdominal CT scans. The original dataset contains 3,515

Table A.1: Question counts and image multiplicity in the evaluated test subsets.

Dataset	Questions	Closed	Open	Images per question	Test subset
SLAKE	2,094	836	1,258	1	Full test
VQA-RAD	451	251	200	1	Official test
VQA-Med 2019	500	64	436	1	ImageCLEF 2019 test
MedXpertQA-MM	2,000	2,000	0	1–6	MM test
Total	5,045	3,151	1,894	–	–

questions, each associated with a single image. They address topics such as imaging modality, anatomical structures, abnormalities, location, and size. Clinicians wrote questions in natural language and also provided paraphrases, capturing different ways to ask about similar visual content. We use the 451-question test split and group its questions by answer type, as summarized in Table A.1. Under the grouping used in this study, 251 questions are closed-ended and typically call for a constrained answer such as yes or no and the remaining 200 are open-ended.

VQA-Med 2019. The ImageCLEF 2019 VQA-Med benchmark uses radiology images to ask about imaging modality, imaging plane, organ system, and abnormality (Ben Abacha et al., 2019). These four categories describe the subject of a question rather than whether its answer is closed-ended or open-ended. The training set contains 12,792 question and answer pairs over 3,200 images, while the validation set contains 2,000 pairs over 500 images. We use the official test set, which has 500 questions paired with 500 images. Its answers were reviewed by a medical doctor and a radiologist, and we retain accepted alternatives and optional answer components when constructing the reference answers. For our common scoring protocol, we classify a test question as closed-ended only when every accepted reference becomes either `yes` or `no` after normalization. All other test questions are treated as open-ended. This rule yields 64 closed-ended and 436 open-ended questions.

MedXpertQA-MM. MedXpertQA is an expert-level medical question answering benchmark containing 4,460 questions across 17 specialties and 11 body systems (Zuo et al., 2025a). It includes separate text and multimodal subsets. The multimodal subset brings together medical images, clinical histories, examination findings, and questions drawn from advanced medical examinations. Its visual material extends beyond radiology to pathology images, clinical photographs, diagrams, charts, and other documents. Answering a question may therefore require relating findings across several figures to information in the clinical description rather than identifying a feature in one image alone. We use the 2,000-question MedXpertQA-MM test split and exclude its five development examples. Each test question provides one to six images and five answer choices, labeled A through E. We treat these multiple-choice questions as closed-ended and score the selected option letter.

A.2 Evaluation Metrics

We use a common question-level scoring protocol across the four medical VQA datasets rather than reproducing each benchmark's original challenge metrics. Closed-ended and multiple-choice questions are evaluated by accuracy, while open-ended questions are evaluated by Recall and ROUGE-1 F1. MedXpertQA-MM contains only closed-ended questions, so only accuracy is reported for that dataset. Scores are computed for individual questions, averaged within the applicable question type, and presented as percentages in the main tables.

Answer extraction and normalization. For each generated response, the scorer first extracts the CONCLUSION field and removes common introductory phrases. If the field is absent, it searches for an answer, conclusion, or diagnosis pattern and then falls back to the last nonempty line. Format validity is recorded separately from answer quality. Let $\hat{y}_i$ denote the extracted prediction for question i, and let $\mathcal{R}_i$ be its nonempty set of accepted reference answers, including all accepted alternatives supplied by the dataset. The normalization function $n(s)$ lowercases text, replaces punctuation with spaces, and collapses whitespace. It does not perform stemming or semantic matching. An extracted prediction that is empty after normalization receives zero credit.

Question types and closed-ended accuracy. We follow the answer types available for each dataset. Because the released VQA-Med 2019 test set does not provide open-ended and closed-ended annotations, we classify a question as closed-ended only when every normalized reference answer is either "yes" or "no". All other VQA-Med questions are treated as open-ended. Let $\mathcal{D}_c$ be the resulting set of closed-ended questions. A prediction is counted as correct if it exactly matches any accepted reference after normalization:

$$\mathrm{Acc} = \frac{1}{|\mathcal{D}_c|} \sum_{i \in \mathcal{D}_c} \max_{y \in \mathcal{R}_i} \mathbf{1}\big[n(\hat{y}_i) = n(y)\big]. \tag{5}$$

For multiple-choice questions, the same rule compares the extracted option letter with the reference. Each question contributes at most one correct prediction, regardless of the number of references.

Open-ended Recall and ROUGE-1 F1. Valid open-ended answers may differ from the reference in wording, making normalized exact match too restrictive. We report both Recall and ROUGE-1 F1 because a single overlap measure cannot show both reference coverage and the effect of additional words. Recall measures the proportion of distinct reference tokens recovered by the prediction. Let $T(s)$ be the set of whitespace-delimited tokens in the normalized text $n(s)$. For question i, we take the highest Recall over its accepted references:

$$\mathrm{Rec}_i = \max_{y \in \mathcal{R}_i} \frac{|T(\hat{y}_i) \cap T(y)|}{|T(y)|}. \tag{6}$$

Using token sets means that repeating a matched word provides no additional credit. Recall also does not penalize words that appear only in the prediction, so a long response can receive a high score without being concise. ROUGE-1 F1 complements Recall by using token counts and balancing unigram precision with unigram recall (Lin, 2004). Let $c_s(w)$ count occurrences of token w in $n(s)$, let $L(s) = \sum_w c_s(w)$, and let $O(s, y) = \sum_w \min\{c_s(w), c_y(w)\}$ be the unigram overlap with reference y. The question-level score is

$$\mathrm{R1}_i = \max_{y \in \mathcal{R}_i} \frac{2O(\hat{y}_i, y)}{L(\hat{y}_i) + L(y)}. \tag{7}$$

The best reference is selected separately for each metric and may differ between Recall and ROUGE-1 F1. We average the two metrics separately over the open-ended question set $\mathcal{D}_o$, giving each question equal weight, and report the averages as percentages. Recall indicates whether the reference terms are covered, while ROUGE-1 F1 also reduces credit for additional or repeated words.

B Compared Methods

The compared methods differ in both the signal they use to guide improvement in medical VQA and the stage at which they apply it. EM-INF adjusts token logits locally during inference without updating underlying model weights. TTRL and TTRV update the policy using signals derived from the empirical distribution of answers generated for each input. Intuitor instead rewards the model's own certainty over generated tokens, while ECHO combines answer voting with entropy and confidence to guide adaptive exploration and subsequent policy updates. The following sections describe each method's learning signal, optimization objective, and procedure for generating trajectories, then explain how the methods are applied within the shared medical VQA workflow.

B.1 Baseline Definitions

EM-INF. EM-INF (Agarwal et al., 2025) adjusts the distribution over the next token during inference while keeping the model weights frozen. At each decoding position, it copies the backbone model's logits z_0 into a temporary vector z and optimizes only that vector. The original paper minimizes $\mathcal{L}_{\text{EM-INF}}(z) = \max\{\mathcal{H}(\mathrm{softmax}(z)), \delta\}$, where $\mathcal{H}$ is Shannon entropy and δ is a minimum entropy threshold. Above this threshold, minimizing the objective reduces the entropy of the token distribution. Below it, the objective is constant and no longer encourages further reduction. The authors' released implementation also includes a KL constraint. Our medical VQA baseline follows this implementation and minimizes $\mathcal{H}(p_z) + \beta_{\mathrm{KL}} D_{\mathrm{KL}}(p_0 \| p_z)$, where $p_0 = \mathrm{softmax}(z_0)$ and $p_z = \mathrm{softmax}(z)$. The KL term limits how far the adjusted distribution moves from the original one.

We use $\delta = 0.3$ and $\beta_{\mathrm{KL}} = 0.05$, with at most three optimization steps per position. Optimization stops early when the entropy falls below δ. A token is then sampled from the adjusted distribution and appended to the response. The model computes new logits from the updated context, and the procedure repeats at the next position. The temporary adjustments do not update model weights. EM-INF uses neither answer voting nor a reward model.

TTRL. TTRL (Zuo et al., 2025b) trains a model on unlabeled test inputs by turning agreement among its own sampled answers into a reward signal. For each input, it generates a group of trajectories $\{\tau_i\}_{i=1}^{K}$ and extracts a final answer a_i from each trajectory. The most frequent answer, $\hat{a} = \arg\max_a \sum_i \mathbf{1}[a_i = a]$, serves as the pseudo label. Each trajectory then receives the binary reward $R_i^{\mathrm{TTRL}} = \mathbf{1}[a_i = \hat{a}]$. Here, majority voting selects the most frequent answer even if it is supported by no more than half of the group. All trajectories with that answer receive the same reward, regardless of how they arrived at it, while the remaining trajectories receive zero reward. In the GRPO instantiation, these rewards are converted into group relative advantages for policy optimization. The pseudo label is estimated again from newly generated trajectories as the model adapts, allowing repeated sampling and policy updates without access to ground truth answers.

TTRV. TTRV (Singh et al., 2026) adapts vision and language models on unlabeled visual inputs by deriving rewards from the distribution of responses generated through repeated model sampling for each input. In answer-level notation, let $p(a) = K^{-1} \sum_j \mathbf{1}[a_j = a]$ be the empirical frequency of answer a, and let $p_i = p(a_i)$ be the frequency of trajectory i's answer. TTRV assigns each trajectory a reward $R_i^{\mathrm{TTRV}} = p_i - \beta_H H_A$, where $H_A = -\sum_{a:p(a)>0} p(a) \log p(a)$ is the Shannon entropy of the empirical answer distribution and β_H controls its contribution. The frequency term gives higher rewards to answers repeated more often while still assigning nonzero support to minority answers. The negative entropy term lowers the reward when answers are widely dispersed and is intended to encourage the model to converge toward more stable predictions. The resulting rewards are used for policy optimization with GRPO. However, the entropy term has a limitation under group relative optimization. Because H_A has the same value for every trajectory in a group, subtracting the group mean cancels its contribution to the relative advantages. It therefore does not distinguish trajectories within that group when advantages are computed solely by standard GRPO normalization. Moreover, the answer frequency measures the agreement among the predictions without checking whether the trajectories relied on consistent image regions.

Intuitor. Intuitor (Zhao et al., 2026b) trains a model using an intrinsic reward derived from its own predictive distributions rather than answer voting or external correctness feedback. Given an input x and a generated response $o = (o_1, \ldots, o_{|o|})$, it measures self-certainty by averaging the divergence between a uniform distribution U over the full vocabulary and the model's distribution at each generation position:

$$R^{\mathrm{SC}}(o) = \frac{1}{|o|} \sum_{t=1}^{|o|} D_{\mathrm{KL}}(U \,\|\, p_\theta(\cdot \mid x, o_{<t})) \,. \tag{8}$$

A larger score means that the predictive distributions depart more strongly from uniformity across the response. Dividing by $|o|$ makes this a per-position average rather than a sum that grows directly with response length. This divergence measures how strongly the model's full vocabulary distribution departs from uniformity at each generation position. For each input, Intuitor generates multiple responses and standardizes their self-certainty scores within the group to obtain GRPO advantages. It can therefore update the policy without labeled answers or an external correctness verifier. The original work studies this internal feedback in language reasoning tasks, including training on MATH. Unlike rewards based on agreement between final answers, self-certainty evaluates the model's predictive distribution throughout the generated response.

ECHO. ECHO (Zhao et al., 2026a) retains the binary reward obtained from majority voting while changing how candidate trajectories are generated and how the policy is updated. During rollout generation, it uses a tree structure in which branches can share an earlier reasoning prefix. At each potential branching point, the method considers both the entropy of the next-token distribution and confidence aggregated over the rollout group. It allocates more branches when uncertainty is high and confidence is low, but limits branching when confidence is already high to avoid concentrating

Table C.1: Performance Comparison of TALC-Adapted Qwen2.5-VL-7B Against Domain-Specific Medical Large Models.

	SLAKE			VQA-RAD			VQA-Med			MedXpertQA-MM
Method	Acc	Recall	ROUGE-1	Acc	Recall	ROUGE-1	Acc	Recall	ROUGE-1	Acc
Medical-Specific VLMs										
Med-R1-2B (Lai et al., 2026)	45.69	14.15	16.32	38.25	28.52	13.44	43.75	14.68	9.83	16.70
MedGemma-4B (Sellergren et al., 2025)	61.84	27.11	26.54	60.96	41.51	36.72	59.38	21.56	16.48	13.60
LLaVA-Med-7B (Li et al., 2023)	47.61	13.20	15.47	47.41	24.53	9.16	48.44	12.16	4.12	17.80
HuatuoGPT-V-7B (Chen et al., 2024)	57.54	22.42	23.68	57.37	38.58	32.19	56.25	20.18	15.34	19.65
Lingshu-7B (Xu et al., 2026)	65.31	30.92	30.17	64.14	43.51	38.63	62.50	23.33	17.85	22.65
General-Purpose VLMs										
Qwen2.5-VL-7B	58.13	19.56	24.05	49.80	43.35	20.68	62.50	24.10	15.39	23.80
Qwen2.5-VL-7B w/ TALC	**71.41**	**39.08**	**37.70**	**74.90**	**53.70**	**49.38**	**68.75**	**26.43**	**22.76**	**26.25**

the sampling budget on a few paths. Confidence is calculated from the mean probability of the top-k candidate tokens rather than directly from raw logits. ECHO also monitors smoothed confidence and changes in entropy as generation proceeds, pruning branches that remain unpromising instead of continuing to spend tokens on them. After complete responses are generated, it extracts their final answers, selects the most frequent answer as a pseudo-label, and assigns a reward of one to matching trajectories and zero to the others. These binary rewards produce group relative advantages for policy optimization. During the update, confidence near the end of each trajectory determines its clipping range, with higher confidence leading to a tighter range. Entropy and token confidence also shape the advantage at each position, changing how strongly different generation steps contribute to the update. Thus, voting determines which trajectories receive positive reward, while uncertainty and confidence guide the allocation of rollout budget and the strength of policy updates.

B.2 Integration into the Medical VQA Workflow

All methods are evaluated on the same medical images and questions, with answer choices included in the prompt when a dataset provides them. Each method uses the corresponding backbone setting and the same evaluation metrics described in Appendix A.2. For every dataset and random seed, methods that update model weights start from the unadapted backbone and process the test inputs in batches. They generate responses, derive their own training signals from those responses, and update the policy before proceeding to the next batch. Reference answers are excluded from reward construction, policy updates, and checkpoint selection. They are used only to calculate the reported evaluation metrics. The methods use this shared workflow in different ways. EM-INF adjusts temporary token logits during decoding and leaves the backbone weights unchanged. TTRL and TTRV derive rewards from the answers produced for each question, using majority voting and empirical answer frequencies, respectively. Intuitor instead computes a self-certainty score from the token distributions along each response and uses the resulting group relative advantages for policy optimization. The ECHO baseline uses entropy and confidence to guide tree branching and prune unpromising continuations during candidate generation. It retains majority voting as its reward signal and uses token uncertainty for advantage shaping and adaptive clipping during policy updates. Our method, TALC, instead introduces explicit region inspection and combines answer support, semantic novelty, and visual agreement in its reward. Each method is run separately, and none uses reference answers as an adaptation signal.

C Additional Experimental Results

Table C.1 shows that Qwen2.5-VL-7B with TALC achieves the best value for every reported metric. Compared with the strongest domain-specific result, TALC improves accuracy by 6.10 points on SLAKE, 10.76 points on VQA-RAD, 6.25 points on VQA-Med, and 3.60 points on MedXpertQA-MM. The accompanying Recall and ROUGE-1 gains show that the improvement extends to open-ended medical responses.

Correct-answer accessibility and consensus formation. Following the rollout-level evaluation protocol, for each closed-ended test question we independently sample $K = 8$ responses. Label Acc. takes the most frequent answer among the eight samples as the final prediction and measures the proportion of questions for which this majority-vote answer matches the reference answer. Pass@8 instead measures the proportion of questions for which at least one of the eight sampled responses matches the reference answer. Formally,

$$\text{LabelAcc} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbf{1}\left[\text{majority}\left(\hat{y}_1, \ldots, \hat{y}_8\right) = y^*\right],$$

$$\text{Pass@8} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathbf{1}\left[\exists i \in \{1, \ldots, 8\} : \hat{y}_i = y^*\right].$$

Label Acc. therefore characterizes whether sampled trajectories can form a correct final consensus, whereas Pass@8 reflects whether the correct answer is accessible anywhere within the rollout set.

Fig. C.1 further analyzes whether performance gains arise from discovering correct candidates or from converting them into a reliable consensus. Frequency Voting exhibits a noticeable gap between Pass@8 and Label Acc., indicating that a correct answer can already appear among the sampled trajectories but still fail to dominate the final vote.

TALC improves both metrics on SLAKE and VQA-RAD. On SLAKE, Pass@8 increases from 0.71 to 0.78, while Label Acc. increases from 0.58 to 0.72, suggesting that TALC substantially expands the coverage of correct candidate trajectories. On VQA-RAD, Pass@8 improves from 0.77 to 0.81 and Label Acc. from 0.57 to 0.62, showing that TALC improves both correct-answer exploration and its conversion into the final consensus. These results demonstrate that TALC not only improves final prediction performance but also increases the accessibility of correct alternative answers within the rollout set.

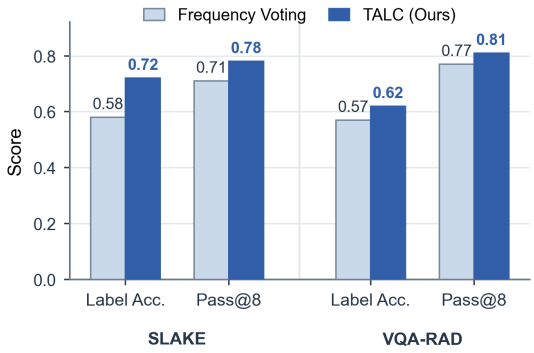


Figure C.1: Correct-answer accessibility and consensus formation on SLAKE and VQA-RAD using $K = 8$ rollouts. Label Acc. measures majority-vote correctness, whereas Pass@8 records whether any sampled response is correct.

D Theoretical Analysis

This section studies how TALC turns agreement among generated conclusions into a structured training signal. Semantic support identifies promising trajectories, while novelty distinguishes their reasoning paths and visual consensus compares the regions they inspect. Look Closer generates alternative inspections according to uncertainty within a finite rollout budget. We show that the chosen reward weights maintain a strict reward gap between trajectories above and below mean support, even after visual credit is added. Under independent localization noise, the error bound for the consensus mask decreases exponentially as more supported inspections contribute. A population surrogate shows when novelty preserves the best baseline trajectories while assigning positive probability to every distinct reasoning mode among them. Finally, the branching rule increases the probability of exploring uncertain region decisions while keeping every rollout within its capacity.

D.1 Reward Hierarchy Within a Rollout Group

Think Again assigns each trajectory a reward based on semantic support, reasoning novelty, and visual credit. We first examine how these signals interact within a group of trajectories generated for one input. Let $\mathcal{V}$ denote the valid trajectories, and let $c_i \in [0, 1]$ be the semantic support of trajectory i. As defined in Section 3.4, the mean support $\bar{c}$ partitions $\mathcal{V}$ into the high-support set $\mathcal{G}_{\text{high}}$ and the low-support set $\mathcal{G}_{\text{low}}$. Let E_i denote the banded variation reward and $C_i \in [0, 1]$ the visual credit. For $\lambda > 0$ and $\lambda_v \geq 0$, define the textual reward T_i and the full reward R_i by

$$T_i = c_i + \lambda E_i, \qquad R_i = T_i + \lambda_v C_i.$$

Proposition D.1 (Support priority under the textual reward). For every $h \in \mathcal{G}_{\text{high}}$ and $\ell \in \mathcal{G}_{\text{low}}$,

$$T_h - T_\ell > \lambda.$$

Thus, every high-support trajectory receives a higher textual reward than every low-support trajectory, regardless of their novelty scores.

Proof. By definition of the mean split, $c_h \geq \bar{c} > c_\ell$. The reward bands give $E_h \in [0.5, 1]$ and $E_\ell \in [-1, -0.5]$, so $E_h - E_\ell \geq 1$. Since $\lambda > 0$,

$$T_h - T_\ell = (c_h - c_\ell) + \lambda(E_h - E_\ell) > \lambda.$$

Corollary D.1 (Support priority under the full reward). If $\lambda \geq \lambda_v$, then for every $h \in \mathcal{G}_{\text{high}}$ and $\ell \in \mathcal{G}_{\text{low}}$,

$$R_h - R_\ell > \lambda - \lambda_v \geq 0.$$

Thus, adding visual credit preserves the ordering between the two support groups. With the reported values $\lambda = 0.5$ and $\lambda_v = 0.25$, the reward gap is strictly greater than 0.25.

Proof. Since $C_h, C_\ell \in [0, 1]$, we have $C_h - C_\ell \geq -1$. Combining this bound with Proposition D.1 yields

$$R_h - R_\ell = (T_h - T_\ell) + \lambda_v(C_h - C_\ell) > \lambda - \lambda_v.$$

This ordering is also preserved by group advantage standardization whenever the empirical reward variance is nonzero, because the same positive affine transformation is applied to every reward in the group. If all rewards are identical, the resulting advantages are zero.

Together, Proposition D.1 and Corollary D.1 show how Think Again combines a stable selection hierarchy with more detailed trajectory assessment. Semantic support separates the two reward strata, and the chosen weights keep this separation strict after visual credit is added. Within each stratum, reasoning novelty and spatial agreement can further distinguish trajectories. TALC can thus favor varied reasoning paths and consistent region inspections among supported candidates without changing which support stratum takes priority.

D.2 Concentration of the Visual Consensus Mask

Look Closer produces region inspections, and Think Again aggregates those from high-support trajectories with valid region requests into a consensus mask. We analyze when this aggregation provides a stable spatial reference for visual credit that captures regions repeatedly inspected across trajectories. Let $S \geq 1$ be the number of selected trajectories, and condition the following probabilities on this selected-set size. For pixel u, let $Z_u \in \{0, 1\}$ indicate membership in a latent spatial pattern shared by the selected trajectories. Let $X_{su} \in \{0, 1\}$ indicate whether trajectory s's inspection mask covers u. The noise assumptions in Proposition D.2 concern these masks after support-based selection. If $S = 0$, TALC forms no consensus mask and assigns zero visual credit.

Proposition D.2 (Concentration of the consensus mask). For a pixel u, assume that the selected masks $X_{1u}, \ldots, X_{Su}$ are independent conditional on Z_u. Suppose every selected trajectory has conditional error probability at most $\varepsilon < \frac{1}{2}$, so that $\Pr(X_{su} \neq z \mid Z_u = z) \leq \varepsilon$ for each s and each $z \in \{0, 1\}$ with positive probability. At the implemented threshold $\tau_c = \frac{1}{2}$, the consensus mask is

$$M_u^\star = \mathbf{1}\left[\frac{1}{S}\sum_{s=1}^{S} X_{su} > \frac{1}{2}\right].$$

A tie is assigned to zero. The pixelwise error probability satisfies

$$\Pr(M_u^\star \neq Z_u) \leq \exp\left[-2S\left(\frac{1}{2} - \varepsilon\right)^2\right].$$

If the same assumptions hold at each of P pixels, the probability of any mask error satisfies

$$\Pr\left(\exists u :\ M_u^\star \neq Z_u\right) \leq \min\left\{1,\ P\exp\left[-2S\left(\frac{1}{2} - \varepsilon\right)^2\right]\right\}.$$

The whole-mask bound does not require independence across pixels.

Proof. Fix $Z_u = z$ and define $Y_s = \mathbf{1}[X_{su} \neq z]$. Conditional on $Z_u = z$, these indicators are independent, lie in $[0, 1]$, and satisfy $\mathbb{E}[Y_s \mid Z_u = z] \leq \varepsilon$. If $z = 1$, a consensus error requires at least half of the trajectories to err. If $z = 0$, it requires more than half to err. Consequently,

$$\{M_u^\star \neq z\} \subseteq \left\{ \frac{1}{S} \sum_s Y_s \geq \tfrac{1}{2} \right\}.$$

Hoeffding's inequality gives

$$\Pr(M_u^\star \neq z \mid Z_u = z) \leq \exp\left[-2S \left(\tfrac{1}{2} - \varepsilon \right)^2 \right].$$

Averaging over Z_u proves the pixelwise bound. Applying a union bound over the P pixels and capping the result at one proves the whole-mask bound.

Proposition D.2 clarifies the role of visual aggregation across trajectories. When selected inspections provide conditionally independent observations of a shared spatial pattern and each observation has an error probability below one half, the pixelwise error bound of the majority mask decreases as more trajectories contribute. This result provides a statistical justification for using the consensus mask as a shared reference for visual credit assignment. Accordingly, the additional inspections generated by *Look Closer* can contribute not only new visual evidence, but also information about which regions are selected consistently across trajectories. Importantly, Z_u denotes an underlying shared spatial pattern in this analysis rather than an externally annotated lesion mask.

D.3 Support Preservation and Reasoning-Mode Coverage

Semantic support identifies trajectories that agree in their conclusions, but agreement alone does not distinguish repeated reasoning from genuinely different paths to a supported answer. We next analyze the role of novelty after baseline quality has been established. Let $\mathcal{Y}$ be a finite set of grammar-valid complete trajectories, and let $\Pi(\mathcal{Y})$ be the probability simplex over this set. Let $b : \mathcal{Y} \to \mathbb{R}$ represent a baseline support and grounding score before novelty is applied. Define its best-scoring set and the gap to the remaining trajectories by

$$b_\star = \max_{y \in \mathcal{Y}} b(y), \qquad \mathcal{G} = \{y : b(y) = b_\star\}, \qquad \Delta = b_\star - \max_{y \notin \mathcal{G}} b(y) > 0,$$

when $\mathcal{G} \neq \mathcal{Y}$. If $\mathcal{G} = \mathcal{Y}$, there are no lower-baseline trajectories to exclude. Let $d(y, y')$ be a symmetric reasoning dissimilarity with $d(y, y) = 0$ and $0 \leq d(y, y') \leq D_{\max}$. For a policy π and novelty weight $\eta > 0$, consider the population surrogate

$$J_\eta(\pi) = \sum_y \pi(y) b(y) + \eta \sum_{y, y'} \pi(y) \pi(y') d(y, y'). \tag{9}$$

The second term is the expected dissimilarity between two independent draws from π. It represents the pairwise diversity encouraged by TALC's novelty score. TALC's empirical reward includes normalization within each support stratum and a nearest-neighbor term. Accordingly, J_η is a surrogate for the design principle, not an exact population rewrite of the implemented reward.

Lemma D.1 (A small novelty weight preserves the supported set). Suppose $\mathcal{G} \neq \mathcal{Y}$. If $D_{\max} > 0$ and $0 < \eta < \Delta/(2D_{\max})$, every global maximizer of J_η over $\Pi(\mathcal{Y})$ is supported on $\mathcal{G}$. If $D_{\max} = 0$, the same conclusion holds for every $\eta > 0$.

Proof. Suppose a maximizer π assigns positive mass to $y_0 \notin \mathcal{G}$, and choose any $y_1 \in \mathcal{G}$. For sufficiently small $\delta > 0$, the distribution $\pi_\delta = \pi + \delta(\mathbf{e}_{y_1} - \mathbf{e}_{y_0})$ remains feasible. By symmetry of d, the derivative of the objective at $\delta = 0$ is

$$b(y_1) - b(y_0) + 2\eta \sum_{y'} \pi(y') \big[d(y_1, y') - d(y_0, y') \big] \geq \Delta - 2\eta D_{\max} > 0.$$

A sufficiently small transfer from y_0 to y_1 therefore increases the objective, contradicting optimality. Every maximizer must assign zero mass outside $\mathcal{G}$. If $D_{\max} = 0$, the diversity term vanishes and maximizing J_η reduces to maximizing the baseline score.

Lemma D.1 establishes the first part of the intended relationship between selection and variation. When novelty has sufficiently small weight relative to the baseline gap, it cannot make a lower-baseline trajectory preferable at the population optimum. We can therefore examine whether novelty has a useful effect among trajectories that already attain the best baseline score.

Assume that $\mathcal{G}$ is partitioned into $M \geq 2$ distinct reasoning modes, $\mathcal{G} = \bigsqcup_{m=1}^{M} \mathcal{G}_m$, with $N_m = |\mathcal{G}_m|$. Let $w_m = \sum_{y \in \mathcal{G}_m} \pi(y)$ be the policy mass assigned to mode m. In a similarity-gap model, suppose

$$d(y, y') = \begin{cases} 0, & y = y', \\ D_{\text{in}}, & y \neq y' \text{ in the same mode}, \\ D_{\text{out}}, & y, y' \text{ in different modes}, \end{cases} \qquad 0 \leq D_{\text{in}} < D_{\text{out}}.$$

Theorem D.1 (Novelty selects broad supported-mode coverage). Under the conditions of Lemma D.1 and the similarity-gap model, every maximizer of J_η assigns positive probability to every reasoning mode. The mode masses are unique and satisfy

$$w_m^\star = \frac{a_m^{-1}}{\sum_{j=1}^{M} a_j^{-1}}, \qquad a_m = (D_{\text{out}} - D_{\text{in}}) + \frac{D_{\text{in}}}{N_m}.$$

If $D_{\text{in}} > 0$, the maximizing policy is also uniquely uniform within each mode, with $\pi^\star(y) = w_m^\star / N_m$ for $y \in \mathcal{G}_m$. Equal-sized modes then yield the uniform policy over $\mathcal{G}$, whose entropy is $\log |\mathcal{G}|$. If $D_{\text{in}} = 0$, the unique mode masses are $w_m^\star = 1/M$, while the distribution within each mode is unrestricted.

Proof. By Lemma D.1, an optimizer is supported on $\mathcal{G}$, where the baseline score equals $b_\star$. Separating pairs from the same mode and different modes gives

$$\sum_{y, y' \in \mathcal{G}} \pi(y)\pi(y')d(y, y') = D_{\text{out}} + (D_{\text{in}} - D_{\text{out}}) \sum_{m=1}^{M} w_m^2 - D_{\text{in}} \sum_{y \in \mathcal{G}} \pi(y)^2.$$

Maximizing this quantity is equivalent to minimizing

$$F(\pi) = (D_{\text{out}} - D_{\text{in}}) \sum_{m} w_m^2 + D_{\text{in}} \sum_{y \in \mathcal{G}} \pi(y)^2.$$

For fixed mode masses, the within-mode sum of squares is minimized by assigning $\pi(y) = w_m / N_m$ to every $y \in \mathcal{G}_m$. This within-mode minimizer is unique when $D_{\text{in}} > 0$. Substitution gives $F = \sum_m a_m w_m^2$. Since every $a_m > 0$, this is a strictly convex quadratic over the mode-mass simplex. Its unique minimizer has the stated masses $w_m^\star$, and every mode receives positive probability. When $D_{\text{in}} > 0$ and all modes have the same size, all a_m are equal. The policy is then uniform over $\mathcal{G}$ and has entropy $\log |\mathcal{G}|$. When $D_{\text{in}} = 0$, the mode masses remain uniquely uniform, but F does not depend on the allocation within each mode.

The lemma and theorem clarify the complementary roles of selection and variation underlying TALC's novelty reward. With a sufficiently small novelty weight, the population surrogate assigns no probability to trajectories below the optimal baseline score. Among the remaining trajectories, it promotes coverage of distinct reasoning modes rather than concentration on a single repeated pattern. Novelty therefore serves a specific role after support-based selection: encouraging diverse yet well-supported lines of reasoning. Because the implemented reward is not identical to this quadratic surrogate, Theorem D.1 characterizes the intended diversity pressure, while Proposition D.1 gives the exact support-ordering result for realized rollout groups.

D.4 Budgeted Uncertainty-Driven Branching

Look Closer uses uncertainty to decide where an additional regional inspection may be valuable, but it must operate within a finite rollout budget. Index potential branch decisions by $t = 1, \ldots, T_{\max}$, where $T_{\max} < \infty$ because the per-trajectory turn limit and rollout capacity are finite. Let E_t indicate that decision t is reached and eligible. Let $H_{0,t}$ be the initial-region entropy for the trajectory containing that decision, and define:

$$\Delta H_t = H_t - H_{0,t}, \qquad q_t = \text{clip}(q_0 + \gamma \Delta H_t, 0, 1).$$

Let $\mathcal{F}_{t-1}$ contain the history available before the branch draw. Assume that E_t and q_t are measurable with respect to $\mathcal{F}_{t-1}$, and let Z_t indicate that decision t creates a branch.

Proposition D.3 (Monotone branch probabilities and budget compliance). For $\gamma \geq 0$, the branch probability q_t is nondecreasing in relative uncertainty ΔH_t. If two eligible decisions lie inside the unclipped interval, then $q_s - q_t = \gamma(\Delta H_s - \Delta H_t)$. If they have the same reference entropy, this becomes $q_s - q_t = \gamma(H_s - H_t)$. For branch capacity B, the number of created branches satisfies

$$\sum_t Z_t \leq B \quad \text{almost surely,} \qquad \mathbb{E}\left[\sum_t Z_t\right] \leq \min\left(B, \sum_t \mathbb{E}[E_t q_t]\right).$$

If eligibility and branch probabilities are fixed deterministically across the potential decisions, the second term becomes $\sum_t E_t q_t$.

Proof. The clipping function is nondecreasing, which proves the probability statements. Let A_t indicate that branch capacity remains before decision t. Under the Bernoulli branch draw,

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = E_t A_t q_t \leq E_t q_t.$$

The tower property and linearity of expectation yield $\mathbb{E}[\sum_t Z_t] \leq \sum_t \mathbb{E}[E_t q_t]$. The scheduler enforces $\sum_t Z_t \leq B$ on every rollout. Combining expected and pathwise bounds proves the claim.

Proposition D.3 explains how Look Closer allocates opportunities for additional inspection. A region decision with greater relative uncertainty has a weakly higher probability of producing an alternative branch, while the scheduler keeps the number of generated trajectories within the rollout capacity. In the reported setting with a total budget of $K = 4$ trajectories and $I = 2$ existing trajectories, branches and fresh-root refills share the two remaining trajectory slots. This rule gives uncertain eligible decisions more opportunities to produce alternative visual evidence, then leaves the completed trajectories to be assessed by Think Again.

Together, these results clarify how TALC coordinates selection and exploration. Semantic support determines the reward hierarchy, reasoning novelty promotes variation within that hierarchy, and visual consensus turns repeated region inspections into a shared comparison signal. Uncertainty guides opportunities for additional inspections without exceeding the rollout budget. Look Closer generates alternative visual evidence, while Think Again determines which resulting trajectories should contribute most strongly to the policy update.

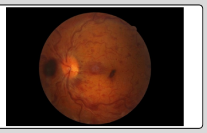
E Limitations and Future Work

A central limitation of answer-level consensus is that trajectories sampled from the same model share semantic and language priors, allowing an incorrect conclusion to be repeatedly reproduced and reinforced. Visual consistency provides an additional source of evidence that is less directly governed by such priors, but it does not eliminate the possibility of correlated errors: multiple trajectories may still converge on the same incorrect region. TALC partially mitigates this risk through its reward construction, such as conditioning visual credit on semantic support and integrating it with reasoning novelty in the final reward. Nevertheless, shared localization biases may still recur across trajectories motivating us to further investigate finer-grained mechanisms for identifying and correcting such correlated visual failures.

F Qualitative Visualizations

Figures in this section visualize the structured TALC interface on real benchmark inputs. Each case presents the original image and question, the executed regional inspections and their returned crops, and a compact structured final response. Cases C01–C07 are from MedXpertQA-MM, C08–C14 from SLAKE, C15–C21 from VQA-RAD, and C22–C28 from VQA-Med 2019.

Case C01 **Venous engorgement and retinal hemorrhages**

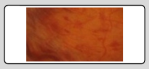


Question. A 76-year-old male with a medical history of hypertension, hyperlipidemia, and poorly managed type 2 diabetes arrives at the emergency department. He reports experiencing sudden vision loss in his left eye that occurred while watching television approximately 90 minutes ago. The patient denies experiencing eye pain, headache, or any other symptoms. Upon examination, a dilated retinal exam is performed, and an image is provided. Based on the retinal examination findings shown in the figure, which of the following is the most probable diagnosis?
Options. (A) Central retinal-artery occlusion (B) Macular degeneration (C) Arteritic ischemic optic neuropathy (D) Central retinal-vein occlusion (E) Severe diabetic retinopathy

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [295, 405, 435, 585], "label": "optic disc and peripapillary veins"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [442, 405, 653, 585] (optic disc and peripapillary veins).
</tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [365, 625, 565, 790], "label": "inferior retinal arcade"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [547, 625, 847, 790] (inferior retinal arcade). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [560, 435, 700, 585], "label": "posterior retinal hemorrhages"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [839, 435, 1050, 585] (posterior retinal hemorrhages).
</tool_response>



<SUMMARY> Identify the retinal pattern accompanying acute painless visual loss. </SUMMARY>
<CAPTION> The disc-adjacent veins are prominent and tortuous; the inferior arcade shows venous tortuosity and adjacent dark hemorrhages; a separate posterior retinal field contains dark hemorrhagic spots. </CAPTION>
<REASONING> Engorged tortuous veins accompany retinal hemorrhages in the inferior field. In the acute painless presentation, the combined vascular pattern favors central retinal-vein occlusion. </REASONING>
<CONCLUSION> D </CONCLUSION>

Case C02 **Creamy vessels in separate retinal arcades**



Question. Based on the image, which of the following conditions is the correct diagnosis?

Options. (A) Retinitis pigmentosa (B) Cytomegalovirus retinitis (C) Central retinal vein occlusion (D) Lipemia retinalis (E) Diabetic retinopathy

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [140, 265, 300, 430], "label": "optic disc vessels"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [210, 347, 450, 565] (optic disc vessels). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [380, 170, 585, 335], "label": "superior retinal arcade"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [570, 223, 878, 440] (superior retinal arcade). </tool_response>



<SUMMARY> Assess whether vessel pallor is a diffuse retinal finding. </SUMMARY>
<CAPTION> Pale vessels emerge around the optic disc; the superior arcade contains conspicuously creamy vessels. </CAPTION>
<REASONING> Pale vessels in the superior arcade reinforce the disc-region finding. This diffuse vascular discoloration favors lipemia retinalis.
</REASONING>
<CONCLUSION> D </CONCLUSION>

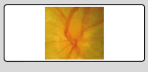
Case C03 **Cotton-wool lesions and a separate hemorrhage**



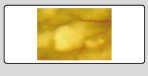
Question. Based on the image, which of the following conditions best explains the observed pathology?

Options. (A) Cytomegalovirus retinitis (B) Glaucoma (C) Retinal vein thrombosis
(D) Panretinal photocoagulation (E) Malignant hypertension

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [675, 285, 850, 490], "label": "optic disc and adjacent vessels"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [1012, 324, 1275, 558] (optic disc and adjacent vessels).
</tool_response>



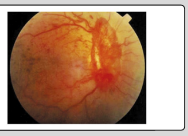
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [365, 130, 555, 305], "label": "superior cotton-wool spots"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [547, 147, 833, 348] (superior cotton-wool spots). </tool_response>



<SUMMARY> Compare white retinal lesions with hemorrhagic changes. </SUMMARY>
<CAPTION> The disc region and surrounding vessels appear abnormal; Fluffy white retinal patches are visible superiorly. </CAPTION>
<REASONING> The numerous fluffy white lesions, together with the abnormal disc-region appearance, favor severe hypertensive retinopathy among the options. </REASONING>
<CONCLUSION> E </CONCLUSION>

Figure F.1: MedXpertQA-MM examples of central retinal-vein occlusion (C01), lipemia retinalis (C02), and malignant hypertension (C03).

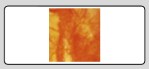
Case C04 **Proliferative retinal vascular changes**



Question. Which condition is demonstrated?

Options. (A) Papilledema (B) Profilerative diabetic retinopathy (C) Chorioretinitis (D) Hypertensive retinopathy (E) Central retinal vein occlusion

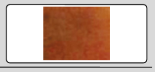
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [570, 185, 760, 425], "label": "peripapillary neovascular tissue"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [855, 226, 1140, 522] (peripapillary neovascular tissue).
</tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [610, 445, 790, 655], "label": "inferior vascular-hemorrhagic complex"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [915, 545, 1185, 804] (inferior vascular-hemorrhagic complex).
</tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [230, 390, 435, 590], "label": "peripheral retinal hemorrhages"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [345, 478, 653, 724] (peripheral retinal hemorrhages).
</tool_response>



<SUMMARY> Inspect the vascular proliferation and accompanying retinal bleeding. </SUMMARY>
<CAPTION> Irregular fine vessels and pale tissue lie near the disc region; a prominent red vascular and hemorrhagic complex extends inferiorly; a separate retinal region shows scattered dark hemorrhagic changes. </CAPTION>
<REASONING> Fine irregular vascular tissue accompanies the hemorrhagic complex. This combination is compatible with proliferative diabetic retinopathy.
</REASONING>
<CONCLUSION> B </CONCLUSION>

Case C05 **A focal arterial plaque and downstream whitening**



Question. Which of the following conditions best explains the observed findings?

Options. (A) Cholesterol embolism (B) Toxoplasmosis (C) Temporal arteritis (D) Central retinal vein occlusion (E) Retinoschisis

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [280, 300, 455, 385], "label": "retinal arterial plaque"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [420, 342, 683, 440] (retinal arterial plaque). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [235, 180, 435, 300], "label": "superior sectoral retinal pallor"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [352, 205, 653, 343] (superior sectoral retinal pallor).
</tool_response>

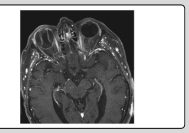


<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [505, 430, 745, 620], "label": "macular-side retinal pallor"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [757, 490, 1118, 708] (macular-side retinal pallor).
</tool_response>



<SUMMARY> Relate a focal arterial finding to the distribution of retinal pallor. </SUMMARY>
<CAPTION> A small bright focus lies along a retinal arterial branch; Pale retina occupies the superior part of the affected arterial territory; a second pale region extends toward the macula along the same territory. </CAPTION>
<REASONING> The focal arterial plaque and downstream sectoral pallor favor an embolic arterial process; cholesterol embolism is the matching option.
</REASONING>
<CONCLUSION> A </CONCLUSION>

Case C06 **Detached retinal membrane on orbital MRI**



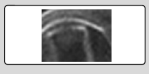
Question. What is the most likely diagnosis?

Options. (A) Graves" ophthalmopathy (B) Angle closure glaucoma (C) Ocular implant (D) Choroidal melanoma (E) Retinal detachment

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [180, 190, 305, 355], "label": "intraocular detached membrane"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [270, 283, 458, 531] (intraocular detached membrane).
</tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [175, 155, 295, 245], "label": "superior retinal membrane"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [262, 231, 443, 367] (superior retinal membrane). </tool_response>



<SUMMARY> Inspect the intraocular membrane and compare the opposite globe. </SUMMARY>
<CAPTION> A thin curved membrane is visible inside the image-left globe; the superior portion of the globe contains the abnormal internal membrane.
</CAPTION>
<REASONING> The curved intraocular membrane and its continuation within the globe support retinal detachment among the listed diagnoses.
</REASONING>
<CONCLUSION> E </CONCLUSION>

Figure F.2: MedXpertQA-MM examples of proliferative diabetic retinopathy (C04), cholesterol embolism (C05), and retinal detachment (C06).

Case C07 **Retinal vascular tumor and feeder vessels**



Question. An 18-year-old female presents with visual loss in her right eye. Based on the ophthalmoscopic findings, what is the most likely diagnosis?

Options. (A) Cystic fibrosis (B) Sickle cell disease (C) Marfan syndrome (D) Tay-Sachs disease (E) Von Hippel-Lindau disease

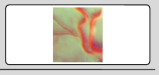
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [440, 15, 680, 205], "label": "peripheral retinal vascular lesion"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [660, 22, 1020, 308] (peripheral retinal vascular lesion). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [495, 275, 660, 485], "label": "tortuous feeder vessels"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [742, 411, 990, 727] (tortuous feeder vessels). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [500, 545, 675, 740], "label": "disc-directed vascular channels"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [750, 816, 1013, 1109] (disc-directed vascular channels). </tool_response>



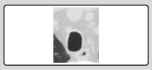
<SUMMARY> Connect the peripheral retinal lesion to its abnormal feeding vessels. </SUMMARY>
<CAPTION> A raised yellow-orange peripheral retinal lesion is visible; Prominent tortuous vessels travel toward the lesion; the enlarged vascular channels can be followed toward the disc. </CAPTION>
<REASONING> A retinal vascular tumor with prominent feeder vessels is compatible with a hemangioblastoma, making von Hippel-Lindau disease the best listed association. </REASONING>
<CONCLUSION> E </CONCLUSION>

Case C08 **Paired lung fields on chest CT**



Question. What is the largest organ in the picture?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [440, 485, 545, 615], "label": "trachea and central thorax"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [225, 248, 280, 315] (trachea and central thorax). </tool_response>



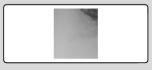
<SUMMARY> Identify the largest visible organ in this chest slice. </SUMMARY>
<CAPTION> The central air-filled trachea establishes the thoracic setting. </CAPTION>
<REASONING> The inspected central air-filled trachea establishes the thoracic setting. In the original image, the paired aerated lung fields occupy the largest visible organ area. </REASONING>
<CONCLUSION> Lung </CONCLUSION>

Case C09 **Cardiac silhouette on chest radiography**



Question. What diseases are included in the picture?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [455, 500, 705, 785], "label": "central cardiac silhouette"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [465, 512, 722, 804] (central cardiac silhouette). </tool_response>



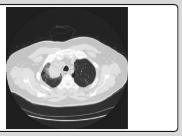
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [340, 490, 460, 710], "label": "image-left cardiac margin"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [348, 501, 472, 728] (image-left cardiac margin). </tool_response>



<SUMMARY> Inspect both margins of the cardiac silhouette in the full chest context. </SUMMARY>
<CAPTION> The central cardiac silhouette is broad across the lower chest; the image-left cardiac margin contributes to the broad silhouette. </CAPTION>
<REASONING> The full chest radiograph shows a broad cardiac silhouette. The local border crop helps trace its transverse extent, supporting cardiomegaly. </REASONING>
<CONCLUSION> Cardiomegaly </CONCLUSION>

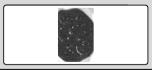
Figure F.3: A MedXpertQA-MM example of von Hippel–Lindau disease indicated by a retinal vascular tumor (C07) and SLAKE examples identifying the lung as the largest visible organ (C08) and cardiomegaly (C09).

Case C10 **Lung visibility beside a focal opacity**



Question. Does the picture contain lung?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [570, 420, 720, 650], "label": "aerated image-right lung"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [291, 215, 369, 333] (aerated image-right lung). </tool_response>



<SUMMARY> Determine whether any lung is present despite the focal dense abnormality. </SUMMARY>
<CAPTION> Aerated pulmonary tissue is clearly visible on the image-right side. </CAPTION>
<REASONING> The inspected image-right field contains aerated pulmonary tissue with vascular markings. This visible lung tissue establishes that lung is present. </REASONING>
<CONCLUSION> Yes </CONCLUSION>

Case C11 **Counting the two kidneys independently**



Question. How many kidneys in this image?

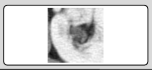
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [250, 435, 380, 610], "label": "image-left kidney"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [128, 222, 195, 313] (image-left kidney). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [620, 440, 770, 645], "label": "image-right kidney"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [317, 225, 395, 331] (image-right kidney). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [620, 500, 745, 625], "label": "image-right renal contour and sinus"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [317, 256, 382, 320] (image-right renal contour and sinus).
</tool_response>



<SUMMARY> Count distinct kidneys using their paired retroperitoneal positions. </SUMMARY>
<CAPTION> An enhancing kidney lies posteriorly on the image-left side; the image-right kidney enhances and has a visible outer contour; a tighter view shows the image-right renal contour and central sinus. </CAPTION>
<REASONING> The first crop identifies the image-left kidney, while the next two views show the image-right kidney at different scales. These are two distinct organs. </REASONING>
<CONCLUSION> 2 </CONCLUSION>

Case C12 **Small lung-base slivers on abdominal CT**



Question. Does the picture contain lung?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [345, 480, 420, 595], "label": "posterior lung-base aeration"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [176, 245, 216, 305] (posterior lung-base aeration).
</tool_response>



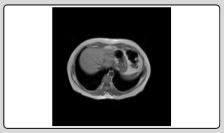
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [160, 595, 355, 700], "label": "image-left basal lung crescent"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [81, 304, 182, 359] (image-left basal lung crescent).
</tool_response>



<SUMMARY> Check whether small lung-base regions are present in an abdominal slice. </SUMMARY>
<CAPTION> A narrow dark aerated region lies posterior to the upper abdominal organ contour; a thin crescent of lung remains at the image-left posterior base. </CAPTION>
<REASONING> The posterior image-left crescent has an aerated lung appearance, confirming that lung is included. </REASONING>
<CONCLUSION> Yes </CONCLUSION>

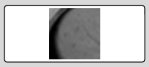
Figure F.4: SLAKE examples of lung presence beside a focal opacity (C10), kidney counting (C11), and basal lung tissue on abdominal CT (C12).

Case C13 **Liver presence on abdominal MRI**



Question. Does the picture contain liver?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [205, 350, 360, 505], "label": "right hepatic parenchyma"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [52, 89, 93, 130] (right hepatic parenchyma). </tool_response>

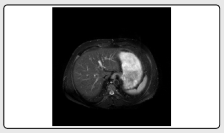


<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [365, 355, 480, 470], "label": "medial liver and vessels"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [93, 90, 123, 121] (medial liver and vessels). </tool_response>



<SUMMARY> Verify the liver from its location and continuity across local regions. </SUMMARY>
<CAPTION> Homogeneous parenchyma occupies the expected right-upper-abdominal liver position; the organ continues medially with internal vascular contours. </CAPTION>
<REASONING> The first parenchymal crop and its medial continuation correspond to the liver in the right upper abdomen. The organ is present. </REASONING>
<CONCLUSION> Yes </CONCLUSION>

Case C14 **Fluid-sensitive abdominal MRI**



Question. What is the mr weighting in this image?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [665, 395, 760, 550], "label": "bright gastric fluid"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [212, 126, 244, 176] (bright gastric fluid). </tool_response>

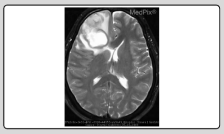


<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [670, 555, 765, 640], "label": "inferior gastric contents"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [214, 177, 245, 205] (inferior gastric contents). </tool_response>



<SUMMARY> Use fluid signal to identify the MR weighting. </SUMMARY>
<CAPTION> Gastric fluid is markedly bright; a separate inferior portion of the gastric contents also remains bright. </CAPTION>
<REASONING> Bright fluid at two gastric locations, relative to darker surrounding soft tissue, supports a T2-weighted image. </REASONING>
<CONCLUSION> T2 </CONCLUSION>

Case C15 **Ventricular horns rather than a peripheral lesion**



Question. what bright white structures form an x?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [435, 310, 555, 470], "label": "central ventricular CSF spaces"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [148, 132, 190, 202] (central ventricular CSF spaces).
</tool_response>



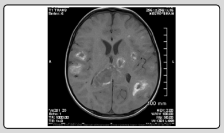
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [250, 545, 410, 640], "label": "image-left posterior ventricular horn"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [85, 233, 140, 274] (image-left posterior ventricular horn).
</tool_response>



<SUMMARY> Identify the central bright structures described as forming an X. </SUMMARY>
<CAPTION> Bright CSF spaces lie near the midline; a bright posterior horn extends on the image-left side. </CAPTION>
<REASONING> The central CSF spaces and image-left posterior horn have the connected configuration of the lateral ventricles. </REASONING>
<CONCLUSION> lateral ventricles </CONCLUSION>

Figure F.5: SLAKE examples of liver presence (C13) and T2-weighting recognition (C14), followed by VQA-RAD lateral-ventricle identification (C15).

Case C16 **Visibility of dark ventricular spaces**

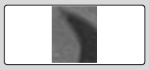


Question. are the lateral ventricles visible?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [465, 365, 520, 475], "label": "midline CSF channel"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [278, 202, 312, 264] (midline CSF channel). </tool_response>

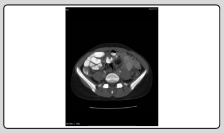


<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [400, 260, 485, 375], "label": "anterior ventricular horn"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [239, 144, 291, 209] (anterior ventricular horn). </tool_response>



<SUMMARY> Check whether the lateral ventricles are visible despite surrounding lesions. </SUMMARY>
<CAPTION> A narrow central CSF channel provides the midline landmark; One dark anterior ventricular horn is visible. </CAPTION>
<REASONING> The visible anterior horn connected to the midline CSF system confirms ventricular visibility. </REASONING>
<CONCLUSION> yes </CONCLUSION>

Case C17 **Dilated small-bowel loops**



Question. is there evidence of small bowel obstruction on this image?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [190, 415, 345, 490], "label": "anterior dilated bowel loop"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [194, 544, 354, 643] (anterior dilated bowel loop). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [355, 360, 465, 450], "label": "enlarged bowel loop with contents"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [363, 471, 477, 590] (enlarged bowel loop with contents).
</tool_response>



<SUMMARY> Inspect the bowel-loop configuration for evidence of obstruction. </SUMMARY>
<CAPTION> A contrast-containing dilated bowel loop is visible anteriorly; Another enlarged loop contains intraluminal material. </CAPTION>
<REASONING> Several enlarged bowel loops are visible in the original image and the second crop. The combined gas- and contrast-containing loop pattern supports small-bowel obstruction. </REASONING>
<CONCLUSION> yes </CONCLUSION>

Case C18 **Local left-lung airspace opacity**



Question. is there airspace consolidation on the left side?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [575, 280, 695, 430], "label": "patient-left upper-mid lung opacity"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [276, 140, 334, 217] (patient-left upper-mid lung opacity).
</tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [645, 430, 770, 550], "label": "adjacent left-lung airspace opacity"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [309, 216, 370, 277] (adjacent left-lung airspace opacity).
</tool_response>



<SUMMARY> Assess the patient-left lung and compare the opposite side. </SUMMARY>
<CAPTION> Patchy opacity is present in the image-right, patient-left upper-mid lung; the opacity extends into a neighboring left-lung region. </CAPTION>
<REASONING> The neighboring patient-left lung crop shows further airspace opacity. Its extent supports left-sided consolidation. </REASONING>
<CONCLUSION> yes </CONCLUSION>

Figure F.6: VQA-RAD examples of ventricular visibility (C16), small-bowel obstruction (C17), and left-sided airspace consolidation (C18).

Case C19 **Bowel anterior to the right kidney**



Question. what are these opacities anterior to the right kidney?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [370, 500, 440, 650], "label": "patient-right kidney"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [378, 278, 451, 363] (patient-right kidney). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [370, 315, 450, 435], "label": "contrast-filled bowel anterior to kidney"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [378, 175, 461, 243] (contrast-filled bowel anterior to kidney).
</tool_response>



<SUMMARY> Identify the opacities using their position relative to the right kidney. </SUMMARY>
<CAPTION> The patient-right kidney supplies the posterior anatomical landmark; Contrast-filled bowel lies anterior to that kidney. </CAPTION>
<REASONING> The opacity is a luminal bowel structure anterior to the kidney, consistent with the small intestines. </REASONING>
<CONCLUSION> the small intestines </CONCLUSION>

Case C20 **Which cardiac border is obscured?**



Question. which side of the heart border is obscured?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [390, 405, 490, 565], "label": "image-left cardiac margin"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [338, 396, 425, 554] (image-left cardiac margin). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [285, 380, 390, 515], "label": "patient-right lung opacity at cardiac interface"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [247, 372, 339, 505] (patient-right lung opacity at cardiac interface). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [655, 425, 745, 570], "label": "opposite cardiac margin"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [567, 416, 646, 559] (opposite cardiac margin). </tool_response>



<SUMMARY> Determine which cardiac margin loses its normal sharp interface. </SUMMARY>
<CAPTION> The image-left cardiac margin blends with adjacent opacity; Neighboring patient-right lung opacity reaches the cardiac interface; the opposite cardiac margin is more clearly delineated. </CAPTION>
<REASONING> Opacity abuts the image-left cardiac interface, corresponding to the patient-right heart border. </REASONING>
<CONCLUSION> right </CONCLUSION>

Case C21 **Air-filled bronchus within a dense right-lung region**

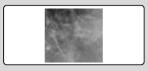


Question. what is this lucency amid densities in the right lung?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [340, 405, 440, 505], "label": "right hilar branching lucency"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [326, 486, 423, 606] (right hilar branching lucency).
</tool_response>



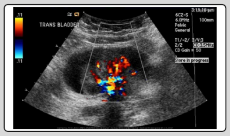
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [340, 345, 420, 405], "label": "central bronchial continuation"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [326, 414, 404, 486] (central bronchial continuation).
</tool_response>



<SUMMARY> Identify the branching lucency within the patient-right lung density. </SUMMARY>
<CAPTION> A dark branching lucency lies within the dense right hilar region; the lucency continues toward the more central bronchial region. </CAPTION>
<REASONING> The dark hilar lumen can be followed toward the central bronchial region. This air-containing configuration is consistent with the requested right bronchus. </REASONING>
<CONCLUSION> the right bronchus </CONCLUSION>

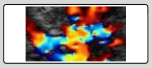
Figure F.7: VQA-RAD examples identifying bowel anterior to the right kidney (C19), right heart-border obscuration (C20), and the right bronchus (C21).

Case C22 **Color-flow signal on a sonographic background**



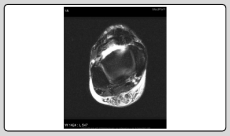
Question. what imaging method was used?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [400, 520, 560, 665], "label": "color-flow signal"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [409, 317, 574, 407] (color-flow signal). </tool_response>



<SUMMARY> Identify the imaging method from the flow overlay and background. </SUMMARY>
<CAPTION> A localized red-blue color-flow signal overlays grayscale anatomy. </CAPTION>
<REASONING> The color-flow signal on ultrasound identifies Doppler ultrasound. </REASONING>
<CONCLUSION> Doppler ultrasound </CONCLUSION>

Case C23 **Bright joint fluid on musculoskeletal MRI**



Question. what is the mr weighting in this image?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [300, 280, 535, 365], "label": "bright joint recess fluid"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [229, 254, 410, 333] (bright joint recess fluid). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [250, 430, 375, 570], "label": "adjacent cortical bone and muscle"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [191, 391, 288, 519] (adjacent cortical bone and muscle).
</tool_response>



<SUMMARY> Use relative fluid and tissue signals to determine MR weighting. </SUMMARY>
<CAPTION> Fluid in the joint recess is markedly bright; Adjacent cortical bone and muscle are relatively dark. </CAPTION>
<REASONING> Bright joint fluid relative to the adjacent dark musculoskeletal structures favors T2 weighting. </REASONING>
<CONCLUSION> T2 </CONCLUSION>

Case C24 **Branching contrast-filled arterial lumen**

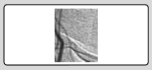


Question. what modality is shown?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [375, 215, 525, 445], "label": "contrast-filled arterial trunk"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [64, 59, 92, 127] (contrast-filled arterial trunk). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [465, 470, 705, 660], "label": "opacified lateral arterial branch"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [81, 133, 124, 187] (opacified lateral arterial branch).
</tool_response>



<SUMMARY> Recognize the vessel-focused projection imaging method. </SUMMARY>
<CAPTION> A strongly contrasted arterial trunk is projected on a low-detail background; a lateral arterial branch is continuously opacified. </CAPTION>
<REASONING> Selective visualization of the arterial trunk and side branch identifies an angiogram. </REASONING>
<CONCLUSION> angiogram </CONCLUSION>

Figure F.8: VQA-Med 2019 examples recognizing Doppler ultrasound (C22), T2-weighted MRI (C23), and an angiogram (C24).

Case C25 **CT identification from bone and soft tissue**



Question. is this a ct scan?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [355, 445, 560, 585], "label": "vertebral cortical margin"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [363, 536, 574, 705] (vertebral cortical margin). </tool_response>



<SUMMARY> Check whether the image has CT attenuation and cross-sectional anatomy. </SUMMARY>
<CAPTION> The vertebra has a very bright cortical margin in a cross-sectional image. </CAPTION>
<REASONING> The bright bone and soft-tissue attenuation pattern are characteristic of CT, so the answer is yes. </REASONING>
<CONCLUSION> yes </CONCLUSION>

Case C26 **Projection radiography of the hand**



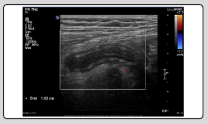
Question. what type of image modality is seen?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [410, 475, 640, 645], "label": "metacarpal cortices"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [419, 622, 656, 845] (metacarpal cortices). </tool_response>



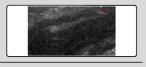
<SUMMARY> Identify the imaging modality from the projected skeletal anatomy. </SUMMARY>
<CAPTION> Projected metacarpal shafts have bright cortices and darker interiors. </CAPTION>
<REASONING> The projected phalanges and metacarpals identify a plain film X-ray. </REASONING>
<CONCLUSION> plain film X-ray </CONCLUSION>

Case C27 **Grayscale sonographic tissue architecture**



Question. what imaging modality is used to acquire this picture?

<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [275, 365, 500, 550], "label": "hypoechoic structure in speckled tissue"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [281, 254, 512, 383] (hypoechoic structure in speckled tissue). </tool_response>



<SUMMARY> Recognize ultrasound from tissue texture and layered interfaces. </SUMMARY>
<CAPTION> A hypoechoic structure lies within speckled soft tissue. </CAPTION>
<REASONING> The returned crop contains a hypoechoic structure within speckled tissue. Together with the superficial layered echoes in the original image, this identifies ultrasound. </REASONING>
<CONCLUSION> ultrasound </CONCLUSION>

Case C28 **IV contrast versus oral contrast on abdominal CT**



Question. was iv contrast given to the patient?

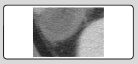
<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [470, 465, 590, 610], "label": "central vascular enhancement"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [235, 166, 297, 219] (central vascular enhancement). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [665, 585, 770, 720], "label": "enhancing renal parenchyma"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [333, 210, 387, 259] (enhancing renal parenchyma). </tool_response>



<tool_call> {"name": "image_zoom_in_tool", "arguments": {"image_index": 0, "bbox_2d": [225, 610, 365, 745], "label": "contralateral renal parenchyma"}} </tool_call>
<tool_response> Figure 1 (image_index=0), zoomed ROI [112, 218, 184, 268] (contralateral renal parenchyma). </tool_response>



<SUMMARY> Look for vascular and renal enhancement rather than relying on bowel brightness. </SUMMARY>
<CAPTION> Central vascular lumina are bright relative to nearby soft tissue; Renal parenchyma enhances strongly; the opposite kidney also has enhanced parenchyma. </CAPTION>
<REASONING> Bright intravascular contrast and renal enhancement support administration of IV contrast. </REASONING>
<CONCLUSION> yes </CONCLUSION>

Figure F.9: VQA-Med 2019 examples recognizing CT (C25), a plain-film X-ray (C26), ultrasound (C27), and IV contrast enhancement (C28).