# $M^3$DT-RL: Multimodal Multidisciplinary Diagnostic Teaming with Reinforcement Learning

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

## Abstract

Multidisciplinary consultation integrates complementary specialty perspectives, but training a shared policy for its distinct diagnostic roles remains challenging. We study reinforcement learning over the full multidisciplinary diagnostic trajectory, termed *MDT-RL*, for multimodal differential diagnosis (DDx) reranking. MDT-RL faces two challenges: trajectory-level rewards provide coarse credit across stages, while uneven learning progress makes fixed rollout allocation inefficient. We introduce **$M^3$DT-RL**, combining *Modular Stage Training* (MST) and *Mastery-Driven Sampling* (MDS). MST jointly optimizes full MDT trajectories and isolated Initial, Expert, and Final tasks with revision-aware credit, while MDS tracks task mastery and adaptively reallocates rollouts toward insufficiently learned tasks. We further construct **AcuteM$^2$DDx**, containing 4,225 acute and emergency-related cases and 22,208 images, with 537 cases published from March 2026 onward reserved for temporal holdout evaluation. Across six multimodal DDx benchmarks, **$M^3$DT-RL**-4B achieves macro-average Hit@1/MRR of 60.89%/75.62%, and **$M^3$DT-RL**-9B reaches 62.93%/77.12%. Both outperform the compared commercial MLLMs, larger open general and medical MLLMs, and prompted multi-agent workflows. These results show that multidisciplinary diagnosis depends not only on model scale or agent count, but also on how diagnostic roles and their coordination are trained. Our code is available at https://anonymous.4open.science/r/M3DT-RL-2460/.

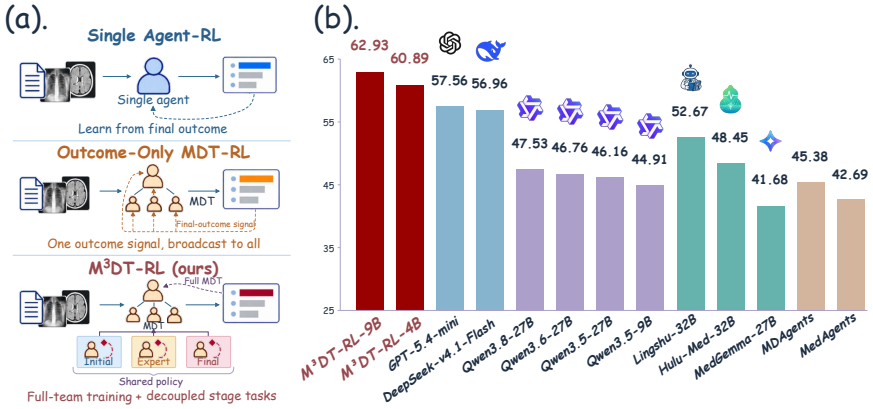



Figure 1: Overview of **$M^3$DT-RL** for multimodal DDx reranking. **(a)** Single-Agent RL, Outcome-Only MDT-RL, and **$M^3$DT-RL** with decoupled stage training. **(b)** Mean Hit@1 across six datasets for **$M^3$DT-RL** (red), commmercial models (blue), open-source general models (purple), medical models (teal), and multi-agent baselines (tan).

## 1 Introduction

Multimodal large language models (MLLMs) enable diagnostic reasoning over clinical histories and medical images (Yao et al., 2026; Pan et al., 2026; Ning et al., 2026). Case-level diagnosis requires comparing competing hypotheses against evidence across patient history and examinations (Heneghan et al., 2009; Brush Jr et al., 2017; Thampy et al., 2019).



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.

This is particularly challenging in acute and emergency cases, where evidence may span organ systems and specialties (Pelaccia et al., 2014; Croskerry, 2002). Multidisciplinary consultation addresses this challenge by integrating specialist assessments into a shared judgment. Medical multi-agent systems (Tang et al., 2024; Kim et al., 2024; Wu et al., 2025b) operationalize this structure through specialist collaboration and adaptive team organization in complex clinical settings. This raises a natural question: how should models be trained for the distinct diagnostic roles within multidisciplinary consultation?

Recent RL-based medical systems have optimized different components of clinical reasoning, including specialty routing and opinion integration (Xia et al., 2026), diagnostic testing and evidence acquisition (Gao et al.; Ren et al., 2025), structured diagnostic inquiry (Guo et al., 2026), and general clinical reasoning (Liu et al., 2026b). Meanwhile, recent multi-agent RL studies explore adaptive agent orchestration and coordinated stage-wise updates (Dang et al., 2026; Xie et al., 2026). We study the intersection of these directions: reinforcement learning over the full multidisciplinary diagnostic trajectory, which we term *MDT-RL*. Here, a shared policy learns the attending and specialist roles within a multistage consultation that reranks candidate diagnoses from clinical history and medical images.

Training an MDT policy introduces two challenges. As illustrated in Fig. 1(a), extending RL from a single agent to a multistage MDT creates a credit-assignment problem: the consultation contains diagnostic responsibilities, where the attending forms an initial judgment, specialists reassess the case from specialty perspectives, and the attending integrates their reports into a final decision. Outcome-only MDT training broadcasts a single trajectory-level reward across these stages, providing coarse supervision and making it difficult to attribute the final outcome to individual decisions (Foerster et al., 2018; Zhang, 2026). Second, the stages differ in learning difficulty and improve at different rates, so allocating rollout budget across stages can waste computation on already-mastered behaviors while undertraining weaker ones (Narvekar et al., 2020; Xiong et al., 2025). These challenges motivate two complementary components: I. *Modular Stage Training* (MST) for stage-aware credit assignment and optimization, and II. *Mastery-Driven Sampling* (MDS) for adaptive rollout allocation.

**Framework.** Accordingly, we introduce ***M³DT-RL***, which improves shared-policy training through targeted learning signals and adaptive rollout allocation while keeping the MDT inference workflow fixed. MST augments full MDT trajectory training with isolated Initial, Expert, and Final tasks. Full trajectories preserve joint learning across roles, while isolated tasks provide direct supervision for each stage. For the Final stage, revision-aware credit rewards correcting an initially wrong decision and penalizes overturning an initially correct one. MDS tracks mastery over Full, Initial, Expert, and Final tasks and dynamically adjusts their sampling probabilities. Unseen or insufficiently mastered tasks receive more rollouts, while mastered tasks are sampled less frequently. Together, MST improves stage-level learning signals, while MDS allocates rollout budget according to mastery.

**Benchmark.** Studying MDT-RL also requires data that connect multimodal case evidence, diagnostic decisions, and consultation-level evaluation. We therefore construct ***AcuteM²DDx***, a multimodal DDx benchmark for acute and emergency-related cases (Fig. 2). Each case combines clinical history, multiple medical images, and candidate diagnoses into a unified reranking task, requiring models to rank diagnostic hypotheses against the available evidence. This formulation provides a common task interface across MDT roles and stages, and supports both SFT and RL under the same diagnostic objective. ***AcuteM²DDx*** contains 4,225 cases and 22,208 images spanning 14 modalities, with 537 independently collected cases first published from March 2026 onward reserved as a temporal holdout test set.

**Experiments.** Across six multimodal DDx benchmarks, ***M³DT-RL*** improves both Qwen3.5 backbones. As shown in Fig. 1(b), ***M³DT-RL***-4B raises macro-average Hit@1 (43.77% → 60.89%) and MRR (60.26% → 75.62%), while ***M³DT-RL***-9B raises Hit@1 (44.91% → 62.93%) and MRR (63.01% → 77.12%). Both variants outperform the compared commercial MLLMs, larger open general and medical MLLMs, and prompted multi-agent workflows in macro-average Hit@1 and MRR. Moreover, at least one ***M³DT-RL*** variant achieves the highest Hit@1 and MRR on five of six benchmarks. These results suggest that effective multidisciplinary diagnosis is not simply a matter of scaling model size or agent count; how diagnostic roles and their coordination are trained also matters.

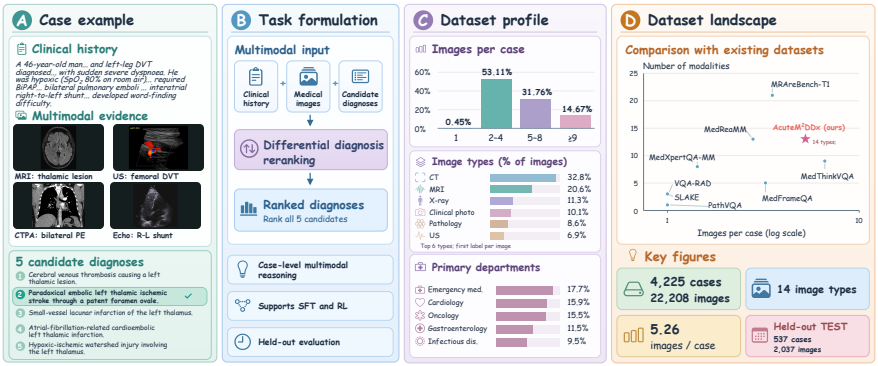



Figure 2: Overview of ***AcuteM²DDx***. **(A)** A representative multimodal case. **(B)** Multimodal differential diagnosis reranking task. **(C)** Image distributions and primary departments. **(D)** Comparison with existing benchmarks and key characteristics.

## 2 ***AcuteM²DDx*** BENCHMARK

***AcuteM²DDx*** evaluates multimodal DDx reranking for acute and emergency-related cases (Fig. 2). Each case contains clinical history, medical images, and a fixed candidate set, enabling the same diagnostic decision to be evaluated across MDT roles and stages.

**Task and metrics.** Following prior DDx reranking settings (Miyachi et al., 2023; Groza et al., 2026), each case is $x = (h, \mathcal{V}, \mathcal{C})$, where $h$ is the clinical history, $\mathcal{V}$ the medical images, and $\mathcal{C}$ five candidate diagnoses containing one reference diagnosis $c^\star$. The model outputs a complete ranking $\rho$. For case $d$, let $r_d^\star = \mathrm{rank}_{\rho_d}(c_d^\star)$. We report Hit@1 and MRR:

$$\text{Hit@1} = \frac{1}{D}\sum_{d=1}^{D}\mathbb{I}[r_d^\star = 1], \qquad \text{MRR} = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{r_d^\star}. \tag{1}$$

Candidate sets are fixed before inference and shared across all evaluated models.

**Data construction.** ***AcuteM²DDx*** contains 4,225 cases and 22,208 images, including 3,019 SFT, 669 RL, and 537 TEST cases. SFT and RL are derived from MedThinkVQA and MultiCaRe (Yao et al., 2026; Nievas Offidani et al., 2025). Training cases first undergo Qwen3.5-9B mastery screening, and cases solved correctly in all three trials are removed to focus on unresolved cases. Qwen3.5-27B, Qwen3.6-27B, and Qwen3.8-27B then generate diagnostic trajectories for the remaining cases, with successful trajectories retained for training. RL additionally requires each case to involve at least two specialties beyond Emergency Medicine, increasing the representation of cross-specialty cases.

**Temporal TEST.** TEST comprises independently collected PMC and Eurorad cases published from March 2026 onward, disjoint from the training splits. Inputs include only evidence available before diagnostic confirmation. Candidate diagnoses are checked for synonymy, granularity, and ambiguity; physicians review uncertain cases.

## 3 METHOD

***M³DT-RL*** trains a shared policy through MDT consultation, MDS, and MST (Fig. 3). MDS samples Full and isolated-stage tasks (B), and the policy executes the corresponding consultation stages (A). MST derives learning signals from rollout outcomes to update the policy (C), while observed success updates MDS weights for subsequent sampling.

**MDT consultation.** An ED attending and $K$ specialists share policy $\pi_\theta$, conditioned on their roles and stage-specific contexts. Given case $x$, the attending produces an initial ranking $\rho^{(I)}$ during INITIAL, then selects specialists $\mathcal{E} = \{\sigma_k\}_{k=1}^{K}$ from a predefined pool during PLAN. In EXPERT, each specialist receives the clinical history, all case images, and $\rho^{(I)}$, and returns an assessment $a^{(k)}$ and ranking $\rho^{(k)}$. Specialists work in parallel with separate contexts, without access to one another's reports. In FINAL, the attending uses the

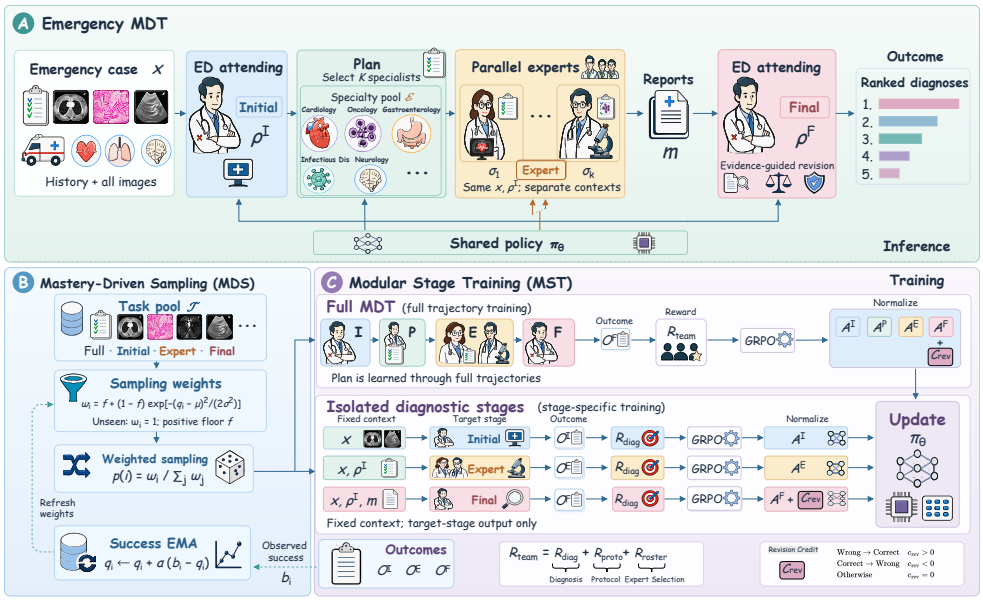



Figure 3: ***M³DT-RL*** overview. (A) One shared policy performs initial ranking, specialist selection, parallel consultation, and final revision. (B) MDS uses a smooth success-dependent weight with a positive floor, assigning maximal weight to unseen instances. (C) MST combines full trajectories with isolated diagnostic stages. Valid full trajectories receive diagnostic, protocol, and roster rewards. Expert-conditioned revision scores are centered within the prompt group and added to Final advantages after GRPO normalization, with segment-weight correction.

collected reports $m = \{(\sigma_k, a^{(k)}, \rho^{(k)})\}_{k=1}^{K}$ to revise $\rho^{(I)}$ into the final ranking $\rho^{(F)}$:

$$x \xrightarrow{\text{Initial}} \rho^{(I)} \xrightarrow{\text{Plan}} \mathcal{E} \xrightarrow{\text{Expert}} m \xrightarrow{\text{Final}} \rho^{(F)}. \tag{2}$$

Revisions in the Final stage require case-specific evidence rather than agreement alone.

For training, each consultation is represented as a trajectory with segments $\mathcal{S} = \{\mathrm{I}, \mathrm{P}, \mathrm{E}_1, \ldots, \mathrm{E}_K, \mathrm{F}\}$. MST combines full-trajectory training with isolated-stage training: the former trains all stages, while the latter trains Initial, Expert, and Final under fixed contexts. MDS allocates rollouts across these task instances based on their observed success.

### 3.1 MDS: Allocating Training Rollouts

MDS adapts rollout allocation using observed success at the task-instance level (Fig. 3B). The task pool $\mathcal{T}$ contains Full and isolated-stage instances, each with a success estimate $q_i$. The observed success $b_i \in \{0, 1\}$ indicates top-1 correctness at the Final stage for a Full task or at the target stage for an isolated task. After each observation, MDS updates the estimate using an exponential moving average:

$$q_i \leftarrow (1 - \alpha) q_i + \alpha b_i, \tag{3}$$

where $\alpha \in (0, 1]$ controls the weight of the latest observation.

MDS uses these estimates to determine subsequent sampling distributions. At step $t$, let $q_i^{(t)}$ denote the latest estimate and $\mathcal{U}_t$ the set of instances with no observed outcomes. The sampling weight $\omega_i^{(t)}$ follows a smooth kernel centered at an intermediate success rate $\mu$, with width $\sigma$ and a positive floor $f$; unseen instances receive weight one. Normalizing these weights gives the sampling probability $p_t(i)$:

$$\omega_i^{(t)} = \begin{cases} 1, & i \in \mathcal{U}_t, \\ f + (1 - f) \exp\left[ -\dfrac{(q_i^{(t)} - \mu)^2}{2\sigma^2} \right], & i \in \mathcal{T} \setminus \mathcal{U}_t, \end{cases} \qquad p_t(i) = \frac{\omega_i^{(t)}}{\sum_{j \in \mathcal{T}} \omega_j^{(t)}}. \tag{4}$$

Here $0 < f < 1$, $0 < \mu < 1$, and $\sigma > 0$. This mechanism favors observed instances whose estimated success is near $\mu$, while preserving a nonzero sampling probability for every instance. Sampling spans the pool, allowing the proportions of Full and isolated tasks to vary with their weights.

### 3.2 MST: Learning Diagnostic Stages

**Task rewards and group advantages.** Let $r^\star(\rho)$ denote the reference diagnosis's rank and $b(\rho) = \mathbb{I}[r^\star(\rho) = 1]$ indicate top-1 correctness. The diagnostic reward is

$$R_{\text{diag}}(\rho) = \frac{1}{r^\star(\rho)^2} + \lambda_{\text{hit}} b(\rho), \tag{5}$$

where $\lambda_{\text{hit}}$ controls the top-1 correctness bonus. For a Full trajectory $\tau$, the terminal reward combines diagnosis, protocol compliance, and specialist selection:

$$R_{\text{team}} = v_\tau \big[ R_{\text{diag}}(\rho^{(F)}) + \lambda_{\text{prot}} u_\tau + R_{\text{roster}} \big], \tag{6}$$

where $v_\tau$ indicates protocol validity, $u_\tau$ indicates first-pass success throughout the workflow, and $\lambda_{\text{prot}}$ controls the corresponding protocol bonus. The roster reward compares specialist roles selected during Plan with mapped case-level department labels:

$$R_{\text{roster}} = \begin{cases} \lambda_{\text{roster}}, & \text{primary-department match,} \\ \eta \lambda_{\text{roster}}, & \text{secondary-only match,} \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where $\lambda_{\text{roster}} > 0$ sets the primary-match reward and $\eta \in (0, 1)$ discounts secondary-only matches. Isolated tasks use diagnostic and first-pass rewards but no roster reward.

For rollouts sharing a task input, GRPO (Shao et al., 2024) converts task rewards $R_g$ into group-normalized advantages:

$$\widehat{A}_g = \frac{R_g - \overline{R}}{\text{std}(R) + \varepsilon}, \tag{8}$$

where $\overline{R}$ and $\text{std}(R)$ are the group mean and standard deviation, and $\varepsilon$ stabilizes normalization. Full tasks share $\widehat{A}_g$ across all stages; isolated tasks apply it to the target stage.

**Revision credit.** MST supplements the base advantage with local credit for revisions in the Final stage, based on initial correctness and specialist reliability. Let $b_{s,g}$ indicate top-1 correctness at stage $s$ for rollout $g$, and let $h_g$ be the fraction of specialists with incorrect top-ranked diagnoses. For isolated Final tasks, initial correctness and specialist reliability are evaluated from the fixed context. For a protocol-valid rollout with an accepted Final decision, define

$$c_g = \begin{cases} \lambda_{\text{res}} h_g, & b_{\text{I},g} = 0,\ b_{\text{F},g} = 1,\ h_g > 0, \\ \lambda_{\text{adopt}}, & b_{\text{I},g} = 0,\ b_{\text{F},g} = 1,\ h_g = 0, \\ -\lambda_{\text{corrupt}}, & b_{\text{I},g} = 1,\ b_{\text{F},g} = 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

The positive coefficients $\lambda_{\text{res}}$, $\lambda_{\text{adopt}}$, and $\lambda_{\text{corrupt}}$ control correction credit when some specialists are incorrect, adoption credit when all specialists are correct, and the penalty for overturning a correct diagnosis, respectively. Other transitions receive zero raw local credit.

Let $\mathcal{G}_{\text{F}}$ contain the protocol-valid rollouts with accepted Final decisions in a prompt group, and let $w_{g,\text{F}} > 0$ be the Final segment's optimization weight. For nonempty $\mathcal{G}_{\text{F}}$, MST centers the local scores and adjusts the Final advantages after GRPO normalization:

$$\overline{c}_{\text{F}} = \frac{1}{|\mathcal{G}_{\text{F}}|} \sum_{j \in \mathcal{G}_{\text{F}}} c_j, \qquad A_{g,s} = \widehat{A}_g + \mathbb{I}[s = \text{F}] \, \frac{c_g - \overline{c}_{\text{F}}}{w_{g,\text{F}}}, \quad g \in \mathcal{G}_{\text{F}}. \tag{10}$$

Centering makes the adjustment relative to the group mean; dividing by $w_{g,\text{F}}$ compensates for segment weighting in the loss. The adjustment applies to Full and isolated Final tasks, while other accepted segments retain their base advantages. Ineligible Final decisions receive no local adjustment, and identical eligible scores yield zero adjustment.

**Optimization.** Training consists of SFT followed by RL, both using the same consultation-stage decomposition. SFT initializes the shared policy by maximizing the conditional likelihood of supervised outputs:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(c,y)\sim\mathcal{D}_{\mathrm{SFT}}}\left[\sum_{t=1}^{|y|}\log \pi_\theta(y_t \mid c, y_{<t})\right], \tag{11}$$

where $c$ is the stage context and $y$ is the supervised output; the loss is computed only over output tokens. RL assigns equal total weight to each trajectory, distributed uniformly over its trainable segments $\mathcal{L}_g$:

$$w_{g,s} = \frac{1}{|\mathcal{L}_g|}, \qquad \mathcal{L}_{\mathrm{RL}}(\theta) = \frac{1}{|\mathcal{B}|}\sum_{g\in\mathcal{B}}\sum_{s\in\mathcal{L}_g} w_{g,s}\,\ell_{g,s}^{\mathrm{GRPO}}(\theta), \tag{12}$$

where $\mathcal{B}$ is the trajectory batch and $\ell_{g,s}^{\mathrm{GRPO}}$ uses the stage advantage $A_{g,s}$.

# 4 Experiments

We evaluate diagnostic performance, identify the contributions of training components, and examine expert selection and final revision: **RQ1:** Does ***M³DT-RL*** improve diagnostic ranking across benchmarks? **RQ2:** How do MDS and MST contribute to these improvements? **RQ3:** How does expert selection affect consultation quality? **RQ4:** How effectively does the attending use expert opinions to revise its diagnosis?

## 4.1 Experimental Setup

**Implementation Details. Implementation Details.** We use Qwen3.5-9B as the base model and train it with Adam at a learning rate of $10^{-6}$. Both rollout and training batch sizes are set to 16, with 4 rollouts per sample.

**Datasets.** We evaluate on ***AcuteM²DDx***-TEST, MedThinkVQA (Yao et al., 2026), DDx-TRACE (Pan et al., 2026), MedFrameQA (Yu et al., 2025), MRareBench-Diagnosis (Ning et al., 2026), and MedReaMM (Wei et al., 2026). All use the five-candidate ranking interface; the main comparison reports Hit@1 and MRR and their unweighted macro-average. Invalid outputs and execution failures receive zero credit.

**Baselines.** The main table includes commercial, open general, open medical, and multi-agent models. Commercial models include GPT-5.4-mini[1] and DeepSeek-V4.1-Flash (Xu et al., 2026a). Open general models include Qwen3.5[2] at 2B, 4B, 9B, 27B, Qwen3.6-27B[3], and Qwen3.8-27B[4]. Open medical models include MedGemma-1.5-4B-IT and MedGemma-27B-IT (Sellergren et al., 2025; 2026), Lingshu-I-8B and Lingshu-32B (Xu et al., 2026b), and Hulu-Med at 7B and 32B (Jiang et al., 2025a). Multi-agent models are MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024), use Qwen3.5-9B as backbone.

## 4.2 RQ1: Diagnostic Ranking Performance

Table 1 shows that ***M³DT-RL***-9B achieves the highest macro-averaged Hit@1 and MRR at 62.93% and 77.12%. It exceeds GPT-5.4-mini by 5.37 and 4.93 percentage points and DeepSeek-v4.1-Flash by 5.97 and 4.78 points, respectively. The 4B variant ranks second on both metrics, also exceeding these commercial models and all evaluated 27B and 32B open models. With the same Qwen3.5-9B backbone, MDAgents changes macro-averaged Hit@1 and MRR by +0.47 and -0.30 points relative to single-model inference, while MedAgents lowers both. ***M³DT-RL***-9B instead improves them by 18.02 and 14.11 points. This comparison measures the complete method's gains over the evaluated workflows; it does not isolate

[1] https://openai.com/index/introducing-gpt-5-4-mini-and-nano
[2] https://qwen.ai/blog?id=qwen3.5
[3] https://qwen.ai/blog?id=qwen3.6-27b
[4] https://qwen.ai/blog?id=qwen3.8

Table 1: DDx reranking performance on five multimodal benchmarks and ***AcuteM²DDx***. We report Hit@1 and MRR (%), with the top five distinct scores per column highlighted by rank (ties share the same rank). **Avg.** denotes the macro-average across six benchmarks.

| Model | MedThinkVQA | | DDx-TRACE | | MedFrameQA | | MRareBench | | MedReaMM | | *AcuteM²DDx* | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR |
| *Commercial MLLMs* | | | | | | | | | | | | | | |
| GPT-5.4-mini | **37.92** | **59.57** | **69.63** | **80.90** | **51.32** | **69.56** | **49.50** | **65.96** | **73.12** | **81.40** | **63.87** | **75.76** | **57.56** | **72.19** |
| DeepSeek-v4.1-Flash | **33.99** | **57.54** | **65.97** | **77.92** | **51.25** | **69.77** | **51.17** | **67.45** | **75.52** | **84.33** | **63.87** | **77.06** | **56.96** | **72.34** |
| *Open General MLLMs* | | | | | | | | | | | | | | |
| Qwen3.5-2B | 17.56 | 30.35 | 20.94 | 31.56 | 21.50 | 37.29 | 15.38 | 26.12 | 17.60 | 25.12 | 19.18 | 30.24 | 18.69 | 30.11 |
| Qwen3.5-4B | 31.32 | 50.61 | 50.79 | 64.39 | 34.27 | 52.68 | 37.46 | 56.95 | 59.04 | 71.19 | 49.72 | 65.71 | 43.77 | 60.26 |
| Qwen3.5-9B | 29.78 | 51.84 | 54.45 | 68.91 | 39.57 | 59.95 | 36.45 | 56.26 | 58.72 | 72.62 | 50.47 | 68.48 | 44.91 | 63.01 |
| Qwen3.5-27B | 30.34 | 50.73 | 53.40 | 68.07 | 39.39 | 57.83 | **43.48** | **61.63** | 59.52 | 71.34 | 50.84 | 66.02 | 46.16 | 62.60 |
| Qwen3.6-27B | 26.97 | 46.16 | 54.45 | 67.11 | 41.07 | 57.86 | 42.14 | 59.36 | 61.76 | 71.06 | 54.19 | 67.14 | 46.76 | 61.45 |
| Qwen3.8-27B | 31.74 | 53.31 | 53.93 | 68.77 | 41.84 | 61.47 | **43.48** | 61.41 | 62.40 | 74.47 | 51.77 | 68.18 | 47.53 | 64.60 |
| *Open Medical MLLMs* | | | | | | | | | | | | | | |
| Lingshu-I-8B | 28.65 | 52.41 | 29.32 | 47.95 | 29.81 | 49.55 | 27.76 | 50.50 | 40.80 | 58.85 | 36.50 | 56.90 | 32.14 | 52.69 |
| Lingshu-32B | **35.81** | **58.09** | **60.21** | **74.49** | **53.35** | **70.71** | 40.47 | 59.77 | **66.56** | **78.38** | **59.59** | **74.28** | **52.67** | **69.29** |
| Hulu-Med-7B | 24.72 | 47.85 | 42.93 | 61.62 | 23.15 | 40.88 | 30.77 | 49.74 | 37.28 | 52.08 | 35.20 | 55.57 | 32.34 | 51.29 |
| Hulu-Med-32B | **33.99** | 55.73 | 56.02 | 70.79 | 38.06 | 54.36 | 42.47 | 61.08 | 63.36 | 75.68 | **56.80** | 71.64 | 48.45 | 64.88 |
| MedGemma-1.5-4B | 23.03 | 47.43 | 35.08 | 57.09 | 8.14 | 17.51 | 24.41 | 47.73 | 22.72 | 46.19 | 28.86 | 51.01 | 23.71 | 44.49 |
| MedGemma-27B | 23.46 | 47.70 | 53.40 | 68.75 | 38.41 | 58.39 | 29.77 | 52.61 | 56.80 | 71.83 | 48.23 | 65.79 | 41.68 | 60.85 |
| *Multi-Agent Workflows (Qwen3.5-9B Base)* | | | | | | | | | | | | | | |
| MedAgents | 26.26 | 49.68 | 48.17 | 64.00 | 36.93 | 59.45 | 38.46 | 59.00 | 59.20 | 72.87 | 47.11 | 64.75 | 42.69 | 61.63 |
| MDAgents | 31.04 | 52.65 | 50.79 | 65.02 | 39.00 | 59.02 | 38.80 | 58.48 | 61.60 | 74.17 | 51.02 | 66.89 | 45.38 | 62.71 |
| **Ours** *(Qwen3.5-4B Base & 9B Base)* | | | | | | | | | | | | | | |
| ***M³DT-RL*-4B** | **54.58** | **71.56** | **73.30** | **84.13** | **45.53** | **63.90** | **48.49** | **68.27** | **71.36** | **82.96** | **72.07** | **82.89** | **60.89** | **75.62** |
| ***M³DT-RL*-9B** | **54.03** | **71.07** | **71.20** | **82.75** | **50.02** | **67.69** | **51.51** | **70.30** | **77.44** | **86.87** | **73.37** | **84.01** | **62.93** | **77.12** |

individual training components. ***M³DT-RL*-9B** exceeds every baseline in both metrics on five of six benchmarks. On ***AcuteM²DDx***, its Hit@1 reaches 73.37%, compared with 63.87% for both GPT-5.4-mini and DeepSeek-v4.1-Flash. Its MRR reaches 84.01%, compared with 75.76% and 77.06%, respectively. On MRareBench, the improvement over DeepSeek-v4.1-Flash is 0.34 points in Hit@1 and 2.85 points in MRR, indicating a larger change in ranking quality than in top-1 correctness. Lingshu-32B remains the strongest model on MedFrameQA.

### 4.3 RQ2: Contributions of Training Components

Table 2 shows that Outcome-Only MDT-RL performs similarly to Single-Agent RL. The complete method reaches 65.51% Hit@1 and 79.00% MRR, exceeding Outcome-Only MDT-RL by 8.41 and 6.43 percentage points. Adding MST improves average Hit@1 and MRR by 4.13 and 3.15 points. In Fig. 4(a), adding the Final objective changes the final-minus-initial Hit@1 gap from -0.47 to +0.60 points. Further adding Expert and Initial objectives raises oracle Hit@1 from 66.40% to 69.11%, without monotonically increasing the consultation gain (Fig. 4b). Removing revision credit lowers Hit@1 and MRR by 2.68 and 1.39 points (Fig. 4c). Adding MDS to MST yields another 1.77 points in Hit@1 and 0.89 points in MRR, with the largest gains on ***AcuteM²DDx*** (+4.84/+3.36 points). The MDS run also maintains higher Hit@1 at every displayed checkpoint (Fig. 4d). SFT initialization adds 2.51 and 2.39 points over MST+MDS. Conversely, MST+MDS improves SFT-initialized MDT-RL by 5.82 and 5.47 points, with gains in both metrics on all five benchmarks.

### 4.4 RQ3: Expert Selection

Across the five datasets, 1,169 cases have department labels that can be mapped to the predefined pool of 12 specialist roles. On these cases, Plan selects the reference expert first in 69.2% of cases and includes it in the selected pair in 80.4%, compared with 13.2% and 19.0% under within-dataset team reassignment (Fig. 5b). Matching varies across departments: either-expert matching reaches 99% for oncology/blood cases but only 62% for pediatrics and 12% for general surgery (Fig. 5a).

Removing roster reward lowers five-dataset mean Hit@1 from 65.51% to 65.02% and MRR from 79.00% to 78.62% (Fig. 5c). The selection analysis therefore shows strong alignment with mapped department labels, while the reward ablation measures a modest contribution to diagnostic performance.

Table 2: Ablation results for DDx reranking across five multimodal benchmarks. We report Hit@1 and MRR (%), with the top three distinct scores in each column highlighted by rank (ties share the same rank). **Avg.** is the macro-average across all five benchmarks.

| Method | MedThinkVQA | | DDx-TRACE | | MRareBench | | MedReaMM | | *AcuteM²DDx* | | Avg. | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR | H@1 | MRR |
| *Baselines* | | | | | | | | | | | | |
| Qwen3.5-9B | 29.78 | 51.84 | 54.45 | 68.91 | 36.45 | 56.26 | 58.72 | 72.62 | 50.47 | 68.48 | 45.97 | 63.62 |
| Single-Agent RL | 41.81 | 62.23 | 67.02 | 78.63 | 45.82 | 63.52 | 72.32 | 81.79 | 62.38 | 75.79 | 57.87 | 72.39 |
| Prompted MDT | 33.33 | 55.13 | 55.50 | 71.27 | 43.48 | 62.89 | 61.92 | 75.75 | 52.33 | 69.30 | 49.31 | 66.87 |
| Outcome-Only MDT-RL | 42.16 | 61.12 | 67.02 | 80.23 | 45.82 | 64.74 | 69.44 | 80.62 | 61.08 | 76.15 | 57.10 | 72.57 |
| **MDT-RL Variants** | | | | | | | | | | | | |
| + MST | **47.50** | **67.06** | **73.82** | **83.77** | 44.82 | 65.19 | **76.32** | **85.18** | 63.69 | 77.42 | **61.23** | **75.72** |
| + MST + MDS | **48.06** | **66.94** | **73.30** | **83.23** | **48.16** | **67.25** | **76.96** | **84.87** | **68.53** | **80.78** | **63.00** | **76.61** |
| + SFT Init. | 44.31 | 62.84 | 67.54 | 79.08 | **48.83** | **66.35** | 71.68 | 81.86 | **66.11** | **77.50** | 59.69 | 73.53 |
| **+ SFT Init. + MST + MDS** | **54.03** | **71.07** | **71.20** | **82.75** | **51.51** | **70.30** | **77.44** | **86.87** | **73.37** | **84.01** | **65.51** | **79.00** |

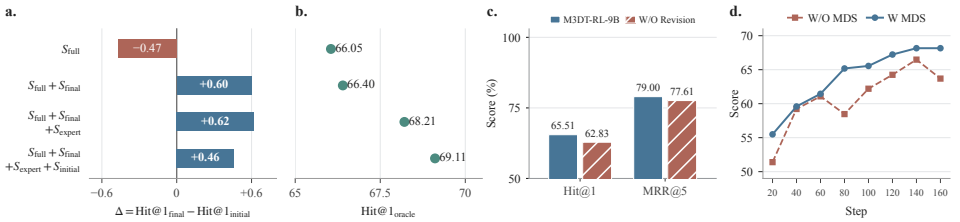



Figure 4: Stage objectives, revision, and adaptive sampling. (a) Final-minus-initial Hit@1 and (b) Oracle Hit@1 under cumulative stage objectives, averaged over MedThinkVQA, DDx-TRACE, MRareBench, MedReaMM, and ***AcuteM²DDx***. (c) Hit@1 and MRR@5 with and without revision. (d) ***AcuteM²DDx*** Hit@1 over training steps with and without MDS.

Increasing expert count improves oracle performance more than final performance (Fig. 5d). From one to eleven experts, oracle Hit@1 rises from 70.94% to 78.08%, whereas final Hit@1 changes from 65.23% to 65.66% and stays within 65.23–66.52% across all tested team sizes. Thus, the additional oracle coverage is not matched by a comparable improvement in the attending's final ranking.

### 4.5 RQ4: Revision and Expert Input

Removing revision reduces five-dataset mean Hit@1 from 65.51% to 62.83% and MRR from 79.00% to 77.61% (Fig. 4c). The resulting drops of 2.68 and 1.39 percentage points show a contribution to both top-1 correctness and diagnostic ranking.

We further replace the experts with GPT-5.4-mini. On ***AcuteM²DDx***, Hit@1 and MRR reach 73.93% and 84.47%, improving over the original 9B configuration by 0.56 and 0.46 points. On MRareBench, they reach 54.18% and 71.90%, with larger gains of 2.67 and 1.60 points. The benefit of replacing the expert model therefore varies between the two benchmarks. Oracle Hit@1 in these runs reaches 85.47% and 68.23%, leaving gaps of 11.54 and 14.05 points to the final decisions. Expert replacement improves final performance on both datasets, but the attending still falls short of the corresponding oracle results.

## 5 Related Work

**Multimodal Medical Reasoning.** Medical multimodal reasoning has progressed from early biomedical vision-language models such as `Med-Flamingo` (Moor et al., 2023) and `LLaVA-Med` (Li et al., 2023) to more general systems including `Med-PaLM M` (Tu et al., 2024), `Med-Gemini` (Saab et al., 2024), `RadFM` (Wu et al., 2025a), and `BioMedGPT` (Luo et al., 2024). Specialized models such as `CheXagent` (Chen et al., 2024b), `MAIRA-2` (Bannur et al., 2024), `CONCH` (Lu et al., 2024a), and `PathChat` (Lu et al., 2024b) further target radiology and pathology. Recent benchmarks increasingly emphasize multi-image reasoning and differential diagnosis, including MedFrameQA (Yu et al., 2025), MedThinkVQA (Yao et al., 2026), DDx-TRACE (Pan et al., 2026), MMRareBench (Ning et al., 2026), and MedReaMM (Wei

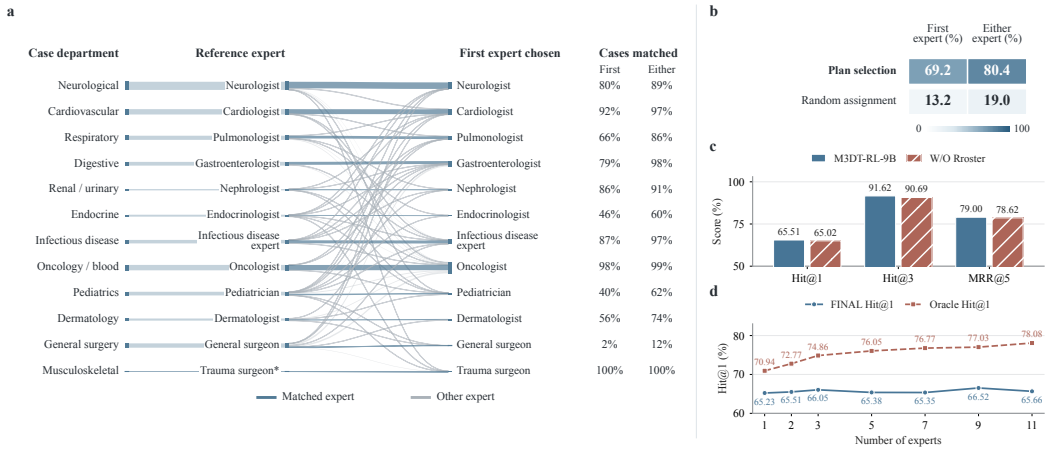



Figure 5: **Expert selection and roster analysis.** (a) Department-to-expert flows and matching on 1,169 cases ($^*n = 2$). (b) Plan selection versus within-dataset team reassignment. (c) Roster ablation. (d) Expert-count scaling. Panels (c–d) report five-dataset means.

et al., 2026). ***M³DT-RL*** formulates multimodal differential diagnosis as multidisciplinary decision-making, combining complementary specialty perspectives to rank diagnoses.

**Multidisciplinary Medical Agents.** Medical agents increasingly adopt role-based collaboration. `MedAgents` (Tang et al., 2024) uses multidisciplinary discussion, while `MDAgents` (Kim et al., 2024) adapts collaboration to case complexity. `AI Hospital` (Fan et al., 2025), `Agent Hospital` (Li et al., 2024), and `AgentClinic` (Schmidgall et al., 2026) study multi-role interaction, while `EHRAgent` (Shi et al., 2024) and `MedAgentBench` (Jiang et al., 2025b) focus on EHR and tool interaction. `RareAgents` (Chen et al., 2026) and `MMedAgent-RL` (Xia et al., 2026) move toward explicit multidisciplinary teams. ***M³DT-RL*** frames multimodal diagnosis as team ranking, combining specialist reranking with attending adjudication.

**Reinforcement Learning for Collaborative Reasoning.** Reinforcement learning has become a major approach for improving reasoning in LLMs, including `DeepSeekMath` (Shao et al., 2024) and `DeepSeek-R1` (Guo et al., 2025), with related medical efforts such as `HuatuoGPT-o1` (Chen et al., 2024a) and `Med-RLVR` (Zhang et al., 2025). `Agentic RL` extends this setting to search and tool use through `Search-R1` (Jin et al., 2025), `ReTool` (Feng et al., 2026), `DeepResearcher` (Zheng et al., 2025), `FlowReasoner` (Gao et al., 2025), and `Agent-R1` (Cheng et al., 2025). Multi-agent optimization has also emerged through `MAPoRL` (Park et al., 2025), `MAGRPO` (Liu et al., 2026c), `SPIRAL` (Liu et al., 2026a), and `Stronger-MAS` (Zhao et al., 2026). ***M³DT-RL*** combines full-trajectory rewards with stage-specific diagnostic training and adaptive task sampling for multimodal multidisciplinary diagnosis.

## 6 Conclusion

We study *MDT-RL*, reinforcement learning over a multidisciplinary diagnostic trajectory in which a shared policy performs distinct attending and specialist roles. To address coarse stage-level credit and uneven learning progress, we introduce ***M³DT-RL***, combining Modular Stage Training (MST) with Mastery-Driven Sampling (MDS). We also construct ***AcuteM²DDx*** to support training and temporal evaluation of multimodal DDx reranking in acute and emergency-related cases. Across six benchmarks, ***M³DT-RL*** consistently improves both 4B and 9B backbones and compares favorably with larger MLLMs and prompted multi-agent workflows. These results suggest that effective multidisciplinary diagnosis depends not only on model scale or agent count, but also on how diagnostic roles and their coordination are trained. Our study is limited to closed-set DDx reranking and a controlled consultation workflow; extending MDT-RL to open-ended diagnosis and clinically grounded evaluation remains important future work.

### AI use statement

In this work, we used generative AI tools to assist with research methodology and experimental design, implementation of the proposed method, data analysis, translation, and interpretation of experimental results. Generative AI was also used to support preliminary reasoning and analytical conclusions.

Generative AI models assisted the construction and review pipeline of ***AcuteM$^2$DDx***. All cases in ***AcuteM$^2$DDx*** were derived from publicly available medical datasets, and generative AI was not used to generate synthetic patient cases. Generative AI was also used to assist with parts of the training and evaluation code, manuscript drafting and editing, literature search and summarization, preparation and refinement of scientific figures, and LaTeX editing.

All AI-assisted outputs used in the final work were reviewed by the authors. Benchmark annotations were subject to human verification and correction, and AI-assisted code was inspected, modified, and tested before use. Reported results, factual claims, and references were checked against the underlying experimental data and original sources. We take responsibility for all content and artifacts included in the final submission.

### Ethics statement

This work studies multimodal diagnostic reasoning and multidisciplinary collaboration for research purposes and is not intended for direct clinical deployment or medical decision-making. ***AcuteM$^2$DDx*** is constructed from publicly available medical datasets and does not involve the collection of new private patient data. We follow the applicable terms and licenses of the source datasets. Generative AI models were used to assist parts of the benchmark construction and review process, with human reviewers verifying and correcting the resulting cases before their inclusion in the benchmark.

As with other medical AI systems, the evaluated models may inherit biases, errors, or limitations from their training data and may exhibit uneven performance across clinical conditions, imaging modalities, or patient populations. Performance on ***AcuteM$^2$DDx*** should therefore not be interpreted as evidence of clinical safety or effectiveness. The benchmark and methods presented in this work are intended to support research on diagnostic reasoning and multidisciplinary collaboration rather than replace qualified clinical judgment.

### Reproducibility statement

We provide the information needed to reproduce our experiments throughout the main paper and appendix. The proposed multidisciplinary training framework, agent communication procedure, and reward design are described in the Method section. The appendix documents the construction and data splits of ***AcuteM$^2$DDx***, its generative-AI-assisted and human-reviewed curation procedure, external evaluation datasets, evaluated models and baselines, system prompts and agent roles, and the evaluation protocol. These details are provided to support consistent reconstruction of both the training and evaluation pipelines.

## References

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

John E Brush Jr, Jonathan Sherbino, and Geoffrey R Norman. How expert clinicians intuitively recognize a medical diagnosis. *The American journal of medicine*, 130(6): 629–634, 2017.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms. *arXiv preprint arXiv:2412.18925*, 2024a.

Xuanzhong Chen, Ye Jin, Xiaohao Mao, Lun Wang, Shuyang Zhang, and Ting Chen. Rareagents: Autonomous multi-disciplinary team for rare disease diagnosis and treatment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 101–109, 2026.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024b.

Mingyue Cheng, Shuo Yu, Daoyu Wang, Qingchuan Li, Xiaoyu Tao, Jie Ouyang, Yucong Luo, Yitong Zhou, Qi Liu, and Enhong Chen. Agent-r1: A unified and modular framework for agentic reinforcement learning. *arXiv preprint arXiv:2511.14460*, 2025.

Pat Croskerry. Achieving quality in clinical decision making: cognitive strategies and detection of bias. *Academic emergency medicine*, 9(11):1184–1204, 2002.

Yufan Dang, Chen Qian, Xueheng Luo, Jingru Fan, Zihao Xie, Ruijie Shi, Weize Chen, Cheng Yang, Xiaoyin Che, Ye Tian, et al. Multi-agent collaboration via evolving orchestration. *Advances in neural information processing systems*, 38:165025–165059, 2026.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10183–10213, 2025.

Jiazhan Feng, Shijue Huang, Xingwei Qu, Ge Zhang, Yujia Qin, Baoquan Zhong, Chengquan Jiang, Jinxin Chi, and Wanjun Zhong. Retool: Reinforcement learning for strategic tool use in llms. In *International Conference on Learning Representations*, volume 2026, pp. 37909–37926, 2026.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Hongcheng Gao, Yue Liu, Yufei He, Longxu Dou, Chao Du, Zhijie Deng, Bryan Hooi, Min Lin, and Tianyu Pang. Flowreasoner: Reinforcing query-level meta-agents. *arXiv preprint arXiv:2504.15257*, 2025.

Shuohao Gao, Xuanzhong Chen, Lingxiao Luo, Zilin Ding, Rong Han, Rui Jiang, and Ting Chen. Salus: Strategic diagnostic testing for complex diagnosis via multi-agent reinforcement learning. In *Forty-third International Conference on Machine Learning*.

Tudor Groza, Freddie Tan, Noah Tian Run Lim, Maheshwaran Windersalam Shanmugasundar, Jhanvi Kappaganthu, Jane Andrea Lieviant, Neerja Karnani, Haichao Chen, Tien Y Wong, and Saumya Shekhar Jamuar. Rd-embed: Unified representations of rare-disease knowledge from clinical records. *medRxiv*, pp. 2026–04, 2026.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Yue Guo, Fanfu Wang, Jianwei Lv, Xincheng Shi, Yuchen Li, Youya Wang, Yunsheng Zeng, Yujing Liu, Yunhao Qiao, Gen Li, et al. Dr. assistant: Enhancing clinical diagnostic inquiry via structured diagnostic reasoning data and reinforcement learning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 36624–36658, 2026.

C Heneghan, P Glasziou, M Thompson, P Rose, J Balla, D Lasserson, C Scott, and R Perera. Diagnostic strategies used in primary care. *Bmj*, 338, 2009.

Songtao Jiang, Yuan Wang, Sibo Song, Tianxiang Hu, Chenyi Zhou, Bin Pu, Yan Zhang, Zhibo Yang, Yang Feng, Joey Tianyi Zhou, et al. Hulu-med: A transparent generalist model towards holistic medical vision-language understanding. *arXiv preprint arXiv:2510.08668*, 2025a.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *Nejm Ai*, 2(9):AIdbp2500144, 2025b.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik S Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae W Park. Mdagents: An adaptive collaboration of llms for medical decision-making. *Advances in Neural Information Processing Systems*, 37:79410–79452, 2024.

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.

Junkai Li, Yunghwei Lai, Weitao Li, Jingyi Ren, Meng Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, et al. Agent hospital: A simulacrum of hospital with evolvable medical agents. *arXiv preprint arXiv:2405.02957*, 2024.

Bo Liu, Simon Yu, Zichen Liu, Leon Guertler, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, et al. Spiral: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 35407–35434, 2026a.

Chi Liu, Yan Shu, Mengzhuo Chen, Hongming Piao, Zhijian Duan, Derek Li, and Bryan Dai. Scaling is not all you need: Clinical-oriented reinforcement learning makes parameter-efficient clinical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15056–15068, 2026b.

Shuo Liu, Zeyu Liang, Xueguang Lyu, and Christopher Amato. Llm collaboration with multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32150–32158, 2026c.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature medicine*, 30(3):863–874, 2024a.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024b.

Yizhen Luo, Jiahuan Zhang, Siqi Fan, Kai Yang, Massimo Hong, Yushuai Wu, Mu Qiao, and Zaiqing Nie. Biomedgpt: An open multimodal large language model for biomedicine. *IEEE journal of biomedical and health informatics*, 2024.

Yasuhiko Miyachi, Osamu Ishii, and Keijiro Torigoe. Design, implementation, and evaluation of the computer-aided clinical decision support system based on learning-to-rank: collaboration between physicians and machine learning in the differential diagnosis process. *BMC medical informatics and decision making*, 23(1):26, 2023.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine learning for health (ML4H)*, pp. 353–367. PMLR, 2023.

Sanmit Narvekar, Bei Peng, Matteo Leonetti, Jivko Sinapov, Matthew E Taylor, and Peter Stone. Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50, 2020.

Mauro Nievas Offidani, Facundo Roffet, María Carolina González Galtier, Miguel Massiris, and Claudio Delrieux. An open-source clinical case dataset for medical image classification and multimodal ai applications. *Data*, 10(8):123, 2025.

Junzhi Ning, Jiashi Lin, Yingying Fang, Wei Li, Jiyao Liu, Cheng Tang, Chenglong Ma, Wenhao Tang, Tianbin Li, Ziyan Huang, et al. Mmrarebench: A rare-disease multimodal and multi-image medical benchmark. *arXiv preprint arXiv:2604.10755*, 2026.

Jiazhen Pan, Weixiang Shen, Jun Li, Julian Canisius, Felix Bitzer, Paula RoĆ mĆ1/4ller, Jiancheng Yang, Virginie Kreutzinger, Daniel Rueckert, and Benedikt Wiestler. Ddx-trace: A benchmark for medical diagnostic trajectories in vlms. *arXiv preprint arXiv:2605.23629*, 2026.

Chanwoo Park, Seungju Han, Xingzhi Guo, Asuman E Ozdaglar, Kaiqing Zhang, and Joo-Kyung Kim. Maporl: Multi-agent post-co-training for collaborative large language models with reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30215–30248, 2025.

Thierry Pelaccia, Jacques Tardif, Emmanuel Triby, Christine Ammirati, Catherine Bertrand, Valérie Dory, and Bernard Charlin. How and when do expert emergency physicians generate and evaluate diagnostic hypotheses? a qualitative study using head-mounted video cued-recall interviews. *Annals of Emergency Medicine*, 64(6):575–585, 2014.

Weijieying Ren, Tianxiang Zhao, Lei Wang, Tianchun Wang, and Vasant G Honavar. Diallms: Ehr-enhanced clinical conversational system for clinical test recommendation and diagnosis prediction. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 25622–25635, 2025.

Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Ji Woong Kim, Eduardo Pontes Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal benchmark for tool-using clinical ai agents. *npj Digital Medicine*, 2026.

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.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, et al. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*, 2026.

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

Wenqi Shi, Ran Xu, Yuchen Zhuang, Yue Yu, Jieyu Zhang, Hang Wu, Yuanda Zhu, Joyce C Ho, Carl Yang, and May Dongmei Wang. Ehragent: Code empowers large language models for few-shot complex tabular reasoning on electronic health records. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22315–22339, 2024.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. Medagents: Large language models as collaborators for zero-shot medical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 599–621, 2024.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, Jianfeng Cai, Xinyuan Cai, Peizhou Cao, Yuxuan Cao, Ziwei Chai, Y Charles, et al. Kimi k3: Open frontier intelligence. *arXiv preprint arXiv:2607.24653*, 2026.

Harish Thampy, Emma Willert, and Subha Ramani. Assessing clinical reasoning: targeting the higher levels of the pyramid. *Journal of general internal medicine*, 34(8):1631–1636, 2019.

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

Lai Wei, Yuchao Chen, Zhenbiao Cao, Xiaojin Zhang, Zhongyu Wei, Bangting Wang, Wei Chen, and Xiang Bai. Medreamm: Evaluating large multimodal models on expert-level clinical diagnostic synthesis. *arXiv preprint arXiv:2608.22323*, 2026.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 16(1):7866, 2025a.

Xiao Wu, Ting-Zhu Huang, Liang-Jian Deng, Yanyuan Qiao, Imran Razzak, and Yutong Xie. A knowledge-driven adaptive collaboration of llms for enhancing medical decision-making. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33483–33500, 2025b.

Peng Xia, Jinglu Wang, Yibo Peng, Kaide Zeng, Zihan Dong, Xian Wu, Xiangru Tang, Hongtu Zhu, Yun Li, Linjun Zhang, et al. Mmedagent-rl: Optimizing multi-agent collaboration for multimodal medical reasoning. In *International Conference on Learning Representations*, volume 2026, pp. 81513–81546, 2026.

Yi Xie, Siao Liu, Falong Fan, Yuanqi Yao, Yue Zhao, and Bo Liu. Teamtr: Trust-region fine-tuning for multi-agent llm coordination. *arXiv preprint arXiv:2605.15207*, 2026.

Wei Xiong, Chenlu Ye, Baohao Liao, Hanze Dong, Xinxing Xu, Christof Monz, Jiang Bian, Nan Jiang, and Tong Zhang. Reinforce-ada: An adaptive sampling framework for reinforce-style llm training. *arXiv e-prints*, pp. arXiv–2510, 2025.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. Deepseek-v4. 1-flash: Pushing the limits of kv cache compression. *arXiv preprint arXiv:2609.19969*, 2026a.

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*, 2026b.

Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Zhichao Yang, Arman Cohan, and Hong Yu. Medical thinking with multiple images. In *International Conference on Learning Representations*, volume 2026, pp. 101924–101988, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *Advances in Neural Information Processing Systems*, 38:113222–113244, 2026.

Suhao Yu, Haojin Wang, Juncheng Wu, Luyang Luo, Jingshen Wang, Cihang Xie, Pranav Rajpurkar, Carl Yang, Yang Yang, Kang Wang, et al. Medframeqa: A multi-image medical vqa benchmark for clinical reasoning. *arXiv preprint arXiv:2505.16964*, 2025.

Chenchen Zhang. From reasoning to agentic: Credit assignment in reinforcement learning for large language models. *arXiv preprint arXiv:2604.09459*, 2026.

Sheng Zhang, Qianchu Liu, Guanghui Qin, Tristan Naumann, and Hoifung Poon. Med-rlvr: Emerging medical reasoning from a 3b base model via reinforcement learning. *arXiv preprint arXiv:2502.19655*, 2025.

Yujie Zhao, Lanxiang Hu, Yang Wang, Minmin Hou, Hao Zhang, Ke Ding, and Jishen Zhao. Stronger-mas: Multi-agent reinforcement learning for collaborative llms. In *International Conference on Learning Representations*, volume 2026, pp. 150619–150651, 2026.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025.

# Appendix

## A Analysis of Modular Stage Training and Mastery-Driven Sampling

We provide further analysis of the two main components of our training framework. For Modular Stage Training (MST), we examine why trajectory-level team rewards, although valid for optimizing the team objective, do not directly evaluate individual diagnostic stages, and how isolated stage objectives provide complementary learning signals. For Mastery-Driven Sampling (MDS), we analyze why instances with intermediate success rates are more likely to yield within-group outcome diversity under group-relative optimization. We additionally characterize the Final-stage revision credit and the estimation and sampling properties used by MDS.

The gradient identities below concern unnormalized, on-policy likelihood-ratio updates with differentiable policies of fixed support and finite second moments. Differentiation and expectation are assumed to commute. These results explain individual components of the training signal; they are not convergence or monotonic-improvement guarantees for clipped, finite-group GRPO (Shao et al., 2024).

### A.1 Why Stage-Specific Objectives Complement Team Rewards

Let $u_s = (o_s, z_s)$ denote the visible context and clinical role at stage $s$, and define the sequence score

$$\psi_s = \nabla_\theta \log \pi_\theta(y_s \mid u_s). \tag{13}$$

For a fixed case, the trajectory probability factors over the generated segments, with each segment conditioned on its visible context. For

$$J_{\text{team}}(\theta) = \mathbb{E}_{\tau \sim p_\theta}\left[R_{\text{team}}(\tau)\right], \tag{14}$$

the likelihood-ratio gradient can therefore be written as

$$\nabla_\theta J_{\text{team}}(\theta) = \sum_{s \in \mathcal{S}} \mathbb{E}\left[R_{\text{team}} \psi_s\right]. \tag{15}$$

Thus, broadcasting a shared terminal reward to all generated stages is a valid likelihood-ratio update for the team objective. The motivation for MST is not that this update is incorrect, but that a single terminal reward does not directly evaluate the diagnostic output produced at each stage.

To make this distinction explicit, define the conditional team value

$$Q_s(u_s, y_s) = \mathbb{E}\left[R_{\text{team}} \mid u_s, y_s\right]. \tag{16}$$

This quantity is used only for analysis; MST neither learns nor explicitly estimates $Q_s$.

**Proposition 1 (Conditional decomposition of stage feedback).** For each stage $s$,

$$\mathbb{E}\left[R_{\text{team}} \psi_s\right] = \mathbb{E}\left[Q_s \psi_s\right], \tag{17}$$

and

$$\begin{aligned} \operatorname{Cov}\left(R_{\text{team}} \psi_s\right) = \operatorname{Cov}\left(Q_s \psi_s\right) \\ + \mathbb{E}\left[\operatorname{Var}\left(R_{\text{team}} \mid u_s, y_s\right) \psi_s \psi_s^{\mathsf{T}}\right]. \end{aligned} \tag{18}$$

The second term is positive semidefinite.

*Proof.* Because $\psi_s$ is determined by $(u_s, y_s)$,

$$\mathbb{E}\left[R_{\text{team}}\psi_s \mid u_s, y_s\right] = Q_s\psi_s. \tag{19}$$

Equation (17) follows from iterated expectation. Applying the law of total covariance gives Eq. (18), whose second term is an expectation of a nonnegative scalar multiplied by an outer product. □

**Implication for MST.** The additional covariance term captures variation in the team outcome that remains unresolved after conditioning on the stage context and generated output. For example, the reward assigned to an Initial or Expert decision can still depend on later decisions in the MDT trajectory.

MST complements this trajectory-level signal with an isolated stage reward $R_s$ that directly evaluates the diagnostic output produced at stage $s$. In general, $R_s$ is not equal to $Q_s$. Under a matched context distribution,

$$\mathbb{E}\left[R_s\psi_s\right] - \mathbb{E}\left[R_{\text{team}}\psi_s\right] = \mathbb{E}\left[(R_s - Q_s)\psi_s\right]. \tag{20}$$

By Cauchy–Schwarz,

$$\|\mathbb{E}\left[(R_s - Q_s)\psi_s\right]\|_2 \le \sqrt{\mathbb{E}\left[(R_s - Q_s)^2\right]}\sqrt{\mathbb{E}\left[\|\psi_s\|_2^2\right]}. \tag{21}$$

Historical or perturbed contexts can additionally change the stage-context distribution. Isolated stage tasks should therefore be viewed as auxiliary conditional objectives rather than unbiased replacements for the team objective. MST retains full-trajectory optimization so that the shared policy continues to receive direct supervision from team outcomes while also receiving stage-specific diagnostic feedback.

### A.2 Final-Stage Revision Credit

The Final stage has an additional role: deciding whether the evidence from specialist consultation justifies revising the Initial diagnosis. We distinguish the evaluation-side consultation gain from the training-side revision credit.

**Evaluation-side consultation gain.** Let $b_{\text{I}}$ and $b_{\text{F}}$ indicate whether the Initial and Final top-ranked diagnoses are correct. Define

$$\Delta_{\text{rev}} = b_{\text{F}} - b_{\text{I}}. \tag{22}$$

Then

$$\begin{aligned} H_{\text{F}} - H_{\text{I}} &= \mathbb{E}\left[\Delta_{\text{rev}}\right] \\ &= \Pr\left(b_{\text{I}} = 0, b_{\text{F}} = 1\right) - \Pr\left(b_{\text{I}} = 1, b_{\text{F}} = 0\right). \end{aligned} \tag{23}$$

For $0 < H_{\text{I}} < 1$, this can equivalently be written as

$$\begin{aligned} H_{\text{F}} - H_{\text{I}} = (1 - H_{\text{I}})\Pr\left(b_{\text{F}} = 1 \mid b_{\text{I}} = 0\right) \\ - H_{\text{I}}\Pr\left(b_{\text{F}} = 0 \mid b_{\text{I}} = 1\right). \end{aligned} \tag{24}$$

Thus, the net consultation gain is determined by the balance between correcting initially wrong decisions and corrupting initially correct ones. The same identities hold for empirical paired predictions.

**Training-side revision credit.** The local training score $c_g$ is distinct from $\Delta_{\text{rev}}$. It additionally incorporates specialist correctness and is defined only for protocol-valid rollouts with accepted Final decisions. For rollout $g$, define the expert miss fraction

$$h_g = \frac{1}{K}\sum_{k=1}^{K}\left(1 - b_{\text{E}_k,g}\right). \tag{25}$$

The rescue/corrupt rule in Eq. (9) assigns positive raw credit when the Final stage corrects an initially wrong diagnosis, negative raw credit when it overturns an initially correct diagnosis, and zero otherwise. Thus, the training signal directly targets the direction of revision rather than rewarding revision itself.

**Centered Final-stage adjustment.** Let $\mathcal{V}$ denote the protocol-valid rollouts with accepted Final decisions in one prompt group. Let $n_g$ be the number of live trainable segments in rollout $g$. Under `shared_policy_units`, each such segment receives weight

$$w_{g,s} = \frac{1}{n_g}. \tag{26}$$

**Proposition 2 (Centered revision credit under segment weighting).** For nonempty $\mathcal{V}$, define the Final-stage adjustment as

$$\delta A_{g,\mathrm{F}} = \frac{c_g - \bar{c}_{\mathcal{V}}}{w_{g,\mathrm{F}}} = n_g \left(c_g - \bar{c}_{\mathcal{V}}\right), \qquad g \in \mathcal{V}, \tag{27}$$

where

$$\bar{c}_{\mathcal{V}} = \frac{1}{|\mathcal{V}|} \sum_{g \in \mathcal{V}} c_g. \tag{28}$$

Then

$$\sum_{g \in \mathcal{V}} w_{g,\mathrm{F}} \delta A_{g,\mathrm{F}} = 0. \tag{29}$$

*Proof.* For every $g \in \mathcal{V}$,

$$w_{g,\mathrm{F}} \delta A_{g,\mathrm{F}} = c_g - \bar{c}_{\mathcal{V}}. \tag{30}$$

Therefore,

$$\sum_{g \in \mathcal{V}} w_{g,\mathrm{F}} \delta A_{g,\mathrm{F}} = \sum_{g \in \mathcal{V}} (c_g - \bar{c}_{\mathcal{V}}) = 0. \tag{31}$$

□

The reciprocal weight correction ensures that the effective contribution of the Final-local signal is the centered revision score even when trajectories contain different numbers of trainable segments. If all eligible Final decisions receive the same raw score, their local adjustments vanish after centering. An empty eligible set receives no local adjustment.

**Two reward routes.** The team reward and the revision credit enter optimization through separate routes. For a valid Full trajectory, the terminal team reward is group-normalized and its advantage is shared across generated optimization segments. In contrast, the rescue/corrupt score is not added to the terminal reward. Its centered, weight-corrected adjustment is applied only to the Final segment. Isolated Final tasks use the same local revision route but do not receive credit for a replayed specialist roster.

This separation avoids directly assigning the revision signal to earlier stages. Because all roles share parameters, Final-stage updates can still indirectly affect behavior at other stages. These algebraic properties do not imply monotonic improvement under the clipped GRPO objective.

### A.3 Why MDS Favors Intermediate Mastery

Group-relative optimization depends on variation among sampled rewards (Shao et al., 2024; Yu et al., 2026). MDS therefore uses recent task success as a heuristic for identifying instances that are more likely to produce heterogeneous outcomes within a rollout group.

Fix an instance $i$ and a behavior policy. Let $\zeta_i$ denote the probability that one rollout places the reference diagnosis first. This true success probability is distinct from the EMA estimate $q_i$ used by MDS and from the sampling probability $p_t(i)$.

Assume that the $G \geq 2$ rollouts in a group share the same task context and behavior policy and differ only through independent sampling randomness:

$$b_{i,g} \overset{\text{iid}}{\sim} \text{Bernoulli}(\zeta_i). \tag{32}$$

For Full tasks, independence is assumed between complete trajectories rather than between roles within a trajectory.

Define the mixed-correctness event

$$\mathcal{M}_i = \left\{ 0 < \sum_{g=1}^{G} b_{i,g} < G \right\}. \tag{33}$$

**Proposition 3 (Probability of mixed-correctness groups).** The probability that a rollout group contains both successful and unsuccessful trajectories is

$$\Pr(\mathcal{M}_i) = F_G(\zeta_i) := 1 - \zeta_i^G - (1-\zeta_i)^G. \tag{34}$$

For $G \geq 2$, $F_G(\zeta_i)$ is maximized at $\zeta_i = 1/2$.

*Proof.* A group is not mixed exactly when all $G$ rollouts succeed or all $G$ rollouts fail. These disjoint events have probabilities $\zeta_i^G$ and $(1-\zeta_i)^G$, respectively, which gives Eq. (34). Moreover,

$$F_G'(\zeta) = G\left[(1-\zeta)^{G-1} - \zeta^{G-1}\right]. \tag{35}$$

The derivative is positive for $\zeta < 1/2$ and negative for $\zeta > 1/2$, so the maximum occurs at $\zeta = 1/2$. □

This result gives a simple interpretation of the MDS sampling rule. Very easy instances tend to produce uniformly correct groups, while very hard instances tend to produce uniformly incorrect groups. Instances at intermediate mastery levels are more likely to produce both outcomes within the same group.

The MDS kernel is centered around this intermediate regime, but it should not be interpreted as a calibrated estimator of $F_G(\zeta_i)$. It operates on an online EMA estimate, includes a positive sampling floor, and assigns maximal initial weight to unseen instances. It therefore biases sampling toward regions associated with higher mixed-outcome probability without guaranteeing mixed outcomes in every sampled group.

**Connection to diagnostic reward variation.** Assume valid rankings and nonnegative diagnostic-reward coefficients, and define

$$d_R := \frac{3}{4} + \lambda_{\text{hit}} > 0. \tag{36}$$

Every top-1 correct rollout satisfies

$$R_{\text{diag}} \geq 1 + \lambda_{\text{hit}}, \tag{37}$$

whereas every top-1 incorrect valid rollout satisfies

$$R_{\text{diag}} \leq \frac{1}{4}. \tag{38}$$

Hence, a mixed-correctness group necessarily contains nonconstant diagnostic rewards.

For $\zeta_i \in (0,1)$, let

$$m_b = \mathbb{E}\left[R_{\text{diag}} \mid b_i = b\right]. \tag{39}$$

By the law of total variance,

$$\begin{aligned} \operatorname{Var}\left(R_{\text{diag}} \mid i\right) &= \mathbb{E}\left[\operatorname{Var}\left(R_{\text{diag}} \mid b_i, i\right)\right] \\ &\quad + \zeta_i(1-\zeta_i)(m_1 - m_0)^2 \\ &\geq \zeta_i(1-\zeta_i)d_R^2. \end{aligned} \tag{40}$$

Thus, intermediate-success instances are associated with a larger lower bound on diagnostic reward variability under the stated assumptions. Reward variability, however, does not guarantee a nonzero expected gradient or improved performance. Mixed correctness is also not necessary for a useful update: incorrect rankings can receive different rank-based rewards, and Final-stage revision scores can provide additional within-group variation. MDS should therefore be interpreted as a sampling heuristic for increasing the likelihood of outcome diversity rather than as an exact test for useful gradient updates.

### A.4 EMA Estimation and Sampling Support

MDS does not observe the true success probability $\zeta_i$. Instead, it maintains an online estimate $q_i$ from previously observed outcomes. We characterize this estimate under a simplified stationary setting and then describe the support induced by the positive-floor sampling kernel.

**EMA estimation.** Consider a stationary stream of independent Bernoulli observations with mean $\zeta_i$, deterministic initialization $q_{i,0}$, and EMA coefficient $\alpha$. After $n$ observations,

$$q_{i,n} = (1-\alpha)^n q_{i,0} + \alpha \sum_{j=1}^{n} (1-\alpha)^{n-j} b_{i,j}. \tag{41}$$

Its bias and variance are

$$\mathbb{E}[q_{i,n}] - \zeta_i = (1-\alpha)^n (q_{i,0} - \zeta_i), \tag{42}$$

and

$$\mathrm{Var}(q_{i,n}) = \frac{\alpha}{2-\alpha} \left[1 - (1-\alpha)^{2n}\right] \zeta_i (1-\zeta_i). \tag{43}$$

The limiting variance corresponds to an effective sample size of

$$N_{\mathrm{eff}} = \frac{2-\alpha}{\alpha}. \tag{44}$$

For example, $\alpha = 0.25$ gives $N_{\mathrm{eff}} = 7$.

These expressions describe a stationary filtering process rather than a calibration guarantee for MDS during training. Policy updates, correlated rollouts, adaptive sampling, and delayed asynchronous results violate the assumptions above. If the EMA is updated using a group mean rather than a single Bernoulli outcome, the observation variance must also be adjusted accordingly.

**Positive sampling support.** Let the MDS sampling weight for instance $i$ at training step $t$ be $\omega_i^{(t)}$ as defined in Eq. (4). With a positive floor $0 < f < 1$,

$$f \le \omega_i^{(t)} \le 1. \tag{45}$$

For a finite task pool $\mathcal{T}$ with $N_{\mathcal{T}} = |\mathcal{T}|$,

$$p_t(i) = \frac{\omega_i^{(t)}}{\sum_{j \in \mathcal{T}} \omega_j^{(t)}} \ge \frac{f}{N_{\mathcal{T}}} > 0. \tag{46}$$

Therefore, low-weight instances remain in the sampling support and can be revisited as the shared policy changes. Positive support does not guarantee a bounded realized waiting time, so EMA estimates may still become stale.

**Dynamic allocation across task types.** For task type $\kappa$, let $\mathcal{T}_\kappa \subseteq \mathcal{T}$ denote its instances. Under global weighted sampling,

$$\Pr_t(\kappa) = \sum_{i \in \mathcal{T}_\kappa} p_t(i) = \frac{\sum_{i \in \mathcal{T}_\kappa} \omega_i^{(t)}}{\sum_{j \in \mathcal{T}} \omega_j^{(t)}}. \tag{47}$$

Consequently, the deterministic composition of the training pool determines which Full and isolated stage instances are available, but it does not fix their realized rollout proportions. As mastery estimates change, MDS dynamically reallocates rollout budget across both instances and task types.

**Scope of the analysis.** The identities above isolate properties of the learning signals and sampling mechanism. The implemented GRPO objective additionally uses group-relative reward normalization and a clipped surrogate. Output-length reduction, reference-policy regularization, asynchronous sampling, and behavior-policy mismatch can further alter the realized optimization update. Accordingly, Eqs. (17)–(21) should be read as local explanations of MST rather than exact unbiasedness statements for the complete training algorithm.

In contrast, the weighted-centering identity in Eq. (29) and the mixed-group probability in Eq. (34) are exact under their stated assumptions. Together, the analysis supports the two design principles used in ***M³DT-RL***: MST augments team-level optimization with directly evaluated stage objectives, while MDS reallocates rollout budget toward instances that are more likely to yield diverse group-relative outcomes.

# B ***AcuteM²DDx*** Dataset Construction

## B.1 Dataset Overview

***AcuteM²DDx*** contains 4,225 distinct cases and 22,208 referenced medical images for five-candidate DDx reranking. Table 3 summarizes the subsets used for training and evaluation. For case and image statistics, each `case_id` and its associated images are counted once. The three subsets have pairwise disjoint case identifiers.

| Subset | Records | Cases | Images | Img./Case |
|---|---|---|---|---|
| SFT | 10,305 | 3,019 | 16,383 | 5.43 |
| RL | 4,014 | 669 | 3,788 | 5.66 |
| TEST | 537 | 537 | 2,037 | 3.79 |
| Total | 14,856 | 4,225 | 22,208 | 5.26 |

Table 3: Statistics of the SFT, RL, and TEST subsets. Record counts and case counts differ because training includes multiple records per case.

The SFT subset contains 3,019 cases: 962 from MedThinkVQA (Yao et al., 2026) and 2,057 from MultiCaRe (Nievas Offidani et al., 2025). The RL subset contains 272 and 397 cases from the respective sources. TEST is independently collected from later PMC and Eurorad reports.

The median numbers of images per case are 4, 5, and 3 for SFT, RL, and TEST, respectively. The corresponding numbers of cases containing at least two image types are 2,163 (71.6%), 455 (68.0%), and 344 (64.1%). Figure 2 reports image-type frequencies using the first label of each image and primary-department distributions over the 3,688 training cases; TEST departments are excluded because they use a different taxonomy.

## B.2 Construction of the SFT/RL Training Pool

**STEP 1 Source pool construction**

We first construct MIDDx, a general multi-image diagnostic case pool, from MedThinkVQA and MultiCaRe. MedThinkVQA contributes 7,347 training cases derived from Eurorad.

For MultiCaRe, we reconstruct case-level records by linking case text, article metadata, figures, and figure captions. Among 34,955 indexed cases, we retain 5,177 multi-image cases with multi-specialty signals and an additional 2,000 single-specialty cases with high multi-image diagnostic value. These specialty signals are used only for case selection and do not imply that the patients underwent actual multidisciplinary consultation.

We apply case-validity checks, diagnosis-leakage filtering, image–case consistency checks, duplicate detection, source-overlap filtering, and exclusion of cases overlapping with external

evaluation resources. The resulting MIDDx pool contains 12,387 cases and 76,941 images, including 6,131 cases from MedThinkVQA and 6,256 cases from MultiCaRe.

**STEP 2 Acute-case identification**

The acute-case taxonomy is derived from clinical criteria summarized by physicians and converted into structured labeling instructions. The criteria cover emergency-care context, acute onset, high-risk acute conditions, trauma, fracture, rupture, injury, poisoning, and related diagnostic and ICD evidence.

GPT-5.6 Sol[5] applies these instructions to assign an acute-label (A–C). The labels are then audited, with physician adjudication when required.

| | |
|---|---|
| **Level A** | Explicit emergency-care context or a clear high-risk acute presentation. |
| **Level B** | An acute clinical condition without an explicitly documented emergency-care setting. |
| **Level C** | Insufficient evidence of an acute or emergency-related presentation. |

Levels A and B are retained, yielding 5,536 candidate cases and 31,515 images. After excluding one case without a recoverable final diagnosis, the trainable pool contains 5,535 cases and 31,510 images.

**STEP 3 Leakage-aware SFT/RL partition**

Cases are connected when they share at least one of the following: a source identifier or PMCID, normalized article title, normalized clinical history, source image identifier, exact image SHA-256 hash, or perceptual image signature.

Connected case groups are assigned entirely to either SFT or RL. Source and acute-level marginals are explicitly controlled, while ICD category, image modality, image count, longitudinal status, and related case characteristics are matched as closely as possible.

The upstream partition contains 4,151 cases and 23,630 images on the SFT side and 1,384 cases and 7,880 images on the RL side. The final SFT and RL training subsets in Table 3 are drawn from these larger upstream partitions.

### B.3 Independent Temporal Test Set

**STEP 1 Independent collection**

***AcuteM²DDx***-TEST is constructed separately from PMC and Eurorad and does not reuse cases from MedThinkVQA, MultiCaRe, or MIDDx.

**STEP 2 Post-release temporal isolation**

To reduce potential pretraining overlap, we restrict TEST to reports published after the release of the evaluated Qwen3.5 models, within the following source-specific windows:

| | |
|---|---|
| **PMC** | March 6, 2026 through August 2026. |
| **Eurorad** | May through August 2026. |

Publication dates alone do not establish whether a case was included in pretraining data.

**STEP 3 Diagnostic-timepoint control**

Case reports frequently contain information obtained after the final diagnosis has been established.

[5] https://openai.com/index/gpt-5-6/

For each test case, we therefore identify the last clinically meaningful timepoint before diagnostic confirmation and restrict model inputs to the clinical history, imaging findings, and medical images available by that point. Information available only after the diagnostic cutoff, including post-diagnostic treatment, later confirmatory findings, follow-up outcomes, retrospective diagnostic statements, and post-diagnostic images, is removed when applicable.

**STEP 4 Image-level quality control**

Each retained image is checked to ensure that it belongs to the target patient, falls within the permitted diagnostic window, and does not directly reveal the final diagnosis. For compound figures, image–caption correspondence is rechecked after image splitting. Cases with fewer than two valid images after cleaning are excluded.

The resulting ***AcuteM$^2$DDx***-TEST contains 537 cases and 2,037 images.

### B.4 Unified Gold-Diagnosis Annotation

The SFT, RL, and TEST subsets use the same gold-diagnosis construction protocol.

**STEP 1 Final-diagnosis extraction**

GPT-5.6 Luna reads the original source report and extracts the diagnosis ultimately confirmed by the case-report authors. The resulting text is retained as the raw gold diagnosis.

**STEP 2 MONDO normalization**

The raw diagnosis is mapped to MONDO solely for concept normalization and standardized naming.

Cases are flagged for further review when no reliable mapping can be established, multiple plausible concepts remain, or the mapped concept does not preserve the diagnostic granularity of the source diagnosis.

**STEP 3 Diagnosis-equivalent alias construction**

To account for valid diagnostic naming variation, GPT-5.6 Sol constructs an alias set from the raw diagnosis and its MONDO-normalized concept.

| | |
|---|---|
| **Include** | Standardized alternative names, established clinical names, and widely accepted medical abbreviations. |
| **Exclude** | Parent disease categories, different disease subtypes, clinically related but non-equivalent conditions, and alternative differential diagnoses. |

Accepted aliases must preserve the diagnosis and its clinical granularity.

**STEP 4 Five-candidate construction**

For TEST, GPT-5.6 Luna constructs four distractors around the confirmed diagnosis. Candidates are checked for synonymous labels, inconsistent diagnostic granularity, and diagnostic ambiguity.

### B.5 Dual-Model Verification and Selective Physician Review

**STEP 1 Independent verification**

Each annotation is independently reviewed by GPT-5.6 Sol and Kimi-K3 (Team et al., 2026). Both reviewers assess whether the extracted diagnosis is directly supported by the original source report, whether the MONDO mapping preserves the correct disease concept and granularity, and whether each alias is diagnostically equivalent to the gold diagnosis.

Each reviewer additionally reports an uncertainty level.

**STEP 2 Selective physician review**

Physician adjudication is triggered under any of the following conditions:

| | |
|---|---|
| **Uncertainty** | Either reviewer reports moderate or higher uncertainty. |
| **Model disagreement** | GPT-5.6 Sol and Kimi-K3 disagree on the diagnosis, MONDO mapping, or alias set. |
| **Granularity conflict** | The source diagnosis and MONDO concept differ in disease granularity. |

## C Evaluated Models

We evaluate general-purpose and medical-domain vision–language models across six medical multimodal datasets. The general-purpose models include Qwen3.5[6], Qwen3.6[7], Qwen3.8[8], GPT-5.4-mini[9], and DeepSeek-V4.1-Flash. The medical-domain models include MedGemma, Lingshu, and Hulu-Med. All models are evaluated using the shared inputs, frozen candidate sets, and five-candidate ranking interface described in Appendix E.1.

- **Qwen models** are general-purpose vision–language models developed by Alibaba. We evaluate Qwen3.5 at four parameter scales—2B, 4B, 9B, and 27B—together with Qwen3.6-27B and Qwen3.8-27B. This set enables comparison across model scale and generation within the same family. Qwen3.5-9B is also used as the backbone for the multi-agent baselines.
- **GPT-5.4-mini** is a multimodal model from OpenAI's GPT-5.4 family supporting text and image inputs, and configurable reasoning effort. We include it as a commercial general-purpose model.
- **DeepSeek-V4.1-Flash** (Xu et al., 2026a) is a general-purpose multimodal model developed by DeepSeek with native text and image input support and configurable reasoning. It is evaluated through its API deployment.
- **MedGemma models** (Sellergren et al., 2025; 2026) are medical-domain multimodal models developed by Google based on Gemma 3. We evaluate the instruction-tuned MedGemma-1.5-4B-IT and MedGemma-27B-IT.
- **Lingshu models** (Xu et al., 2026b) are medical vision–language models designed for multimodal clinical understanding and reasoning. We evaluate Lingshu-I-8B and Lingshu-32B, which use InternVL3 and Qwen2.5-VL backbones, respectively.
- **Hulu-Med models** (Jiang et al., 2025a) are generalist medical vision–language models supporting medical text, images. We evaluate Hulu-Med-7B and Hulu-Med-32B.

## D Overview of the Baselines

We compare against two training-free multi-agent medical reasoning frameworks, MedAgents (Tang et al., 2024) and MDAgents (Kim et al., 2024). MedAgents emphasizes iterative specialist consultation around a shared report, whereas MDAgents adapts the collaboration structure to query complexity.

- **MedAgents** (Tang et al., 2024) is a training-free framework for zero-shot medical reasoning through collaboration among agents assigned different medical specialties. Specialists independently analyze the case, their opinions are consolidated into a shared report, and disagreements are resolved through iterative consultation before producing the final decision.

[6] https://qwen.ai/blog?id=qwen3.5
[7] https://qwen.ai/blog?id=qwen3.6-27b
[8] https://qwen.ai/blog?id=qwen3.8
[9] https://openai.com/index/introducing-gpt-5-4-mini-and-nano/

Table 4: External subsets in the evaluation registry.

| Dataset | Cases | Prepared images |
|---|---|---|
| MedThinkVQA | 720 | 5,618 |
| DDx-TRACE | 191 | 1,609 |
| MedFrameQA | 2,851 | 9,237 |
| MRareBench (diagnosis) | 299 | 1,052 |
| MedReaMM | 625 | 1,745 |
| **External total** | **4,686** | **19,261** |

• **MDAgents** (Kim et al., 2024) is an adaptive medical collaboration framework that selects its agent organization according to task complexity. It uses a single clinician for simpler cases and progressively introduces multidisciplinary or multi-team collaboration for more complex cases, with intermediate analyses informing the final decision.

# E External Evaluation Datasets and Preparation

We evaluate on five external multimodal medical benchmarks.Candidate sets and their ordering are prepared before evaluation and stored in a frozen registry. Shared input formatting and image processing are described in Appendix E.1, and the common evaluation prompt and output interface are described in Appendix E.2.

• **MedThinkVQA** (Yao et al., 2026) is an expert-annotated benchmark for medical reasoning across multiple images, emphasizing both individual-image interpretation and integration of cross-view evidence.

• **DDx-TRACE** (Pan et al., 2026) is a physician-adjudicated benchmark for diagnostic trajectories in multimodal neuroradiology. Its original task involves requesting imaging studies, updating differential diagnoses as evidence becomes available, and deciding when to stop.

• **MedFrameQA** (Yu et al., 2025) is a multi-image medical visual question-answering benchmark constructed from medical education videos. It pairs related visual frames with questions requiring comparative medical reasoning.

• **MRareBench** (Ning et al., 2026) is a multimodal rare-disease benchmark constructed from PubMed Central case reports. It evaluates the relationship between diagnostic predictions and patient-specific visual evidence through forward diagnosis and evidence verification.

• **MedReaMM** (Wei et al., 2026) evaluates clinical diagnostic synthesis from patient histories and multiple medical images under a complete-information setting. Its diagnostic cases emphasize integration of heterogeneous clinical evidence, including rare, atypical, and multi-system presentations.

**Shared distractor generation.** We use Gemini-3.7-flash [10] to generate and select up to 4 distractors while keeping the reference diagnosis fixed.

**Prepared subset statistics.** Table 4 summarizes the external subsets in the evaluation registry. Statistics for the ***AcuteM²DDx*** test set are reported separately in Appendix B.

## E.1 Shared Inputs and Image Processing

The public case representation contains only `clinical_history`, five candidate texts, and images. The reference candidate appears as an unmarked option. Direct baselines receive candidate texts. Multi-agent adaptations may use an internal A–E mapping, but their final call receives the full candidate texts and uses the same ranking interface.

[10] https://blog.google/innovation-and-ai/models-and-research/gemini-models/introducing-gemini-3-7-flash/

Image order follows the prepared case list. Model-specific image processors remain part of each serving configuration; shared image files do not imply identical internal visual-token representations.

### E.2 Evaluation Prompt and Output Interface

Direct local and commercial baselines use the same zero-shot prompt, The system message is:

```
You are the ED Attending making the final ranked diagnostic
assessment.
```

The user message starts with a JSON object containing three keys: `clinical_history`, `candidates`, and `instruction`. The first two fields contain the prepared case text and the ordered array of five candidate strings. The fixed value of `instruction` is:

```
Rank the 5 candidate diagnoses from most to least likely for this
case.  Return only the ranking.
```

Image markers `img_01`, `img_02`, and so forth are then appended in order, each followed by its image.

Only the ranking and the name in each ranking item are mandatory and used in scoring. Assessment, findings, reasons, and uncertainty are optional; absent values do not invalidate an otherwise valid ranking.

## F Prompt Configuration and Consultation Trajectory

We present the prompting procedure in the same order as a multidisciplinary diagnostic trajectory. Agent-level system prompts are introduced when the corresponding agent first appears.

### F.1 Step 1: Initial Multimodal ED Assessment

The trajectory begins with the ED Attending, which coordinates the complete multidisciplinary diagnostic process. Its system prompt remains fixed throughout the consultation.

**Initial ED assessment**

**SYSTEM**

```
You are the ED Attending responsible for the full emergency multidisciplinary
diagnostic process. You are a member of a multidisciplinary consultation team.
Together, your goal is to arrive at a more accurate differential diagnosis (DDx)
for this patient. Assess the patient from an emergency medicine perspective, rank
the supplied candidate diagnoses, determine when and which specialists should be
consulted, coordinate the consultation process, evaluate and integrate specialist
opinions against the available clinical evidence, and make the final ranked
diagnostic assessment.
```

**USER TEMPLATE**

```
case:
  clinical history:
    {clinical_history}

  candidates:
    - <candidate name>
    - <candidate name>
    - <candidate name>
    - <candidate name>
    - <candidate name>
```

**Initial ED assessment** CONTINUED

```
  images:
    img_01
    img_02
    ...
    img_N

instruction:
Rank all 5 supplied candidate diagnoses from most to least likely.
Return exactly 5 lines in this format:
1. <candidate name>
2. <candidate name>
3. <candidate name>
4. <candidate name>
5. <candidate name>
Use each supplied candidate exactly once, copying its name verbatim. Return no
heading, explanation, or text outside these numbered lines.
```

The case history, candidate set, and associated medical images jointly form the multimodal input to the ED Attending. The image identifiers above represent image attachments in the actual model request rather than textual image descriptions.

### F.2 Step 2: Consultation Planning

After producing the initial diagnostic ranking, the same ED Attending selects $k$ complementary specialists from the predefined specialist pool. The selected specialists subsequently assess the case independently.

**Consultation planning**

**USER TEMPLATE**

```
available_roles:
  [cardiology,
   oncology,
   gastroenterology,
   infectious_disease,
   neurology,
   pediatrics,
   nephrology,
   pulmonology,
   endocrinology,
   trauma_surgery,
   pathology,
   general_surgery,
   dermatology]

budget:
  experts: k

instruction:
Select exactly k complementary, pairwise distinct specialists from available_roles
by their 1-based position numbers (the first entry in available_roles is 1). Each
specialist judges independently; do not assign questions. Call plan_consultations
exactly once. Its arguments must have this structure:
{
  "consultations": [
    {
      "expert_index": <i_1>
```

**Consultation planning** CONTINUED

```
    },
    {
      "expert_index": <i_2>
    },
    ...
    {
      "expert_index": <i_k>
    }
  ]
}
consultations must be a native JSON array of exactly k objects, not a single
object, an array of strings, or a quoted JSON string. Replace every placeholder
number with the position of a different available role. Do not return a text
answer or a Markdown code block.
```

### F.3 Step 3: Specialist Consultation

Once the consultation plan is determined, each selected specialist is instantiated with a fixed specialty-specific system prompt. Specialist prompts follow the same structural organization but differ in their specialty identity, clinical scope, and specialty-specific guidance. We show cardiology as a representative example rather than repeating all 13 role-specific prompts.

**Specialist Consultation**

**SYSTEM**

```
You are the cardiology specialist in an emergency multidisciplinary diagnostic
consultation. You are a member of a multidisciplinary consultation team. Together,
your goal is to arrive at a more accurate differential diagnosis (DDx) for this
patient.

Scope: Clinical cardiac differential diagnosis using symptoms, ECG and supplied
cardiovascular evidence.

Integrate cardiac symptoms with supplied examination, ECG, biomarkers and
cardiovascular evidence. Compare ischemic, structural, inflammatory and arrhythmic
explanations. Do not infer unprovided ECG patterns, pressures or functional
measurements.
```

**USER TEMPLATE**

```
case:
  clinical history:
    {clinical_history}

  candidates:
    - <candidate name>
    - <candidate name>
    - <candidate name>
    - <candidate name>
    - <candidate name>

  images:
    img_01
    img_02
    ...
    img_N

current ranking:
  1. <candidate name>
```

**Specialist Consultation** CONTINUED

```
  2. <candidate name>
  3. <candidate name>
  4. <candidate name>
  5. <candidate name>

instruction:
Using the clinical history and all case images, review the current working ranking
of the 5 candidate diagnoses from your specialty perspective and return your own
permutation. This is a closed-set ranking task: do not introduce, rename, or
abbreviate diagnoses. Use only supplied findings; do not invent missing tests or
repeat the full history. When you keep the working order, name the supplied
finding that supports it; when you change a position, the Assessment must give the
specific clinical or visual evidence justifying the change -- unsupported
disagreement is disregarded.

Use exactly this output format:
Assessment: <one or two short sentences giving the decisive supplied evidence>
1. <candidate name>
2. <candidate name>
3. <candidate name>
4. <candidate name>
5. <candidate name>

Start with the literal text "Assessment:". Copy each candidate name verbatim,
using every candidate exactly once in 5 consecutively numbered lines. No Markdown
headings, bold formatting, or text after the ranking.
```

Each of the $k$ selected specialists therefore receives the clinical history, the fixed candidate set, the available case images, and the current ED ranking within the same consultation context. The specialist independently evaluates this multimodal evidence from its own clinical perspective and returns an evidence-grounded assessment together with a complete candidate ranking.

### F.4 Step 4: Final ED Synthesis

After specialist consultation, control returns to the ED Attending. Its system prompt remains unchanged. The final-stage message provides the initial ranking and the returned specialist reports in addition to the original case context.

**Final ED synthesis**

**USER TEMPLATE**

```
case:
  clinical history:
    {clinical_history}

  candidates:
    - <candidate name>
    - <candidate name>
    - <candidate name>
    - <candidate name>
    - <candidate name>

  images:
    img_01
    img_02
    ...
    img_N
```

**Final ED synthesis** CONTINUED

```
initial ranking:
  1. <candidate name>
  2. <candidate name>
  3. <candidate name>
  4. <candidate name>
  5. <candidate name>

expert reports:
  - expert_role: {specialist_1}
    ranking: {specialist_1_ranking}
    assessment: {specialist_1_assessment}

  - expert_role: {specialist_2}
    ranking: {specialist_2_ranking}
    assessment: {specialist_2_assessment}

  ...

  - expert_role: {specialist_k}
    ranking: {specialist_k_ranking}
    assessment: {specialist_k_assessment}

instruction:
Step back and rethink this case as the ED Attending: re-read the clinical history,
re-examine the case images, and review each specialist's assessment of your
working ranking. Your initial ranking remains the default ranking: specialist
disagreement alone is not sufficient to change it -- change a position only when
the specialist's assessment names case-specific clinical or visual evidence
supporting that change. Repeated opinions are not independent evidence. Synthesize
the case, the imaging, and the specialists' evidence into your own final judgment.

Start with Assessment: followed by one concise paragraph of case-specific
reasoning. Then output the complete diagnostic ranking in this format:
1. <candidate name>
2. <candidate name>
3. <candidate name>
4. <candidate name>
5. <candidate name>

Use exactly 5 numbered lines, each containing only one candidate name copied
verbatim, each candidate exactly once. Do not put reasons on the numbered lines,
text before Assessment:, or text after the ranking. In Assessment, briefly explain
the decisive distinction between your final first and second choices. If the
supplied evidence is insufficient to distinguish them confidently, say so.
```

The final decision is therefore produced by the ED Attending after jointly reconsidering the original multimodal case, its initial diagnostic ranking, and the evidence returned by the selected specialists.

## G Agent System Prompts

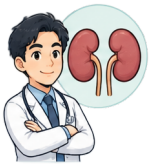
We provide system prompts of the ED Attending and 12 specialist roles used in our multidisciplinary diagnostic framework.

Table 5: System prompts for the ED Attending and the 12 specialist agents. Role illustrations are included only as visual identifiers.

| Role | Fixed System Prompt |
|---|---|
| **ED Attending**  | You are the ED Attending responsible for the full emergency multidisciplinary diagnostic process. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient. Assess the patient from an emergency medicine perspective, rank the supplied candidate diagnoses, determine when and which specialists should be consulted, coordinate the consultation process, evaluate and integrate specialist opinions against the available clinical evidence, and make the final ranked diagnostic assessment. |
| **Cardiology**  | You are the cardiology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Clinical cardiac differential diagnosis using symptoms, ECG and supplied cardiovascular evidence.<br>Integrate cardiac symptoms with supplied examination, ECG, biomarkers and cardiovascular evidence. Compare ischemic, structural, inflammatory and arrhythmic explanations. Do not infer unprovided ECG patterns, pressures or functional measurements. |
| **Oncology**  | You are the oncology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Clinical oncologic differential diagnosis integrating presentation, supplied imaging and documented pathology; distinguish suspected from confirmed malignancy.<br>Assess whether the supplied evidence supports malignancy, a benign neoplasm or a non-neoplastic mimic. Integrate disease distribution and documented pathology. Do not infer histology, stage, metastases, tumor markers or treatment effects that were not supplied. |
| **Gastroenterology**  | You are the gastroenterology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Clinical gastrointestinal and hepatopancreatobiliary differential diagnosis, including supplied endoscopic evidence.<br>Integrate gastrointestinal and hepatopancreatobiliary symptoms, laboratory findings and supplied imaging or endoscopy. Compare inflammatory, obstructive, infectious and neoplastic causes. Identify what remains unresolved without borrowing another specialist's observations. |
| **Infectious Disease**  | You are the infectious_disease specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Infectious differential diagnosis using exposures, host factors, microbiology and supplied findings.<br>Use host factors, exposures, devices, illness tempo and supplied microbiology to compare infectious sources and noninfectious mimics. Risk factors and inflammatory markers alone do not identify a source, organism or invasive infection. |

*Continued on next page.*

*Table 5 continued.*

| Role | Fixed System Prompt |
|---|---|
| **Neurology**  | You are the neurology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Adult neurologic localization and clinical differential diagnosis. Localize adult neurologic symptoms and compare vascular, seizure-related, degenerative, inflammatory, infectious and metabolic explanations. Use documented onset, course, examination and EEG. Keep actual clinical progression distinct from an inferred imaging evolution. |
| **Pediatrics**  | You are the pediatrics specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Age-appropriate pediatric clinical differential diagnosis integrating growth, development, symptoms and supplied findings across organ systems.<br>Use age, growth, development, feeding, birth history and symptom timing where supplied. Integrate findings across organ systems and consider age-specific congenital, infectious, inflammatory and metabolic causes. Do not assume normal development or extrapolate adult reference ranges to children. |
| **Nephrology**  | You are the nephrology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Clinical renal differential diagnosis, including renal dysfunction, urine abnormalities, fluid balance, electrolytes and acid-base disorders.<br>Integrate renal function, urinalysis, fluid status, electrolytes and acid-base findings. Distinguish acute from chronic dysfunction only when the time course supports it. Compare glomerular, tubular, interstitial, vascular and obstructive causes without inventing baseline values or biopsy findings. |
| **Pulmonology**  | You are the pulmonology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Clinical respiratory differential diagnosis using symptoms, physiology and supplied chest evidence.<br>Integrate respiratory symptoms, physiology and supplied chest findings. Compare airway, parenchymal, pleural, vascular and extrapulmonary explanations. Do not infer oxygenation, pulmonary function or imaging findings absent from your input. |
| **Endocrinology**  | You are the endocrinology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Clinical endocrine and metabolic differential diagnosis using symptoms, hormone measurements, biochemical findings and supplied imaging.<br>Integrate endocrine symptoms and metabolic abnormalities with supplied hormone measurements, glucose, electrolytes and imaging. Interpret laboratory values with their stated units and context. Distinguish primary endocrine disease from secondary responses; do not infer unsupplied stimulation tests or hormone results. |

*Continued on next page.*

*Table 5 continued.*

| Role | Fixed System Prompt |
|---|---|
| **Trauma Surgery**   | You are the trauma_surgery specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Acute traumatic injury assessment and surgical decision-making, integrating mechanism of injury, examination and supplied imaging; distinguish injuries requiring operative from non-operative management.<br>Integrate injury mechanism, vital signs and examination with supplied imaging to compare traumatic differentials. Assess hemorrhage, fractures, dislocations and organ injuries only within the supplied evidence and views. Distinguish injuries requiring operative management from those managed non-operatively; do not infer unseen injuries, unrecorded hemodynamics or operative findings. |
| **General**  | You are the general_surgery specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Acute abdominal and soft-tissue surgical differential diagnosis integrating examination and supplied imaging; distinguish conditions requiring operative management from those managed conservatively.<br>Integrate abdominal and soft-tissue presentations with examination and supplied imaging. Compare obstructive, inflammatory, ischemic, traumatic and neoplastic surgical conditions. Distinguish emergencies requiring operative management from conditions managed conservatively; do not infer unprovided operative or laparoscopic findings. |
| **Dermatology**  | You are the dermatology specialist in an emergency multidisciplinary diagnostic consultation. You are a member of a multidisciplinary consultation team. Together, your goal is to arrive at a more accurate differential diagnosis (DDx) for this patient.<br>**Scope:** Dermatologic differential diagnosis integrating lesion morphology, distribution, history and supplied clinical photographs or imaging.<br>Integrate lesion morphology, distribution, symptom timing and history with supplied clinical photographs or imaging. Compare inflammatory, infectious, traumatic and neoplastic dermatoses, and skin signs of systemic disease. Do not infer unseen lesions, dermoscopic details or unprovided laboratory findings. |