PASS: Probabilistic Agentic Supernet Sampling for Interpretable and Adaptive Chest X-Ray Reasoning

Yushi Feng[1], **Junye Du**[1], **Yingying Hong**[1], **Qifan Wang**[2], **Lequan Yu**[1*]

[1] School of Computing and Data Science, The University of Hong Kong, Hong Kong SAR, China
[2] Faculty of Engineering, The University of Hong Kong, Hong Kong SAR, China
{fengys@connect., junyedu@connect., yyhong@, wqf040701@connect., lqyu@}hku.hk

Abstract

Existing tool-augmented agentic systems are limited in the real world by (i) black-box reasoning steps that undermine trust of decision-making and pose safety risks, (ii) poor multimodal integration, which is inherently critical for healthcare tasks, and (iii) rigid and computationally inefficient agentic pipelines. We introduce **PASS** (**P**robabilistic **A**gentic **S**upernet **S**ampling), the first multimodal framework to address these challenges in the context of Chest X-Ray (CXR) reasoning. PASS adaptively samples agentic workflows over a multi-tool graph, yielding decision paths annotated with interpretable probabilities. Given the complex CXR reasoning task with multimodal medical data, PASS leverages its learned task-conditioned distribution over the agentic supernet. Thus, it adaptively selects the most suitable tool at each supernet layer, offering probability-annotated trajectories for post-hoc audits and directly enhancing medical AI safety. PASS also continuously compresses salient findings into an evolving personalized memory, while dynamically deciding whether to deepen its reasoning path or invoke an early exit for efficiency. To optimize a Pareto frontier balancing performance and cost, we design a novel three-stage training procedure, including expert knowledge warm-up, contrastive path-ranking, and cost-aware reinforcement learning. To facilitate rigorous evaluation, we introduce CAB-E, a comprehensive benchmark for multi-step, safety-critical, free-form CXR reasoning. Experiments across various benchmarks validate that PASS significantly outperforms strong baselines in multiple metrics (e.g., accuracy, LLM-Judge, semantic similarity, etc.) while balancing computational costs, pushing a new paradigm shift towards interpretable, adaptive, and multimodal medical agentic systems.

1 Introduction

Chest X-Ray is the most commonly performed diagnostic imaging procedure worldwide, widely regarded as a cornerstone of modern radiology (Johnson et al. 2019). However, interpreting CXRs demands careful multi-structure assessment that is time-consuming and expertise-intensive (Bahl, Ramzan, and Maraj 2020). While specialized AI tools for tasks like classification (Rajpurkar, Irvin, and Zhu 2017), segmentation (Ma et al. 2024) or report generation (Tanno and Barrett 2024; Chambon and Delbrouck 2024) etc. have shown promise in improving turnaround time and diagnostic consistency (Baltruschat et al. 2021; Ahn et al. 2022; Pham 2022; Shin 2023), their narrow specialization hinder their use in complex clinical reasoning scenarios (Erdal 2023; Fallahpour et al. 2024).

Large-scale foundation models (FMs) in recent years like GPT-4o (OpenAI 2024), LLaVA-Med (Li et al. 2023a), and CheXagent (Chen et al. 2024c) offer a more unified approach by integrating visual and textual reasoning. However, these monolithic systems often hallucinate (Eriksen, Möller, and Ryg 2024), lack domain-specific robustness (Chen et al. 2024c), and operate as uninterpretable "black boxes", making them unsuitable for high-stakes medical deployment.

Motivated by the need for more reliable, generalized, and autonomous solutions, recent efforts have explored *multi-agent medical AI systems* that coordinate domain-specific tools utilizing the capability of large language models (LLMs) and vision language models (VLMs). Recent progress in general-purpose agent systems (Li et al. 2023b; Wu et al. 2024; Zhuge et al. 2024) demonstrate the potential of collaborative LLM agents to outperform single-agent baselines through structured communication and role specialization (Du et al. 2023; Liang et al. 2024). Despite these advances, most systems rely on manually-defined and rigid workflows (Qian et al. 2025; Zhang et al. 2025b), which cannot adapt to the varying complexity of clinical queries and are computationally inefficient.

To address these challenges, recent methods have aimed to automate the design of multi-agent workflows. Works such as DsPy (Khattab et al. 2024) and EvoPrompt (Guo et al. 2024) optimize prompts, while G-Designer (Zhang et al. 2025a) and AutoAgents (Chen et al. 2024a) refine inter-agent communication and profiling strategies. In the medical domain, MedRAX (Fallahpour et al. 2025) exemplifies this direction by orchestrating multiple CXR tools via ReAct-style prompting (Yao et al. 2023), achieving improved accuracy over end-to-end models. However, these methods largely rely on black-box LLMs for the decision-making of invoking agents, leaving the concerns regarding trustworthiness and safety risks as open questions.

The most recent advance, agentic supernets like MaAS (Zhang 2025), introduced a paradigm shift by learning a distribution over possible workflows, enabling adap-

*Corresponding author

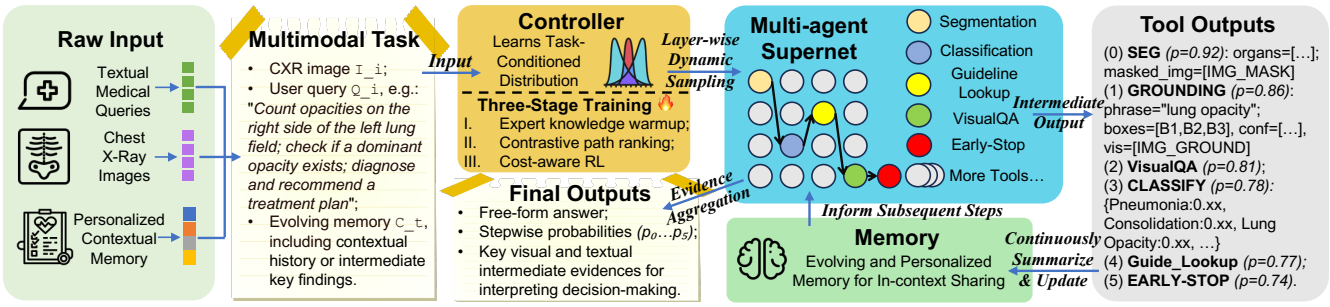

Figure 1: An overview of PASS. Given a multimodal complex reasoning task (CXR image, textual comprehensive query, multimodal personalized context), our probabilistic controller learns a continuous task-conditioned distribution over the agentic supernet (i.e. a directed acyclic graph of medical agent containers). At each step, it samples an action, yielding a workflow annotated with interpretable probabilities for post-audits and directly enhances clinical AI safety. Tool outputs, which can be both text and images, are summarized and fed into an evolving personalized memory and shared in-context to inform subsequent steps. The controller is trained via a principled three-stage strategy (expert knowledge warm-up, contrastive path ranking, cost-aware reinforcement learning) to optimize the accuracy-cost trade-off. Eventually, PASS is enabled to answer multimodal medical questions in free-form text via an interpretable, adaptive, and efficient agentic reasoning process.

tive, cost-aware reasoning. However, this approach has two fundamental flaws for medical applications. First, it is designed for text-only reasoning and lacks multimodal integration, which is inherently a core requirement in clinical reasoning. Second, while its textual gradient mechanism enables workflow optimization, it operates implicitly within the LLM's internal prompt space during multi-turn conversations, providing limited interpretability and traceability in high-stakes use.

These challenges highlight a critical need for a medical agentic system that is not only multimodal and truly interpretable, but also adaptive and efficient. To this end, we propose **PASS** (**P**robabilistic **A**gentic **S**upernet **S**ampling). To the best of our knowledge, PASS is the *first* framework for interpretable and adaptive CXR reasoning via multimodal agentic workflow sampling. Given a CXR image and a complex free-form clinical reasoning task, PASS manages an evolving contextual memory, operates over a directed acyclic graph consisting of multiple specialized medical agent containers (i.e., agentic supernet), and adaptively samples layer-wise tool sequences from the graph. Crucially, we design a Controller module to learn the task-conditioned continuous distribution over the supernet, yielding decision paths annotated with interpretable probabilities. This provides transparent trajectories for post-hoc audits, directly enhancing medical AI safety. We design a principled three-stage regimen for the training of PASS: (1) expert knowledge-guided warm-up aligns tool usage with clinical best practices; (2) contrastive path-ranking sharpens ordering preferences among tool sequences; and (3) cost-aware reinforcement learning trains the controller to learn the optimized accuracy-cost Pareto frontier with an early-exit mechanism.

To systematically evaluate such agentic systems, where existing CXR benchmarks largely focus on simplified classification or short-form QA and are thus poorly aligned with this paradigm, we introduce CHESTAGENTBENCH-E (CAB-E), a new challenging new benchmark comprising 2,550 comprehensive and safety-critical CXR reasoning cases annotated with free-form QA pairs, image inputs, and queries that demand highly complex rationales[1]. CAB-E expands the scope of prior evaluations (Fallahpour et al. 2025; Liu et al. 2021), emphasizing multi-step and clinically grounded queries that require adaptive tool orchestration. It also evaluates free-form answering and safety-critical cases.

Our key contributions can be summarized as follows:

- We propose PASS, the first framework to our knowledge to instantiate a probabilistic agentic supernet for multimodal medical reasoning, representing a paradigm shift towards building trustworthy, adaptive, transparent, and cost-aware agentic systems.
- We design a principled three-stage training strategy including expert knowledge guided warm-up, contrastive path ranking, and cost-aware reinforcement learning.
- We introduce CAB-E, a comprehensive public benchmark to evaluate multi-hop and safety-critical agentic reasoning for CXR with free-form answers.
- Extensive experiments validate that PASS outperforms strong baselines among various benchmarks, while maintaining the balanced computational cost and providing interpretable agentic workflows.

2 Methodology

In this paper, we propose a probabilistic framework for **PASS** that interprets workflow construction as a latent decision-making process governed by a multimodal generative policy. In this section, we first formulate a probabilistic controller over tool trajectories and answers, and derive a cost-aware objective grounded in expected utility maximization. We then introduce the architecture and parameterization of the controller π_θ, followed by a theoretically

[1]Code, data, and benchmarks are available for research purposes at https://github.com/ys-feng/PASS.

motivated multi-phase training algorithm that combines expert knowledge warm-up, contrastive path ranking and cost-aware reinforcement learning.

2.1 Preliminary

Problem formulation and notations. Let $\mathcal{Q} = \{(q_i, I_i, C_i)\}_{i=1}^{N}$ be a collection of *multimodal diagnostic queries*, where $q_i \in \mathcal{T}$ is a free-form text question, $I_i \in \mathbb{R}^{H\times W\times 3}$ is a chest X-ray image and $C_i \in \mathcal{C}$ denotes personalized contextual memory, including summarized information like structured demographic factors, clinical results, and previous analysis outputs. PASS answers q_i by sampling a *workflow* τ over a directed acyclic *multi-container graph* and executing the tools in corresponding containers in the selected sequence. We frame workflow generation as sampling from a probability distribution π_θ based on multimodal evidence:

$$\tau \sim \pi_\theta(\cdot \mid q, I, C), \quad \hat{a} = \textsc{Execute}(\tau) \tag{1}$$

where the *workflow* $\tau = (a_1, a_2, \ldots, a_T)$ is the trajectory of the actions and $\hat{a}$ is a free-form answer (e.g., finding, measurement, report section, etc.) returned to the clinician. PASS must simultaneously maximize diagnostic utility $\mathcal{U}$ and minimize a composite cost $\mathcal{L}$ capturing latency, token usage and privacy risk. The hyperparameter λ, configured by the user or deploying institution, controls the trade-off between performance and operational constraints. Under the above settings, the goal of our model could be stated as:

$$\max_{\theta}\ \mathbb{E}_{\substack{(q,I,C)\sim\mathcal{Q}\\ \tau\sim\pi_\theta}}\Big[\mathcal{U}(\hat{a}, a^\star) - \lambda\cdot\mathcal{L}(\tau)\Big] \tag{2}$$

Agentic supernet. Supernet $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ contains agent containers as nodes and legal tool invocations as edges. Each container $v \in \mathcal{V}$ is typed by one of SEGMENTATION, CLASSIFY, GROUNDING, REPORT, VQANALYZE, GUIDELINELOOKUP and MKG. The container v also stores a mutable set of tool models $T_v = \{t_{v,1}, \ldots\}$ that share identical I/O signature but may differ in backbone architecture, patch size or training epoch. The detailed tool model descriptions are in the Appendix[2]. Edges $e = (v \to v') \in \mathcal{E}$ are labeled with a routing policy ρ_e specifying which fields of the current memory are forwarded to the next container.

Formal Interface. Every container v adheres to a unified formal interface, defining its input x_v and output y_v as:

$$\begin{cases} x_v = \big(q^{(\mathrm{sub})},\, I^{(\mathrm{roi})},\, C^{(\mathrm{sub})},\, \eta\big) \\ y_v = \big(\rho_v,\, \ell_v,\, \kappa_v\big) \end{cases} \tag{3}$$

The input consists of a textual sub-query $q^{(\mathrm{sub})}$, an optional region-of-interest image tensor $I^{(\mathrm{roi})}$, a relevant slice of personalized contextual memory $C^{(\mathrm{sub})}$, and tool-specific hyperparameters η. The output comprises the *primary multimodal payload* ρ_v—which may be a JSON object for structured data (e.g., TEXT, BBOX, PROB) or an image tensor for visual data (e.g., a segmentation mask)—along with the measured *latency* ℓ_v and *token cost* κ_v, both utilized in the overall objective (Eq. (2)). By strictly enforcing this interface across all containers, our design ensures seamless plug-and-play integration and maintenance.

[2]Appendices are available at https://arxiv.org/abs/2508.10501

Algorithm 1: PASS: Training Procedure

Require: Expert demonstrations $\mathcal{D}_{\mathrm{exp}}$, unlabeled data $\mathcal{D}_{\mathrm{ul}}$, supernet $\mathcal{G}$, state encoder ψ, policy π_θ, answer generator p_ϕ, heuristic reward R_h, cost weights λ, entropy weight γ.

Procedure 1

 # Phase I: Expert Knowledge Warm-up
for $(s, a^\star) \in \mathcal{D}_{\mathrm{exp}}$ **do**
  $\theta \leftarrow \theta - \eta_1 \nabla_\theta\big(-\log \pi_\theta(a^\star \mid s)\big)$
end for
 # Phase II: Heuristic-Guided Path Ranking
for $(q, I, C) \in \mathcal{D}_{\mathrm{ul}}$ **do**
  $\{\tau_k\}_{k=1}^{K} \sim \pi_\theta(\cdot \mid q, I, C)$
  $p(\tau_k) \leftarrow \frac{\exp(R_h(\tau_k)/\alpha_{\mathrm{cpr}})}{\sum_{j=1}^{K}\exp(R_h(\tau_j)/\alpha_{\mathrm{cpr}})}$
  $\mathcal{L}_{\mathrm{CPR}} \leftarrow -\sum_{k=1}^{K} p(\tau_k)\log\pi_\theta(\tau_k)$
  Update θ *using* $\nabla_\theta\mathcal{L}_{\mathrm{CPR}}$
end for
 # Phase III: Cost-aware Reinforcement Learning
for $n = 1$ **to** N_{RL} **do**
  $\tau \sim \pi_\theta(\cdot \mid q, I, C) \in \mathcal{D}_{\mathrm{ul}}$
  $\hat{a} \sim p_\phi(\cdot \mid \tau, q, I, C)$
  $R(\tau) \leftarrow \mathcal{U}(\hat{a}, a^\star) - \lambda\mathcal{L}(\tau) - \gamma H(\hat{a})$
  $\theta \leftarrow \theta + \eta_3 R(\tau)\nabla_\theta \log\pi_\theta(\tau)$
end for

Action space. The space spanned by legal actions at state s_t is defined as:

$$\mathcal{A}(s_t) = \big\{(v, T_{v,k}) \mid (v_t \to v) \in \mathcal{E}\big\} \cup \{\textsc{EarlyExit}\}$$

where $(v, T_{v,k})$ denotes executing tool $T_{v,k}$ inside container v. Sampling the EARLYEXIT action, a special action that halts the execution trajectory early to conserve resources, thus initiating answer synthesis in advance.

2.2 Multi-agent Workflows

PASS models diagnostic reasoning as structured decision-making in a latent space of tool-based workflows. Given an input triplet (q, I, C), the agent sequentially builds a trajectory $\tau = (a_1, a_2, \ldots, a_T)$ by sampling actions $a_t \in \mathcal{A}(s_t)$, where s_t is the multimodal reasoning state at step t. The agent's final output is a multimodal package, consisting of a final textual answer $\hat{a} \in \mathcal{T}$ and any visual artifacts (e.g., annotated images) produced during the workflow τ.

The core of PASS is the workflow policy $\pi_\theta(a_t \mid s_t)$, which we aim to learn. This policy, combined with a fixed answer synthesis module p_ϕ, defines the full generative process for the textual answer $\hat{a}$:

$$\begin{aligned} &p_\theta(\tau, \hat{a} \mid q, I, C) \\ &= \underbrace{p_\phi(\hat{a} \mid \tau, q, I, C)}_{\text{Answer generator}} \cdot \underbrace{\prod_{t=1}^{T}\pi_\theta(a_t \mid s_t)}_{\text{Workflow policy}} \end{aligned} \tag{4}$$

Algorithm 2: PASS: Inference

Require: Policy π_θ, generator p_ϕ, summarizer $\mathcal{S}$, state encoder ψ, supernet $\mathcal{G}$, max steps $T_{\max}$.

Procedure 2

$M, \tau \leftarrow \emptyset, []$ *// Initialize memory and trajectory*
for $t = 1$ **to** $T_{\max}$ **do**
  $a_t \sim \pi_\theta(\cdot \mid \psi(q, I, C, M))$
  if $a_t = \text{EARLYEXIT}$ **then**
    break
  end if
  $\rho_t \leftarrow \text{EXECUTETOOL}(a_t)$
  $M \leftarrow M \cup \mathcal{S}(\rho_t)$
  $\tau \leftarrow \tau \cdot a_t$
end for
$\hat{a} \sim p_\phi(\cdot \mid q, I, C, \tau)$
RETURN $(\hat{a}, \tau)$ *// Return final answer and full workflow*

where p_ϕ is a frozen synthesis module (e.g., a large language model) responsible for generating the final text answer $\hat{a}$ based on the evidence gathered in τ. All learning is concentrated in the policy parameters θ, ensuring improvement stems from discovering better workflow decisions, not from fine-tuning the generator. This decomposition makes two key assumptions: (i) the tool sampling is a Markov process over the state space $\mathcal{S}$, and (ii) the final textual answer is conditionally independent of the internal policy decisions, given the full trajectory τ.

Policy-induced answer distribution. By virtue of marginalizing out the latent tool trajectory τ, we obtain the model-implied distribution over answers:

$$p_\theta(\hat{a} \mid q, I, C) = \sum_{\tau \in \mathcal{T}(q,I,C)} \pi_\theta(\tau \mid q, I, C) \cdot p_\phi(\hat{a} \mid \tau, q, I, C) \tag{5}$$

in which $\pi_\theta(\tau \mid q, I, C) = \prod_{t=1}^{|\tau|} \pi_\theta(a_t \mid s_t)$ and $\mathcal{T}(q, I, C)$ is the set of legal trajectories under $\mathcal{G}$ from initial state s_0. Although this marginal distribution is intractable to compute exactly due to the combinatorial size of $\mathcal{T}$, it can be approximated with Monte Carlo sampling, which we exploit both for training and for uncertainty estimation.

Expected utility and cost regularization. Given a ground-truth answer $a^\star$ and a reward function $\mathcal{U}(\hat{a}, a^\star)$ measuring the clinical utility of the predicted answer, our goal is to maximize the expected utility of our policy. This objective must also be balanced against the cost of the workflows it generates. Formally, the goal is to find optimal parameters θ for the policy π_θ:

$$\max_\theta \ \mathbb{E}_{(q,I,C)\sim\mathcal{Q}}\Big[\mathbb{E}_{\hat{a}\sim p_\theta(\cdot|q,I,C)}\mathcal{U}(\hat{a}, a^\star) - \lambda \cdot \mathbb{E}_{\tau\sim\pi_\theta(\cdot|q,I,C)}\mathcal{L}(\tau)\Big] \tag{6}$$

This formulation can be viewed as a constrained variational inference problem over the latent workflow τ with an amortized inference network π_θ.

Uncertainty-aware generation. The posterior entropy of the answer distribution, $H_\theta(\hat{a} \mid q, I, C) = -\mathbb{E}_{\hat{a}} \log p_\theta(\hat{a} \mid q, I, C)$, can be utilized to quantify the epistemic uncertainty of the model. Since the answer generator p_ϕ is frozen, this entropy is solely induced by the sampling variability in the workflow trajectory $\tau \sim \pi_\theta$. In practice, we estimate H_θ via Monte Carlo rollouts of the policy and use it both as a proxy for answer confidence and as a regulariser during policy learning (Sec. 2.4) to discourage high-entropy outputs in high-risk settings.

2.3 Controller Architecture

The controller $\pi_\theta(a_t \mid s_t)$ is designed as a masked categorical distribution over permissible actions, with its parameters determined by a state encoder ψ. Its logits are produced by a policy network head that processes the state representation h_t. Let $s_t = (q, I, C, M_t)$ denote the current multimodal state. The state encoder maps this input into a shared representation $h_t \in \mathbb{R}^d$:

$$h_t = \psi(s_t) = \text{LN}(z_t) \quad s.t. \quad z_t = W_I \cdot \xi(I) \parallel W_Q \cdot \zeta(q, C) \parallel W_M \cdot \mu(M_t) \tag{7}$$

where $\xi(I)$ is a frozen ViT-B/16 image encoder with final-layer CLS token projected to $\mathbb{R}^{256}$, $\zeta(q, C)$ is a Sentence-BERT-style text encoder for (q, C), projected to $\mathbb{R}^{128}$, $\mu(M_t)$ encodes dynamically updating memory over its summaries, with pooled final hidden state $\in \mathbb{R}^{128}$, $\text{LN}(\cdot)$ denotes layer normalization, and $\parallel$ denotes concatenation. The policy head is a feed-forward network with a single hidden layer and ReLU activation:

$$\pi_\theta(a_t \mid s_t) = \text{Softmax}\left(\text{mask}_{\mathcal{A}(s_t)}\left[W_2 \cdot \sigma(W_1 h_t)\right] / \alpha\right) \tag{8}$$

where the legal-action mask $\text{mask}_{\mathcal{A}(s_t)}$ zeroes out infeasible transitions in the supernet $\mathcal{G}$ and α is a temperature parameter annealed during training from 2.0 to 0.8.

Personalized contextual memory. At step t, what the controller observes are stated as:

$$s_t = \Big(q, I, C, M_t\Big), \quad M_t = \big\{(v_j, \tilde{y}_{v_j})\big\}_{j=1}^{t-1}$$

where the memory M_t is a bounded-size first-in-first-summarized (FIFS) buffer. After each tool call, in order to save the computational cost, the JSON response y_v is summarized to a compressed vector $\tilde{y}_v$ using a frozen language model prompted to function only as paraphrasing. These textual summaries are appended to a FIFO memory M_t along with image outputs (if any). This personalized and evolving memory mechanism enables precise, in-context diagnosis in the wild.

2.4 Three-Stage Training Procedure

We train the workflow policy π_θ to optimize the objective in Eq. (6) via a principled three-stage procedure. This curriculum-based approach progressively refines the policy, starting with strong expert supervision before moving to weaker preference signals and finally to direct reinforcement learning on the end-task reward. The three stages are detailed as follows. Each stage is grounded in a formal objective, allowing for stable and efficient training of π_θ.

Model	CAB-E							CAB-Standard		SLAKE	
	Acc.↑	LLM-J.↑	BLEU↑	METEOR↑	ROUGE-L↑	Sim.↑	Lat.↓	Acc.↑	Lat.↓	Sim.↑	Lat.↓
GPT-4o (zero-shot)	60.06 ± 0.01	45.29 ± 0.07	4.09 ± 0.03	25.63 ± 0.02	25.84 ± 0.01	79.03 ± 0.01	18,37	45.45 ± 0.02	3.10	37.25 ± 0.03	2.25
CoT	59.18 ± 0.01	39.43 ± 0.06	3.83 ± 0.03	23.93 ± 0.02	25.25 ± 0.01	77.62 ± 0.01	20.30	50.51 ± 0.02	3.34	38.78 ± 0.02	2.43
ComplexCoT	63.26 ± 0.01	41.06 ± 0.06	4.22 ± 0.04	25.14 ± 0.02	25.12 ± 0.02	78.03 ± 0.01	22.17	44.44 ± 0.01	3.41	42.86 ± 0.03	2.57
SC (CoT×5)	79.59 ± 0.08	54.13 ± 0.07	5.34 ± 0.01	31.22 ± 0.02	25.83 ± 0.01	76.14 ± 0.03	14.55	43.43 ± 0.02	10.35	44.88 ± 0.02	7.83
GPT-4o (finetuned)	81.82 ± 0.06	75.76 ± 0.02	**18.20** ± 0.01	32.92 ± 0.01	44.49 ± 0.02	88.19 ± 0.01	14.99	62.83 ± 0.01	3.79	81.82 ± 0.01	3.36
o3-mini (+visual tool)	73.73 ± 0.01	68.08 ± 0.04	4.43 ± 0.01	33.09 ± 0.01	24.52 ± 0.01	80.21 ± 0.02	41.91	50.51 ± 0.01	26.18	54.55 ± 0.01	11.63
CheXagent	83.67 ± 0.01	69.47 ± 0.01	2.71 ± 0.01	14.68 ± 0.01	20.78 ± 0.01	82.52 ± 0.01	**2.20**	62.63 ± 0.03	**0.40**	78.80 ± 0.01	**0.65**
LLaVA-Med	86.96 ± 0.05	82.65 ± 0.04	8.28 ± 0.01	29.96 ± 0.01	31.26 ± 0.01	**91.00** ± 0.01	21.43	53.23 ± 0.01	7.79	60.60 ± 0.01	10.14
MedRAX	89.54 ± 0.02	76.94 ± 0.01	5.56 ± 0.02	32.84 ± 0.05	27.11 ± 0.02	88.69 ± 0.02	17.44	63.49 ± 0.02	7.39	74.90 ± 0.02	10.47
PASS (Ours)	**91.22** ± 0.12	**84.28** ± 0.10	8.51 ± 0.05	**33.21** ± 0.05	**31.49** ± 0.09	90.16 ± 0.04	22.06	**66.10** ± 0.03	8.05	**79.55** ± 0.04	11.68

Table 1: Performance across three radiology VQA benchmarks (mean ± standard deviation). Best and runner-up numbers are bold and underlined.

Phase I: Expert knowledge guided warm-up. This initial phase uses imitation learning to bootstrap the policy. We construct a dataset of expert demonstrations, $\mathcal{D}_{\text{exp}}$, not from scratch, but by using a more scalable, two-step process. First, we use a powerful foundation model (GPT-4o) to generate initial workflow sketches for a set of problems. Second, these sketches are then reviewed, corrected, and validated in a human-in-the-loop process by licensed radiologists. This "distill-and-refine" strategy yields a high-quality dataset of one-step decisions $\mathcal{D}_{\text{exp}} = \{(s, a^\star)\}$, where $a^\star$ is the expert-verified action for state s. We warm-start the policy by minimizing the KL divergence from the expert policy (i.e., behavior cloning):

$$\mathcal{L}_{\text{BC}} = \mathbb{E}_{(s,a^\star)\sim\mathcal{D}_{\text{exp}}} \left[-\log \pi_\theta(a^\star \mid s) \right] \tag{9}$$

This phase instills a strong prior in the policy, anchoring it in clinically valid reasoning patterns.

Phase II: Heuristic-guided contrastive path ranking. Expert demonstrations are costly to acquire and cannot cover all scenarios. To generalize beyond $\mathcal{D}_{\text{exp}}$, we introduce a weaker supervisory signal based on heuristic preferences for unlabeled data. For a given query, we sample K candidate workflows $\{\tau_k\}_{k=1}^K$ from the current policy π_θ. We then score each path using a heuristic reward function, $R_h(\tau_k)$, which combines domain-specific priors such as clinical guideline compliance, anatomical coherence, and brevity. The policy is then updated using a contrastive loss (InfoNCE) that encourages it to assign higher probability to higher-scoring paths:

$$\begin{aligned} \mathcal{L}_{\text{CPR}} =& \mathbb{E}_{\{\tau_k\}\sim\pi_\theta} \left[-\sum_{k=1}^{K} p(\tau_k) \log \pi_\theta(\tau_k) \right], \\ \text{where } p(\tau_k) =& \frac{\exp(R_h(\tau_k)/\alpha_{\text{cpr}})}{\sum_{j=1}^{K} \exp(R_h(\tau_j)/\alpha_{\text{cpr}})} \end{aligned} \tag{10}$$

where α_{cpr} is a temperature hyperparameter. This phase teaches the policy to distinguish between good and bad reasoning structures, even without a ground-truth workflow.

Phase III: Cost-aware reinforcement learning. In the final phase, we directly fine-tune the policy π_θ using reinforcement learning to maximize the expected end-task utility. To compute the reward for a generated workflow τ, we first use the fixed answer generator p_ϕ to synthesize a textual answer, $\hat{a} \sim p_\phi(\cdot \mid \tau, q, I, C)$. We then define the reward for the trajectory as:

$$R(\tau) = \mathcal{U}(\hat{a}, a^\star) - \lambda \cdot \mathcal{L}(\tau) - \gamma \cdot H(\hat{a}) \tag{11}$$

where $H(\hat{a})$ is the entropy of generated answers, penalizing uncertainty. We then update the policy parameters θ using a reinforcement learning approach. The objective is to maximize the expected reward over all trajectories sampled from the policy:

$$J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau)] \tag{12}$$

The gradient of this objective, $\nabla_\theta J(\theta)$, can be estimated using sampling via the reinforcement algorithm, with a baseline to reduce variance. This final tuning step aligns the workflow generation directly with the ultimate goals of diagnostic accuracy and computational efficiency.

3 Experiments

We evaluate PASS across three radiology benchmarks of increasing complexity to assess four critical aspects of real-world deployment: clinical accuracy, language fidelity, computational efficiency, and safety. All experiments are conducted on a single NVIDIA H800 (80GB) GPU with access to OpenAI's GPT API for relevant baselines.

3.1 Experiment Setup

Benchmarks. We use the following evaluation suites, with more details described in the Appendices:

- **SLAKE** (Liu et al. 2021): A native free-form medical VQA benchmark with 6,437 image–question pairs, used to assess zero-shot generalization.
- **CAB-Standard** (Fallahpour et al. 2025): A multiple-choice Chest Agent Benchmark (CAB) containing 2,500 diagnostic queries. CAB-Standard is constructed using the generation method proposed by Fallahpour et al.
- **CAB-E:** Our proposed benchmark with 2,550 multi-step CXR reasoning cases, including 500 safety-critical instances. Construction details and summary statistics are provided in Appendices A and F. This benchmark is designed to evaluate free-form, multi-hop reasoning grounded in both imaging data and patient context. The

Model	Acc. ↑	Hallucination (%)↓
GPT-4o (zero-shot)	61.22	7.00
LLaVA-Med	87.75	2.00
MedRAX	89.79	1.60
PASS	93.50	1.60

Table 2: Performance on radiologist-verified safety-critical split from CAB-E.

Configuration	Acc.	ΔCost
Full PASS	91.22	-
– EarlyExit	88.60	94.0
– Path-Rank Pretraining	87.86	8.9
– Expert-Guided Warm-up	88.89	9.5

Table 3: Ablation study on CAB-E. ΔCost reports cost decrease relative to full PASS.

safety-critical subset focuses on complex, high-stakes scenarios that demand careful and transparent decision-making, such as life-threatening anatomical abnormalities and urgent systemic conditions. CAB-E is publicly available at the aforementioned URL.

Metrics. On CAB-E, we report: Accuracy, LLM-as-a-Judge score (LLM-J.) based on human expert-guided rubrics, BLEU, METEOR, ROUGE-L, embedding similarity, and end-to-end latency. CAB-Standard is evaluated by accuracy and latency. SLAKE is evaluated by embedding similarity and latency. We evaluate the hallucination rate on the safety-critical split of CAB-E, report blind human radiologist evaluation, and compare the inference cost against LLM-J. to assess the models' efficiency. We present detailed descriptions of the metrics in Appendix B.

Baselines. We compare PASS against four groups of methods: **(1) general-purpose VLMs**, including GPT-4o (OpenAI 2024), the finetuned version of GPT-4o on the same training data of PASS, and its reasoning-augmented variants CoT (Wei et al. 2022), ComplexCoT (Fu et al. 2023), and SC (CoT×5) (Wang et al. 2023); **(2) reasoning-centric VLMs**, o3-mini (OpenAI 2025) paired with LLaVA-Med (Li et al. 2023a) as a visual captioning front-end due to its lack of image input; **(3) medical/CXR-specialized VLMs**, LLaVA-Med and CheXagent (Chen et al. 2024c); and **(4) agentic systems**, including the multimodal system MedRAX (Fallahpour et al. 2025) and originally single-modality methods (e.g., MaAS (Zhang 2025), AFlow (Zhang et al. 2025b)), which we adapt to the multimodal setting by augmenting them with the same vision tools as PASS, with detailed results for these adapted agents reported in Appendix E.

Implementation details. We optimize the model using the AdamW algorithm, incorporating gradient clipping at 1.0 to ensure numerical stability, a weight decay of 0.01 to prevent overfitting, and a cosine learning rate schedule to facilitate smooth convergence. An entropy bonus of 0.01 is applied to encourage exploration and stabilize training. For RL updates, we employ forward-mode unrolling with a 5-step truncation to balance computational efficiency and gradient accuracy.

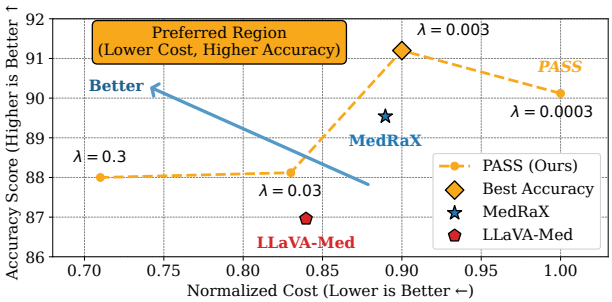

Figure 2: Cost-Accuracy Pareto Frontier analysis. Each orange point on the dashed frontier corresponds to a specific penalty weight (λ) configuration of PASS, enabling flexible cost–accuracy trade-offs at deployment. MedRAX and LLaVA-Med are plotted as additional points for comparison. Lower normalized inference cost and higher accuracy are preferred; the arrow indicates the desired direction toward the top-left preferred region.

3.2 Performance Analysis

Table 1 presents the results on CAB-E, CAB-Standard, and SLAKE. PASS achieves an accuracy of 91.22, outperforming the strongest baseline MedRAX (89.54) by +1.68, surpassing CheXagent by +7.55 and LLaVA-Med by +4.26, demonstrating substantial improvement in diagnostic accuracy through probabilistic multi-tool reasoning. This suggests that adaptively sampled agentic trajectories, rather than single-pass VLMs or black-box agent planners, offer superior coverage and reliability on diverse CXR cases. We also observed that a specific version of GPT-4o that is finetuned on the same training dataset of PASS lags behind PASS, suggesting that probabilistic, query-dependent tool trajectories are the key factor, not merely domain-specific training.

PASS also achieves the highest LLM-J. score (84.28), METEOR (33.21), ROUGE-L score (31.49), and the second-best BLEU (8.51) among all strong baselines. This indicates that the answers provided by PASS align better with ground truth clinical solutions, validating the controller's ability to coordinate image grounding, clinical reasoning, and textual fluency across multi-hop tool outputs.

3.3 Latency and Cost Analysis.

Table 1 shows that while PASS exhibits higher latency than single-pass models like LLaVA-Med, this is a direct and strategic trade-off for its superior accuracy, driven by a more comprehensive reasoning process. Figure 2 illustrates the empirical cost–accuracy Pareto frontier of PASS by varying the penalty weight λ, where the x-axis denotes normalized inference cost (relative to $\lambda = 0.0003$) and the y-axis reports accuracy. As λ increases, PASS traverses a smooth frontier that substantially reduces cost with only modest accuracy degradation, exposing multiple deployment-ready operating points. The highest accuracy (91.2%) is achieved at an intermediate setting $\lambda = 0.003$, where PASS outperforms

MedRAX and LLaVA-Med by 1.66 and 4.24 absolute accuracy points, respectively, at comparable cost. For more aggressive cost-saving, larger λ values (e.g., $\lambda = 0.03$) further reduce cost by roughly 20% while still retaining around 88% accuracy. Overall, PASS learns a well-structured frontier, enabling practitioners to tune λ at deployment time to match latency and budget constraints without retraining.

3.4 Safety-Critical Subset Evaluation

On this safety-critical CAB-E subset, PASS achieves an accuracy of 93.50%, surpassing MedRAX by 3.71 percentage points and LLaVA-Med by 5.75 percentage points. Notably, PASS and MedRAX share the lowest hallucination rate, representing a substantial improvement over the GPT-4o baseline and highlighting PASS's robustness in minimizing errors on safety-critical CXR cases. A blind human radiologist review further corroborates the superiority of PASS, with details provided in the Appendix. Taken together, these results underscore PASS's reliability in safety-critical clinical reasoning scenarios.

3.5 Ablation Study

Ablation results (Table 3) confirm critical design choices: Removing early-exit causes a significant accuracy drop (from 91.22 to 88.60) and a 94% relative cost decrease. Removing path-rank pretraining and warm-up also demonstrates their role in convergence acceleration and performance improvements.

4 Related Work

Tool-augmented LLMs. Tool use in LLMs has evolved from basic augmentation (Schick et al. 2023; Yao et al. 2023; Feng et al. 2025b) to modular agent frameworks (Wu et al. 2024; Li et al. 2023b; Chen et al. 2024b; Zhuge et al. 2024) with specialized roles and communication. Yet, most rely on static or handcrafted workflows, limiting adaptability and efficiency in real-world deployment. Recent work begins to automate tool strategies and workflows via reinforcement learning or structured search (Feng et al. 2025a; Zhang et al. 2025b), but typically commits to a single, task-agnostic pipeline and offers little support for uncertainty-aware or dynamically adaptive inference.

Autonomous agent workflows. Recognizing the limitations of fixed pipelines, a new wave of research seeks to automate agentic system design. Prompt optimization (Khattab et al. 2024; Guo et al. 2024), inter-agent communication tuning (Zhang et al. 2025a), and modular profiling (Chen et al. 2024a) are key directions. Notably, MaAS (Zhang 2025) introduces an agentic supernet that learns a distribution over multi-agent architectures and samples query-dependent workflows, improving accuracy–cost trade-offs and transferability beyond static designs. However, these approaches remain largely confined to text-only domains and offer limited interpretability and explicit uncertainty modeling, which is particularly problematic in high-stakes applications such as medicine.

Multimodal reasoning in medical AI. Multimodal foundation models (e.g., GPT-4V (Liu et al. 2024b), LLaVA-Med (Li et al. 2023a), CheXagent (Chen et al. 2024c)) promise unified vision-language understanding and have shown zero-shot capabilities across radiological tasks. Still, they often hallucinate (Eriksen, Möller, and Ryg 2024), lack task specificity (Chen et al. 2024c), and remain opaque. Domain-specific systems like MedRAX (Fallahpour et al. 2025) and MDAgents (Kim et al. 2024) attempt to integrate medical tools with LLMs via ReAct-style (Yao et al. 2023) prompting, offering partial medical multimodal reasoning capabilities. Yet, their decision-making still largely relies on black-box LLMs, hindering real-world application due to critical concerns about trust and potential risks.

Safety and interpretability in clinical deployment. Clinical settings demand more than performance: they require transparency, controllability, and regulatory compliance (Lundervold and Lundervold 2019). Beyond saliency-based explanations, methods like MedCoT (Liu et al. 2024a) and BoxMed-RL (Jing et al. 2025) leverage chain-of-thought or RL-enhanced generation to increase reliability. PASS extends these efforts with per-step, probability-annotated execution traces and interpretable early exits, allowing for post-hoc audits and fine-grained trust calibration, which are crucial features for safe medical AI deployment.

5 Conclusion

In this paper, we introduce PASS, the first multimodal framework to address the critical challenges of interpretability, adaptability, and efficiency in complex chest X-ray reasoning. Existing agentic systems are often limited by their black-box nature, poor integration of multimodal data, and rigid, inefficient workflows. PASS overcomes these limitations by leveraging a probabilistic controller to adaptively sample workflows from a multi-tool supernet, yielding decision paths annotated with transparent probabilities that are crucial for clinical trust and post-hoc audits. Our novel three-stage training strategy performs expert knowledge warm-up, contrastive path-ranking, and cost-aware reinforcement learning to optimize the performance-cost trade-off, balancing diagnostic accuracy with computational cost via a dynamic early-exit mechanism. Through extensive experiments on our newly curated CAB-E and other public benchmarks, we have demonstrated that PASS not only achieves superior accuracy over strong baselines but also provides interpretable and efficient reasoning. Ultimately, we believe that PASS represents a paradigm shift towards the next generation of multimodal, trustworthy, adaptive, and resource-aware agentic systems, grounded in medical reasoning yet potentially broadly applicable to other multimodal or high-stakes domains.

Limitations. PASS deliberately uses a fixed container set to ensure clinical safety and interpretability as a strategic trade-off between safety and flexibility. Future works will scale this robust foundation by expanding the supernet to new imaging types like MRI or CT, and enriching its agentic containers and tools, thereby further enhancing the diagnostic utility and adaptability of PASS.

Acknowledgements

This work was supported in part by the Research Grants Council of Hong Kong (27206123, 17200125, C5055-24G, and T45-401/22-N), the Hong Kong Innovation and Technology Fund (GHP/318/22GD), the National Natural Science Foundation of China (No. 62201483), and Guangdong Natural Science Fund (No. 2024A1515011875).

References

Ahn, J. S.; Ebrahimian, S.; McDermott, S.; Lee, S.; Naccarato, L.; Di Capua, J. F.; Wu, M. Y.; Zhang, E. W.; Muse, V.; Miller, B.; et al. 2022. Association of artificial intelligence–aided chest radiograph interpretation with reader performance and efficiency. *JAMA Network Open*, 5(8): e2229289–e2229289.

Bahl, S.; Ramzan, T.; and Maraj, R. 2020. Interpretation and documentation of chest X-rays in the acute medical unit. *Clinical Medicine*, 20(2): s73.

Baltruschat, I.; Steinmeister, L.; Nickisch, H.; Saalbach, A.; Grass, M.; Adam, G.; Knopp, T.; and Ittrich, H. 2021. Smart chest X-ray worklist prioritization using artificial intelligence: a clinical workflow simulation. *European radiology*, 31(6): 3837–3845.

Chambon, P.; and Delbrouck, J. e. 2024. CheXpert Plus: Augmenting a Large Chest X-Ray Dataset with Text Radiology Reports, Patient Demographics and Additional Image Formats. *arXiv preprint arXiv:2405.19538*.

Chen, G.; Dong, S.; Shu, Y.; Zhang, G.; Sesay, J.; Karlsson, B.; Fu, J.; and Shi, Y. 2024a. AutoAgents: A Framework for Automatic Agent Generation. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, 22–30. ijcai.org.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Chan, C.; Yu, H.; Lu, Y.; Hung, Y.; Qian, C.; Qin, Y.; Cong, X.; Xie, R.; Liu, Z.; Sun, M.; and Zhou, J. 2024b. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. In *The Twelfth International Conference on Learning Representations*. OpenReview.net.

Chen, Z.; Varma, M.; Xu, J.; Paschali, M.; Veen, D. V.; Johnston, A.; Youssef, A.; Blankemeier, L.; Bluethgen, C.; Altmayer, S.; Valanarasu, J. M. J.; Muneer, M. S. E.; Reis, E. P.; Cohen, J. P.; Olsen, C.; Abraham, T. M.; Tsai, E. B.; Beaulieu, C. F.; Jitsev, J.; Gatidis, S.; Delbrouck, J.-B.; Chaudhari, A. S.; and Langlotz, C. P. 2024c. A Vision-Language Foundation Model to Enhance Efficiency of Chest X-ray Interpretation. arXiv:2401.12208.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 40th International Conference on Machine Learning*.

Erdal, B. S. e. 2023. Integration and Implementation Strategies for AI Algorithm Deployment with Smart Routing Rules and Workflow Management. *arXiv preprint arXiv:2311.10840*.

Eriksen, A. V.; Möller, S.; and Ryg, J. 2024. Use of GPT-4 to Diagnose Complex Clinical Cases. *NEJM AI*, 1(1): 2300031.

Fallahpour, A.; Alinoori, M.; Ye, W.; Cao, X.; Afkanpour, A.; and Krishnan, A. 2024. EHRMamba: Towards Generalizable and Scalable Foundation Models for Electronic Health Records. In *Machine Learning for Health, ML4H@NeurIPS 2024*, volume 259 of *Proceedings of Machine Learning Research*, 291–307. PMLR.

Fallahpour, A.; Ma, J.; Munim, A.; Lyu, H.; and Wang, B. 2025. MedRAX: Medical Reasoning Agent for Chest X-ray. In *Proceedings of the 42nd International Conference on Machine Learning*.

Feng, J.; Huang, S.; Qu, X.; Zhang, G.; Qin, Y.; Zhong, B.; Jiang, C.; Chi, J.; and Zhong, W. 2025a. Retool: Reinforcement learning for strategic tool use in llms. *arXiv preprint arXiv:2504.11536*.

Feng, Y.; Chan, T. H.; Yin, G.; and Yu, L. 2025b. Democratizing large language model-based graph data augmentation via latent knowledge graphs. *Neural Networks*, 191: 107777.

Fu, Y.; Peng, H.; Sabharwal, A.; Clark, P.; and Khot, T. 2023. Complexity-Based Prompting for Multi-step Reasoning. In *The Eleventh International Conference on Learning Representations*. OpenReview.net.

Guo, Q.; Wang, R.; Guo, J.; Li, B.; Song, K.; Tan, X.; Liu, G.; Bian, J.; and Yang, Y. 2024. Connecting Large Language Models with Evolutionary Algorithms Yields Powerful Prompt Optimizers. In *The Twelfth International Conference on Learning Representations*. OpenReview.net.

Jing, P.; Lee, K.; Zhang, Z.; Zhou, H.; Yuan, Z.; Gao, Z.; Zhu, L.; Papanastasiou, G.; Fang, Y.; and Yang, G. 2025. Reason Like a Radiologist: Chain-of-Thought and Reinforcement Learning for Verifiable Report Generation. *arXiv preprint arXiv:2504.18453*.

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1): 317.

Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T. B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; and Amodei, D. 2020. Scaling Laws for Neural Language Models. *CoRR*, abs/2001.08361.

Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; Zaharia, M.; and Potts, C. 2024. DSPy: Compiling Declarative Language Model Calls into State-of-the-Art Pipelines. In *The Twelfth International Conference on Learning Representations*. OpenReview.net.

Kim, Y.; Park, C.; Jeong, H.; Chan, Y. S.; Xu, X.; McDuff, D.; Lee, H.; Ghassemi, M.; Breazeal, C.; and Park, H. W. 2024. Mdagents: An Adaptive Collaboration of LLMs for Medical Decision-Making. *Advances in Neural Information Processing Systems*, 37: 79410–79452.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2023a. Llava-Med: Training A Large Language-and-Vision Assistant for Biomedicine in One Day. *Advances in Neural Information Processing Systems*, 36: 28541–28564.

Li, G.; Hammoud, H.; Itani, H.; Khizbullin, D.; and Ghanem, B. 2023b. Camel: Communicative Agents For "Mind" Exploration of Large Language Model Society. *Advances in Neural Information Processing Systems*, 36: 51991–52008.

Liang, T.; He, Z.; Jiao, W.; Wang, X.; Wang, Y.; Wang, R.; Yang, Y.; Shi, S.; and Tu, Z. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17889–17904. Association for Computational Linguistics.

Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; and Wu, X.-M. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, 1650–1654. IEEE.

Liu, J.; Wang, Y.; Du, J.; Zhou, J.; and Liu, Z. 2024a. MedCoT: Medical Chain of Thought via Hierarchical Expert. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 17371–17389.

Liu, Y.; Li, Y.; Wang, Z.; Liang, X.; Liu, L.; Wang, L.; Cui, L.; Tu, Z.; Wang, L.; and Zhou, L. 2024b. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. *Meta-Radiology*, 2(4): 100099.

Lundervold, A. S.; and Lundervold, A. 2019. An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift fuer medizinische Physik*, 29(2): 102–127.

Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; and Wang, B. 2024. Segment anything in medical images. *Nature Communications*, 15(1): 654.

OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276.

OpenAI. 2025. OpenAI o3-mini System Card. Technical report, OpenAI.

Pham, H. H. e. 2022. An Accurate and Explainable Deep Learning System Improves Interobserver Agreement in the Interpretation of Chest Radiograph. *IEEE Access*, 10: 104512–104531.

Qian, C.; Xie, Z.; Wang, Y.; Liu, W.; Zhu, K.; Xia, H.; Dang, Y.; Du, Z.; Chen, W.; Yang, C.; Liu, Z.; and Sun, M. 2025. Scaling Large Language Model-based Multi-Agent Collaboration. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net.

Rajpurkar, P.; Irvin, J.; and Zhu, K. e. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Hambro, E.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Advances in Neural Information Processing Systems*, 36: 68539–68551.

Shin, H. J. e. 2023. The Impact of Artificial Intelligence on the Reading Times of Radiologists for Chest Radiographs. *NPJ Digital Medicine*, 6: 82.

Tanno, R.; and Barrett, D. G. e. 2024. Collaboration between Clinicians and Vision–Language Models in Radiology Report Generation. *Nature Medicine*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *The Eleventh International Conference on Learning Representations*. OpenReview.net.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems*, 35: 24824–24837.

Wu, Q.; Bansal, G.; Zhang, J.; Wu, Y.; Li, B.; Zhu, E.; Jiang, L.; Zhang, X.; Zhang, S.; Liu, J.; et al. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Zhang, G.; Yue, Y.; Sun, X.; Wan, G.; Yu, M.; Fang, J.; Wang, K.; Chen, T.; and Cheng, D. 2025a. G-Designer: Architecting Multi-Agent Communication Topologies via Graph Neural Networks. In *ICLR 2025 Workshop on Foundation Models in the Wild*.

Zhang, G. e. 2025. Multi-Agent Architecture Search via Agentic Supernet. *arXiv preprint arXiv:2502.04180*.

Zhang, J.; Xiang, J.; Yu, Z.; Teng, F.; Chen, X.; Chen, J.; Zhuge, M.; Cheng, X.; Hong, S.; Wang, J.; Zheng, B.; Liu, B.; Luo, Y.; and Wu, C. 2025b. AFlow: Automating Agentic Workflow Generation. In *The Thirteenth International Conference on Learning Representations*. OpenReview.net.

Zhuge, M.; Wang, W.; Kirsch, L.; Faccio, F.; Khizbullin, D.; and Schmidhuber, J. 2024. Gptswarm: Language agents as optimizable graphs. In *Proceedings of the 41st International Conference on Machine Learning*.