DIAGNO-SV: DECOUPLING PERCEPTION, REASONING, AND KNOWLEDGE FOR TRUSTWORTHY MEDICAL VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

While Large Vision-Language Models (VLMs) show immense promise in emulating the complex cognitive processes of clinical diagnostics, ensuring the reliability of their multimodal reasoning remains a critical and unsolved challenge. Current Chain-of-Thought approaches in VLMs typically conflate visual perception, logical deduction, and medical knowledge, creating an opaque "black box." This ambiguity fosters a critical failure mode we identify as **retrospective forgery**, a phenomenon where models fabricate visual findings or distort facts post-hoc to justify flawed diagnostic hypotheses. To address this, we introduce **Diagno-SV** (**Diagno**sis-**S**eparation for medical **V**LM), a novel framework for diagnostic-oriented optimization via trilevel decoupling. Diagnosis-SV forces the model to decompose its reasoning into independently verifiable steps of visual perception, logical deduction, and medical knowledge. Building on this transparency, we design a curriculum-based diagnostic refinement workflow. A dedicated multimodal diagnostic module pinpoints errors, distinguishing between "looking wrong" and "thinking wrong", and constructs preference pairs for Direct Preference Optimization. The resulting Diagno-SV model achieves state-of-the-art performance across six medical VQA benchmarks, including MMMU-Med and PMC-VQA. More importantly, our method significantly mitigates retrospective forgery, reducing the rate of visual and factual fabrication from 42.74% to 26.32%, paving a new path toward trustworthy visual-based medical AI.

1 INTRODUCTION

In recent years, Large Vision-Language Models (VLMs) have demonstrated significant potential in emulating the multimodal cognitive processes of clinicians, integrating visual interpretation with diagnostic decision-making Moor et al. (2023b); Li et al. (2023a). An ideal medical VLM must possess not only accurate visual perception and rigorous logical reasoning, but also precise recall of medical knowledge to navigate the uncertainty inherent in clinical scenarios Moor et al. (2023a); Singhal et al. (2023). However, in most contemporary medical VLMs, the internal processes of perception, logic, and knowledge are deeply entangled and inseparable. This multimodal "black-box" nature stands as a core impediment to building trustworthy medical AI.

This opacity not only undermines the trustworthiness of their conclusions but also conceals intertwined failure modes across modalities Dang et al. (2024). Among these, we identify a particularly subtle and dangerous phenomenon, which we define as **retrospective forgery**. Unlike simple hallucinations, this phenomenon occurs when, after making a flawed logical leap or hypothesis, the model "retrospectively" fabricates visual descriptions or distorts medical facts to make its Chain-of-Thought (CoT) appear plausible. For instance, to support an incorrect diagnosis of "pneumonia", a model might hallucinate "consolidations" in a clear chest X-ray. Within a traditional, monolithic reasoning output, such perceptual and factual fabrications are seamlessly integrated with the logical flow, rendering them nearly undetectableRajaraman et al. (2025). This enables an even more dangerous outcome: **fortuitous correctness**, where a model arrives at the correct answer through a path containing distorted visual perception or factual inaccuracies Geirhos et al. (2020). Optimizing the model based on such "correct" outcomes reinforces these dishonest reasoning patterns, making the model vulnerable in high-stakes clinical applications Amodei et al. (2016).

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.

To break this impasse, we argue for a paradigm shift from coarse-grained supervision to a **fine-grained, diagnostic intervention** into the multimodal mechanisms, as shown in Figure 1. To this end, we introduce **Diagno-SV** (Diagnosis-Separation for VLM), an innovative decomposed reasoning framework designed to decouple visual perception, reasoning logic, and medical knowledge. Serving as a powerful diagnostic mechanism, it compels the model to explicit decompose its thought process into three independently verifiable components: a **visual perception step** (what I see), a **logical deduction step** (what I infer), and a **medical knowledge recall step** (what I know). A key contribution of this design is its ability to render the model's internal cognitive path transparent. This transparency makes long-hidden issues explicit, analyzable, and correctable, specifically the causality between logical errors and visual or knowledge fabrications.

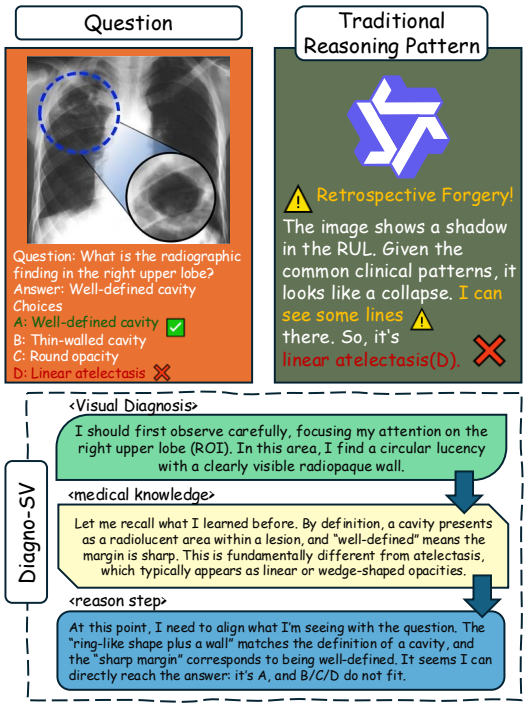


Figure 1: Overview of Diagno-SV against traditional VLM reasoning.

Building on this diagnostic framework, we introduce a complete diagnostic-oriented optimization workflow. First, as a cold-start preparation, the model is fine-tuned on high-quality, decomposed trajectory data generated by a multi-agent system, which now includes a dedicated **visual perceptor** agent to ground reasoning in pixel-level evidence. This adapts the model to the structured "Perception-Reason-Knowledge" thinking pattern. Second, the model enters an iterative refinement loop guided by curriculum learning. Within this loop, a core diagnostic module performs a **tri-pronged check**: it scrutinizes the faithfulness of the "visual perception" against the image, evaluates the coherence of the "reason step", and cross-validates the "medical knowledge". The flaws pinpointed by this module, including visual misinterpretations and knowledge hallucinations, are formulated into preference pairs. The model is then fine-tuned via Direct Preference Optimization (DPO) Rafailov et al. (2023), allowing it to simultaneously learn how to see faithfully, think rigorously, and recall knowledge honestly.

Experiments validate the effectiveness of Diagno-SV. The model achieves new state-of-the-art (SOTA) performance across six multimodal medical benchmarks, including MMMU-Med, PMC-VQA, and OmniMedVQA, on both in-distribution and challenging out-of-distribution (OoD) tasks. More importantly, our diagnostic-oriented optimization mitigates retrospective forgery, reducing the forgery rate from 42.74% to 26.32%—direct evidence that Diagno-SV aligns its visual perception and reasoning in a more reliable and "honest" manner.

Our main contributions are threefold:

1. We propose **Diagno-SV**, a novel framework that achieves **tri-level decoupling** of visual perception, logical deduction, and knowledge recall in VLMs, transforming the opaque multimodal "black box" into a transparent, diagnosable process.

2. We identify and define **retrospective forgery** in the context of VLMs, revealing how models often hallucinate visual features to justify flawed reasoning logic. We provide a theoretical perspective on how this connects to fortuitous correctness.

3. We introduce a complete diagnostic-oriented optimization workflow that integrates a visual-logic-knowledge diagnostic module with DPO. This approach yields SOTA performance on extensive medical VLM benchmarks while significantly reducing the rate of reasoning and perceptual fabrication.

2 Related Works

2.1 Medical Vision-Language Models

Early adaptations like LLaVA-Med Li et al. (2023a) and Med-Flamingo Moor et al. (2023b) transferred general VLM capabilities to the biomedical domain through continual pre-training and instruction tuning on image-text pairs. More recent unified models, such as Qwen2-VL Wang et al. (2024) and InternVL Chen et al. (2024b), achieve impressive performance on standard benchmarks by scaling visual encoders and training data, while works like HuatuoGPT-V Chen et al. (2024a) add domain-specific Supervised Fine-Tuning (SFT) data. However, despite boosting overall accuracy, these methods structurally treat multimodal reasoning as a holistic "black box": they map pixels and text directly to answers without explicit intermediate steps. Consequently, when errors occur, it is impossible to determine whether the failure stems from visual misperception (e.g., missing a nodule), logical fallacy, or medical knowledge deficiency, severely limiting reliability in high-stakes clinical applications.

2.2 Reasoning and Hallucination in VLMs

Recent research improves VLM reliability by enhancing reasoning transparency and alignment. One line of work applies Reinforcement Learning from Human/AI Feedback (RLHF/RLAIF), such as LLaVA-RLHF Sun et al. (2024) and Silkie Li et al. (2023b), using Direct Preference Optimization (DPO) to penalize hallucinations at the response level. Another explores chain-of-thought (CoT) prompting or multi-turn reasoning to elicit better answers Zhang et al. (2024). However, these approaches operate at the outcome level or focus solely on the language modality, and fail to decouple the *visual perception process* from the *logical reasoning process*: a model penalized for a wrong diagnosis via DPO receives an ambiguous gradient that does not reveal whether it "looked" at the wrong place or "thought" wrongly. In contrast, **Diagno-SV** introduces a tri-level decoupling framework (Perception, Reason, Knowledge) as a prerequisite for diagnostic-oriented optimization. By separating visual grounding from logical deduction and knowledge recall, our "Visual Diagnosis" and "Reasoning Diagnosis" modules pinpoint and correct the root cause of failures, enabling fine-grained, interpretable enhancement of medical VLM reasoning.

3 Methodology

To address the challenge of opacity and hallucination in medical VLM,Diagno-SV aims to decompose diagnostic reasoning as illustrated in Figure 2. The core philosophy of Diagno-SV replaces "confused and verbose reasoning" with a "perceive-reason-recall" cognitive loop built on three aligned pillars: faithful visual perception, rigorous logical deduction, and accurate medical knowledge. Inspired by prior work Xu et al. (2025); DeepSeek-AI (2025), our methodology has two stages: first, multi-agent collaboration constructs high-quality tri-level decomposed trajectories from carefully filtered multimodal data for cold-start SFT; second, a curriculum-based diagnostic refinement process optimizes the model to rectify errors in perception, logic, and knowledge.

3.1 Data Preparation: Filtering for Reasoning-Intensive VQA

Unlike general VQA tasks where answers are often direct visual recognitions (e.g., "What color is the pill?"), medical diagnostic reasoning requires multi-hop logic. To build a robust cold-start dataset, we utilize the PMC-VQA dataset Zhang et al. (2023) as our source but apply strict filtering criteria to select samples suitable for **Chain-of-Thought (CoT)** expansion: ❶ **Modality Necessity:** We filter out questions that can be answered by text blind (e.g., pure knowledge questions independent of the image) to ensure the visual modality is essential; ❷ **Logical Depth:** We select questions requiring at least 2 hops of reasoning (e.g., Symptom Identification $\rightarrow$ Disease Diagnosis). Simple classification samples are discarded; ❸ **Knowledge Density:** We prioritize questions linked to specific medical entities (diseases, anatomical structures) that can be verified against external databases.

From an initial pool of 70k samples, we curate a high-quality subset of **12k reasoning-intensive image-text pairs** for the subsequent trajectory construction.

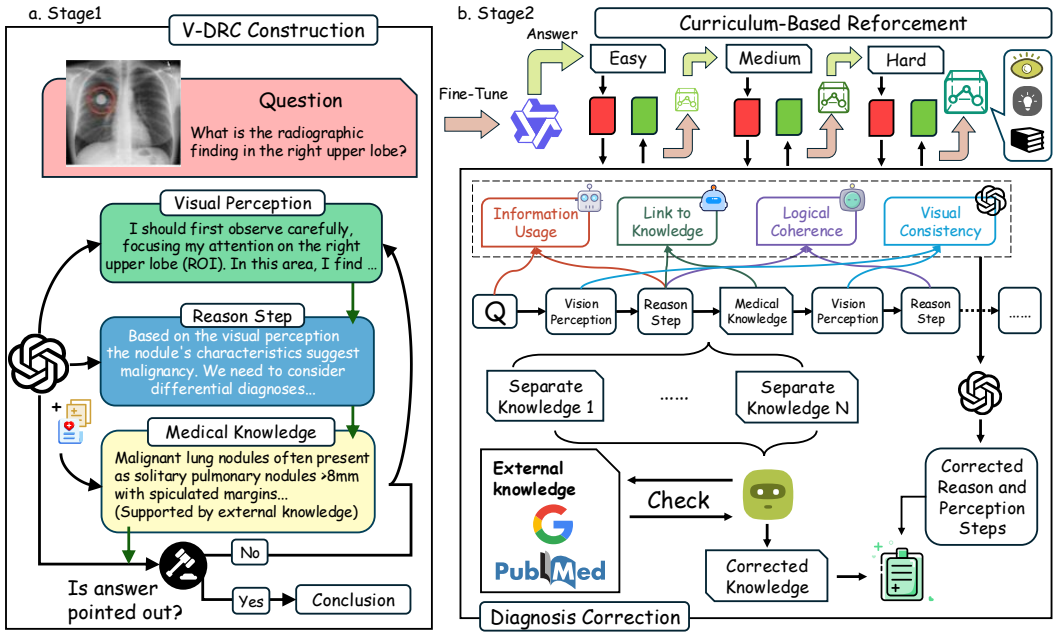


Figure 2: Schematic of the Diagno-SV architecture. (a) Stage 1: V-DRC Synthesis. A recurring loop that fuses visual feature extraction, logical deduction, and knowledge lookup to produce evidence-anchored diagnostic paths until a termination condition is satisfied. (b) Stage 2: Progressive Optimization.

3.2 Diagno-SV: The Tri-Level Decomposed Framework

To open the multimodal "black box," Diagno-SV compels the model to generate a structured output termed the **Visual-Grounded Decomposed Reasoning Chain (V-DRC)**. A complete V-DRC, denoted as $\mathcal{P}$, consists of a sequence of steps, where each step is a triplet of "visual perception" (v_i), "reasoning" (r_i), and "medical knowledge" (k_i):

$$\mathcal{P} = \{(v_1, r_1, k_1), (v_2, r_2, k_2), \ldots, (v_n, r_n, k_n)\} \tag{1}$$

where, $v_i = (\text{Focus}_i, \text{Desc}_i)$ represents the **Visual Perception**. Focus_i explicitly states the region of interest (RoI) the model is attending to (e.g., "upper right lung field"), and Desc_i provides a textual description of the visual findings in that region; r_i represents the **Reasoning Step**, acting as the bridge that synthesizes the visual finding v_i into a diagnostic hypothesis; k_i represents the **Medical Knowledge**, serving as the external factual basis to validate whether the visual finding v_i supports the hypothesis r_i (e.g., "Consolidation in X-ray is a primary indicator of Pneumonia").

3.3 Stage 1: Multi-Agent Collaborative Trajectory Construction

To transform the 12k filtered (Image, Question) pairs into full V-DRC trajectories ($\mathcal{P}^*$), we design a four-agent collaborative system. This system dynamically generates verified trajectories through a **Perception-Hypothesis-Verification** cycle.

Visual Perceptor Module ($\mathcal{A}_{visual}$). This sensory frontend determines the optimal visual attention at step i, given image I, question Q, and history H_{i-1}:

$$v'_i = (\text{Focus}'_i, \text{Desc}'_i) = \mathcal{A}_{\text{visual}}(I, Q, H_{i-1}) \tag{2}$$

This grounds the reasoning in pixel space, preventing hallucination of features absent from the image.

Hypothesis Generation Module ($\mathcal{A}_{reason}$). Taking v'_i as input, this module deduces the logical implication:

$$r'_i = \mathcal{A}_{\text{reason}}(Q, H_{i-1}, v'_i) \tag{3}$$

Two validators ensure quality: ❶ **Visual-Logic Consistency ($\mathcal{V}_{\text{align}}$)** checks that r'_i is strictly derived from v'_i, preventing hallucinated logic that ignores visual evidence; ❷ **Logical Coherence ($\mathcal{V}_{\text{logic}}$)** evaluates the flow from previous steps. A step is accepted only if both pass.

Knowledge Verification Module ($\mathcal{A}_{\text{knowledge}}$). This module queries external databases $\mathcal{S}$ to verify the medical validity of the link between the visual symptom v_i and the inferred condition r_i:

$$k_i = \mathcal{A}_{\text{knowledge}}(v_i, r_i, \mathcal{S}) \tag{4}$$

Conclusion Inference Module ($\mathcal{A}_{\text{conclude}}$). This module judges whether the accumulated evidence H_i suffices for a final diagnosis; the trajectory is saved only if the answer matches ground truth A^*. See Algorithm 1.

Algorithm 1 Stage 1: Multi-Agent V-DRC Trajectory Construction

Input: Image I, Question Q, Ground-truth A^*, Max iterations N_{max}
Output: A verified-correct Visual-Grounded DRC $\mathcal{P}^*$
$H \leftarrow \emptyset$
for $i = 1$ to N_{max} **do**
  repeat
    $v'_i \leftarrow \mathcal{A}_{\text{visual}}(I, Q, H)$; $r'_i \leftarrow \mathcal{A}_{\text{reason}}(Q, H, v'_i)$
  until $\mathcal{V}_{\text{align}}(v'_i, r'_i) \wedge \mathcal{V}_{\text{logic}}(r'_i, r_{i-1}) = \text{True}$
  $v_i \leftarrow v'_i$; $r_i \leftarrow r'_i$; $k_i \leftarrow \mathcal{A}_{\text{knowledge}}(v_i, r_i, \mathcal{S})$
  $H \leftarrow H \cup \{(v_i, r_i, k_i)\}$
  if $\mathcal{A}_{\text{conclude}}$.is_sufficient$(Q, H)$ **then**
    $C_{\text{final}} \leftarrow \mathcal{A}_{\text{conclude}}(Q, H)$
    if $C_{\text{final}} = A^*$ **then**
      return $\mathcal{P}^* \leftarrow H$
    else
      break {Incorrect diagnosis, discard}
    end if
  end if
end for
return FAILURE

3.4 Stage 2: Curriculum-Based Diagnostic Refinement with V-DPO

Although Stage 1 gives the model the structural capability to generate V-DRCs, it may still exhibit *retrospective forgery*—fabricating visual findings to support incorrect hypotheses. We therefore introduce a curriculum-based diagnostic refinement process that optimizes reliability by progressively correcting errors across visual, logical, and factual dimensions.

Curriculum Difficulty Partitioning. We partition the training set into Easy (D_{easy}), Medium (D_{medium}), and Hard (D_{hard}) levels based on the SFT model's consistency α over N rollouts:

$$\text{Difficulty}(I, Q) = \begin{cases} \text{Hard} & \text{if } \alpha < 0.2 \\ \text{Medium} & \text{if } 0.2 \leq \alpha \leq 0.9 \\ \text{Easy} & \text{if } \alpha > 0.9 \end{cases} \tag{5}$$

Starting from D_{easy}, the model must master basic visual recognition before tackling complex multi-hop diagnostic reasoning in D_{hard}.

Tri-Modal Diagnostic Correction and Preference Learning. For each difficulty level, we execute a "rollout-diagnose-correct-learn" closed loop.

❶ **Multimodal Rollout:** The current model (Iter(t)-DiagnoSV) generates K reasoning trajectories $\{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$ for a given image-question pair. Erroneous trajectories leading to incorrect diagnoses are flagged as $\mathcal{P}_{\text{wrong}}$.

❷ **The Tri-Modal Diagnostic Module ($\mathcal{M}_{\text{diag}}$):** This module decomposes error analysis into three strictly ordered phases: **Phase 1: Visual Perception Diagnosis ($\mathcal{M}_{\text{diag}}^{\text{visual}}$):** The most critical step for VLMs. A Teacher VLM verifies the Desc_i in the student's perception step v_i:

> *Diagnostic Logic:* Does the described feature (e.g., "pleural effusion") actually exist in image region Focus$_i$?

If the student "saw" something absent (visual hallucination), the Teacher overwrites v_i with a faithful description grounded in the image. **Phase 2: Reasoning Logic Diagnosis ($\mathcal{M}_{\text{diag}}^{\text{reason}}$):** Once v_i is corrected, we check whether r_i logically follows; derivations contradicting the corrected visual evidence are revised. **Phase 3: Knowledge Verification ($\mathcal{M}_{\text{diag}}^{\text{knowledge}}$):** We verify the medical facts k_i against external bases $\mathcal{S}$, correcting factual fabrications.

This pipeline transforms an error-ridden trajectory $\mathcal{P}_{\text{wrong}}$ (potentially containing visual lies) into a truthful, grounded trajectory $\mathcal{P}_{\text{corrected}}$:

$$\mathcal{P}_{\text{corrected}} = \mathcal{M}_{\text{diag}}^{\text{knowledge}}(\mathcal{M}_{\text{diag}}^{\text{reason}}(\mathcal{M}_{\text{diag}}^{\text{visual}}(\mathcal{P}_{\text{wrong}}, I))) \tag{6}$$

❸ **Visual-Aware Preference Learning:** We construct the preference dataset $\mathbb{D}_{\text{DPO}}$. Crucially, the rejection pair often represents a specific failure mode: **"Right answer for the wrong visual reason"** or **"Hallucinated features."**

$$\mathbb{D}_{\text{DPO}} = \{(\mathcal{P}_{\text{corrected}}, \mathcal{P}_{\text{wrong}})\} \tag{7}$$

By fine-tuning on these pairs, Diagno-SV learns to align its internal visual processing with its verbal outputs, effectively mitigating retrospective forgery.

The complete process is summarized in Algorithm 2.

Algorithm 2 Stage 2: Curriculum-Based Diagnostic Refinement for VLMs

Input: SFT-DiagnoSV, Multimodal Datasets $\{D_{\text{easy}}, D_{\text{medium}}, D_{\text{hard}}\}$
Output: Final DiagnoSV
Model $\leftarrow$ SFT-DiagnoSV
for $d \in \{D_{\text{easy}}, D_{\text{medium}}, D_{\text{hard}}\}$ **do**
  $\mathbb{D}_{\text{DPO}}^{d} \leftarrow \emptyset$
  for $(I, Q) \in D_d$ **do**
    $\{\mathcal{P}_j\}_{j=1}^{K} \leftarrow$ Rollout(Model, I, Q)
    for each incorrect $\mathcal{P}_j$ **do**
      $\mathcal{P}_{\text{chosen}} \leftarrow \mathcal{M}_{\text{diag}}(\mathcal{P}_j, I)$ {Diagnose vision, then logic, then knowledge}
      $\mathbb{D}_{\text{DPO}}^{d} \leftarrow \mathbb{D}_{\text{DPO}}^{d} \cup \{(\mathcal{P}_{\text{chosen}}, \mathcal{P}_j)\}$
    end for
  end for
  Model $\leftarrow$ FineTune_DPO(Model, $\mathbb{D}_{\text{DPO}}^{d}$)
end for
return Model

4 EXPERIMENTS

We design experiments with three objectives:

- **How does Diagno-SV compare in performance against state-of-the-art (SOTA) medical VLMs, including both open-source and proprietary models?**
- **What is the specific contribution of each architectural component—particularly the Visual Diagnosis module—to the overall model performance?**
- **To what extent does Diagno-SV quantitatively mitigate the "retrospective forgery" problem in medical visual question answering?**

4.1 EXPERIMENTAL SETUP

Baselines We compare against two groups of baselines. (1) **Proprietary Foundation Models**: GPT-4.1 OpenAI (2024), Claude Sonnet 4 Anthropic (2024), and Gemini 2.5 Flash Gemini Team, Google (2025), representing the current upper bound of multimodal reasoning. (2) **Open-Source**

Table 1: Performance comparison of Diagno-SV with SOTA Medical VLMs (Accuracy, %). Best open-source results are in **bold**. Proprietary models are listed for reference.

	Reasoning-Heavy	Clinical VQA Tasks				Comprehensive	
Model	**MMMU-Med**	**VQA-RAD**	**SLAKE**	**PathVQA**	**PMC-VQA**	**OmniMedVQA**	**Average**
Proprietary Models (Reference)							
GPT-4.1 OpenAI (2024)	75.2	65.0	72.2	55.5	55.2	75.5	66.4
Claude Sonnet 4 Anthropic (2024)	74.6	67.6	70.6	54.2	54.4	65.5	64.5
Gemini 2.5 Flash Gemini Team, Google (2025)	76.9	68.5	75.8	55.4	55.4	71.0	67.2
Open-Source Models							
LLaVA-Med-7B Li et al. (2023a)	28.7	50.2	50.0	37.0	26.6	39.6	38.7
HuatuoGPT-V-7B Chen et al. (2024a)	48.6	67.3	68.0	46.4	53.2	74.3	59.6
MedGemma-4B-IT Sellergren et al. (2025)	43.5	72.4	77.3	48.5	49.6	70.3	60.3
BioMediX2-8B Pieri et al. (2025)	39.8	52.5	55.9	35.8	42.7	64.7	48.6
InternVL2.5-8B Chen et al. (2024b)	53.5	59.4	69.0	42.1	51.3	81.3	59.4
InternVL3-8B Chen et al. (2024b)	59.2	65.4	72.8	48.6	53.8	79.1	63.2
Lingshu-7B Xu et al. (2025)	54.0	67.9	**83.1**	**61.9**	56.3	82.9	67.7
MedVLM-R1-2B Pan et al. (2025)	45.2	61.5	65.4	42.8	48.9	68.2	55.3
MIMO Chen et al. (2025)	51.8	64.2	70.5	47.3	52.1	73.8	59.9
Qwen3-VL-8B Qwen Team (2025)	55.4	66.8	71.2	49.5	54.0	76.5	62.2
Base: Qwen2.5-VL-7B Wang et al. (2024)	50.6	63.9	67.0	44.1	51.9	63.6	56.9
Diagno-SV (Ours)	**61.8**	**70.1**	78.5	59.8	**59.6**	**83.4**	**68.8**

Models: medical specialists including LLaVA-Med Li et al. (2023a), HuatuoGPT-V Chen et al. (2024a), MedGemma Sellergren et al. (2025), BioMediX2 Pieri et al. (2025), Lingshu-7B Xu et al. (2025), and MedVLM-R1 Pan et al. (2025), which applies reinforcement learning to elicit explicit diagnostic reasoning; general-domain VLMs including InternVL2.5/3 Chen et al. (2024b), MIMO Chen et al. (2025), and our backbone Qwen2.5-VL Wang et al. (2024); and the latest generation Qwen3-VL-8B Qwen Team (2025). This tests whether Diagno-SV can instill specialized, verifiable reasoning into general VLM architectures.

Benchmarks Evaluation spans six medical VQA benchmarks. For **Complex Reasoning** we use the medical subset of MMMU (MMMU-Med) Yue et al. (2024), requiring college-level subject knowledge and multi-step inference; for **Clinical Visual Understanding**, VQA-RAD Lau et al. (2018) and SLAKE Liu et al. (2021); for **Pathology**, PathVQA He et al. (2020). We additionally include PMC-VQA Zhang et al. (2023) (test split) for large-scale modality alignment and OmniMedVQA Hu et al. (2024) for comprehensive evaluation. Consistent with standard protocols, we report Accuracy (%) on the official test splits, or validation splits where test sets are withheld.

Dataset Construction Our training data derives from PMC-VQA, filtered for reasoning depth. For Stage 1 (Cold Start) we constructed **12k high-quality V-DRC trajectories** using our multi-agent system, with **GPT-4o** as the teacher agent generating accurate Visual Perception and Reasoning steps. For Stage 2 (Diagnostic Refinement) we generated **41k preference pairs** for DPO via the curriculum strategy, with GPT-4o also serving as the "Visual Auditor" to correct student hallucinations. External knowledge verification relied on PubMed and reliable medical knowledge graphs.

Model Training We use **Qwen-2.5-7B-VL-Instruct** as the base model. Stage 1 fine-tunes on the 12k V-DRC data for 3 epochs (learning rate $5e^{-6}$, batch size 128). Stage 2 performs DPO for 2 epochs (learning rate $1e^{-5}$, batch size 64, context length 4096 to accommodate high-resolution image tokens). All experiments run on 8 NVIDIA A100 (80GB) GPUs.

4.2 Main Results

Table 1 presents a comprehensive comparison of Diagno-SV against state-of-the-art (SOTA) medical Vision-Language Models (VLMs). Diagno-SV achieves **the highest average performance among all open-source models** (68.8%), establishing a new SOTA for medical VLMs. The key findings are summarized as follows:

❶ **Surpassing Open-Source SOTA:** Diagno-SV outperforms the strong general-purpose baseline, Qwen2.5-VL, by an average margin of **11.9%** (68.8% vs. 56.9%). It also surpasses the newer general-purpose model Qwen3-VL-8B (62.2%), demonstrating that specialized medical architecture yields greater benefits than simply scaling general VLMs.

Table 2: Performance evolution across stages of the diagnostic refinement process (Accuracy, %). Diagno-SV is 3-Iter DPO.

Benchmark	Original	SFT	DPO-1Iter	DPO-2Iter	DPO-3Iter
MMMU-Med	50.60	49.85	54.20	58.90	**61.81**
OmniMedVQA	63.60	65.10	72.40	79.50	**83.43**
VQA-RAD	63.90	64.50	66.80	68.90	**70.12**
PMC-VQA	51.90	53.20	55.40	57.80	**59.61**
SLAKE	67.00	68.50	72.10	75.60	**78.50**

Compared to the previous medical SOTA, Lingshu-7B, Diagno-SV achieves a higher average score (68.8% vs. 67.7%), confirming the effectiveness of our decomposed reasoning framework in enhancing overall generalization.

❷ **Outperforming Proprietary Models on Comprehensive Benchmarks:** Diagno-SV surpasses leading proprietary models in average accuracy, exceeding GPT-4.1 (66.4%) and Gemini 2.5 Flash (67.2%). Specifically, Diagno-SV achieves superior performance on complex visual-medical tasks, outperforming GPT-4.1 on OmniMedVQA (83.4% vs. 75.5%) and VQA-RAD (70.1% vs. 65.0%). This highlights Diagno-SV's strong capability in handling diverse, real-world medical visual queries with high precision.

❸ **Superior Reasoning Robustness:** On the MMMU-Med benchmark, which demands deep multimodal reasoning and complex clinical logic, Diagno-SV achieves 61.8%. This significantly outperforms other 7B medical VLMs, such as LLaVA-Med (28.7%) and MedVLM-R1-2B (45.2%). The substantial gap validates our **decomposed reasoning framework**, demonstrating its superior ability to handle intricate, multi-step medical queries compared to standard end-to-end training approaches.

4.3 Ablation Studies

To investigate the sources of performance gains in Diagno-SV, we conducted ablation experiments on the distinct roles of visual perception, logical reasoning, and medical knowledge.

4.3.1 Contribution Analysis of the Diagnostic Refinement Process

The essence of our method is the multi-stage, tri-modal diagnostic refinement process. To verify the necessity of each stage, we tracked performance from the base Qwen2.5-VL-7B to the final Diagno-SV (3-Iter DPO), as shown in Table 2. Notably, applying only SFT to enforce the V-DRC structure caused a slight dip on complex reasoning tasks like MMMU-Med, suggesting that forcing the model to output a "Visual Description" without feedback can distract it or produce ungrounded verbosity. Once diagnostic optimization with DPO was introduced, performance improved steadily, with the three iterations yielding continuous gains especially on **Reasoning-Heavy** (MMMU-Med) and **Visual-Heavy** (VQA-RAD) benchmarks. This confirms that our iterative optimization—penalizing visual hallucinations and logical fallacies—is the core driver of capability enhancement.

4.3.2 Diagnostic Analysis of Error Types

To understand *why* the model fails, we conducted a fine-grained error analysis. Unlike LLMs, VLMs face a "Perception-Reasoning Gap." We categorize errors into **Visual Perception Error** (hallucination/blindness), **Reasoning Error** (logic gap), and **Knowledge Error** (factual mistake). Table 3 reveals distinct patterns: ❶ **Visual Errors Dominate Easy Tasks:** For easy (recognition-based) problems, failures are predominantly visual (42.15% before Stage 2); if the model "misses" the lesion, no amount of reasoning can save it. ❷ **Logic Takeover in Hard Tasks:** As difficulty increases, logical errors surge (65.22% on Hard) and become the primary bottleneck, as complex cases require multi-hop inference where the visual finding is only the starting point. ❸ **Tri-Modal Correction Efficacy:** Diagno-SV reduces errors across all dimensions, most notably cutting the **Visual Error**

Table 3: Diagnostic analysis of error types by difficulty and refinement stage

		Easy	Medium	Hard
Visual Error	**Before Stage 2**	42.15%	28.40%	18.50%
	After Stage 2	15.30%	9.80%	7.20%
Reasoning Error	**Before Stage 2**	5.10%	38.20%	65.22%
	After Stage 2	2.20%	25.40%	51.10%
Knowledge Error	**Before Stage 2**	8.50%	20.10%	35.40%
	After Stage 2	4.10%	12.50%	22.80%

Table 4: Retrospective Forgery Rate (%) evolution. Forgery includes both Visual Hallucination and Knowledge Fabrication.

	SFT	**1-Iter DPO**	**2-Iter DPO**	**3-Iter DPO**
Forgery Rate	42.74%	37.98%	28.33%	**26.32%**

Rate from 42.15% to 15.30% on Easy tasks, proving that our "Visual Diagnosis" module teaches the model to "look closer" and describe images faithfully.

4.3.3 Verifying the Mitigation of Retrospective Forgery

A core contribution of Diagno-SV is mitigating **Retrospective Forgery**, which we redefine in VLMs as *fabricating visual findings or medical facts to justify a flawed hypothesis*. We used GPT-4o (with vision) to adjudicate whether the generated "Visual Description" (v_i) was grounded in the image and whether the "Medical Knowledge" (k_i) was factual. As shown in Table 4, the SFT model exhibits a 42.74% forgery rate, often "hallucinating nodules" to support a "tumor" diagnosis, which Diagno-SV reduces to 26.32%. This indicates the model has learned to be **visually honest**—admitting when a feature is absent rather than fabricating it to force a fit.

4.4 Limitations

Our cold-start trajectories and preference pairs are both built from PMC-VQA, so the reported gains are measured primarily on pathology and radiology style imagery, and generalization to other modalities remains untested. The diagnostic module also depends on a proprietary teacher model (GPT-4o) for visual auditing and knowledge verification, limiting reproducibility and making the refinement loop sensitive to teacher errors. Finally, although our refinement reduces the retrospective forgery rate from 42.74% to 26.32%, roughly a quarter of the remaining chains still contain ungrounded statements, which is not yet sufficient for unsupervised clinical deployment.

5 Conclusion

This work addresses the challenge of reasoning unreliability and hallucination in medical Vision-Language Models (VLMs). We identify and define a distinct failure mode termed "retrospective forgery," where models fabricate visual features to justify flawed hypotheses. To counter this, we propose **Diagno-SV**, a framework that transforms holistic "black box" optimization into a transparent process by decoupling reasoning into verifiable steps of **visual perception**, logical deduction, and medical knowledge. Experiments demonstrate that Diagno-SV achieves SOTA performance across diverse multimodal benchmarks while significantly reducing forgery rates. By aligning what the model "sees" with what it "thinks," Diagno-SV fosters more honest and trustworthy reasoning, a pivotal step toward the safe deployment of interpretable multimodal AI in clinical practice.

AI USE STATEMENT

Large Language Models (LLMs) were employed solely to improve the clarity, flow, and grammatical correctness of the text. The AI tools were not involved in the conceptualization, experimental design, data analysis, or scientific reasoning of this work. The authors have thoroughly inspected all AI-assisted edits and assume complete accountability for the final published content.

6 IMPACT STATEMENTS

This paper aims to advance machine learning research; we do not believe any societal consequence requires specific highlighting here.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. In *arXiv preprint arXiv:1606.06565*, 2016.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. *Anthropic Technical Report*, 2024.

Junying Chen, Chi Gui, Ruyi Gao, Zhongwei Wan, Hongyi Wang, Wenqi Zhang, Xiang Zhang, Shuang Zhou, Benyou Wang, Lijun Wu, et al. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4301–4321, 2024a.

Yanyuan Chen, Dexuan Xu, Yu Huang, Songkun Zhan, Hanpin Wang, Dongxue Chen, Xueping Wang, Meikang Qiu, and Hang Li. MIMO: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24732–24741, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024b.

Yunkai Dang, Wei Huang, Bin Li, et al. Explainable and interpretable multimodal large language models: A comprehensive survey. *arXiv preprint arXiv:2412.02104*, 2024.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *Nature*, 645:633–638, 2025.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, Pengtao Liu, and Pengtao Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

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

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023b.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023a.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-Flamingo: a multimodal medical few-shot learner. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367, 2023b.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

Sara Pieri, Sahal Shaji Mullappilly, et al. BiMediX2: A unified medical mixture-of-experts model for bilingual text and medical image understanding. *arXiv preprint arXiv:2504.02477*, 2025.

Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53841, 2023.

Ankit Rajaraman et al. The hidden risks of large reasoning models: A safety assessment of R1-based models. *arXiv preprint arXiv:2502.12659*, 2025.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Yiming Xu, Xiaoman Wang, Zheng Wang, et al. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. *arXiv preprint arXiv:2506.07044*, 2025.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024.