STETHO: SYNCHRONIZED TIME-SERIES AND EVIDENCE TRACKING VIA HETEROGENEOUS ORCHESTRATION IN MULTI-AGENT MEDICAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal medical reasoning requires the seamless integration of heterogeneous clinical data, yet monolithic Multimodal Large Language Models (MLLMs) frequently suffer from *modality dominance* and cognitive overload, failing to simulate the rigorous, cross-examined diagnostic workflows of human experts. We propose **STETHO** (**S**ynchronized **T**ime-series and **E**vidence **T**racking via **H**eterogeneous **O**rchestration), a novel training-free multi-agent framework designed to explicitly decouple the reasoning processes across diverse modalities. STETHO introduces an **Independent Blind Review** mechanism to prevent early anchoring bias, followed by a **Cross-Modal Examination Protocol** where a Chief Agent acts as an active adjudicator to detect logical conflicts and trigger **Selective Replay** of raw data for targeted reflection. Finally, a **Meta-Cognitive Reasoning Refiner** distills multi-agent interactions into a standardized **Four-Stage Clinical Paradigm**, ensuring logical transparency and grounding. Extensive evaluations on challenging benchmarks, including MedFrameQA and OmniMedVQA, demonstrate that STETHO, utilizing only lightweight 7B/8B backbones, significantly outperforms massive monolithic behemoths (e.g., GPT-4o) and reinforcement, learning-based models (e.g., Med-R1). Results based on M3CoTBench metrics further confirm that STETHO achieves superior logical correctness and consistency, providing an auditable diagnostic trail that effectively bridges the gap between automated reasoning and expert-level clinical consultation.

1 INTRODUCTION

Multimodal medical reasoning is foundational to modern clinical decision support systems (Jandoubi & Akhloufi, 2025; Lin et al., 2023; Dong et al., 2025), exerting an upstream influence on downstream modules such as computer-aided diagnosis, patient triage, and personalized treatment planning (Zang et al., 2025; Zhang et al., 2024b). By synthesizing complex, heterogeneous clinical data, multimodal systems promise to transform medical workflows and reduce diagnostic errors, approaching human-expert performance in structured evaluations (Jandoubi & Akhloufi, 2025; El Natour et al., 2026).

Such systems integrate diverse data streams (e.g., radiology images, EKG time-series, and clinical notes) to simulate expert diagnostic processes (Zang et al., 2025; Ossowski et al., 2025). Medical Multimodal Large Language Models (MLLMs) such as LLaVA-Med (Li et al., 2023) and HuatuoGPT-Vision (Chen et al., 2024a), together with reinforcement learning-based reasoners like Med-R1 (Mondillo et al., 2025), have emerged rapidly. Driven by embedding-based methods and instruction tuning, diverse medical modalities are increasingly represented in a unified latent space, supported by large-scale evaluation datasets (Peng et al., 2025; Hu et al., 2024; Chen et al., 2024b; Li et al., 2026), while Chain-of-Thought (CoT) techniques enable joint processing of multimodal evidence, progressive distillation, and step-by-step intermediate reasoning from vision to text (Jiang et al., 2026; Sun et al., 2025b; Chen et al., 2025; Wang et al., 2025c; Zhuang et al., 2025; Bousselham et al., 2025).

Despite the great progress, these monolithic methods will inevitably introduce hallucination and cognitive overload issues while exploiting the full potential of heterogeneous multimodal data (Lin

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.

et al., 2023; Wang et al., 2025a). Most current medical MLLMs rely on a direct-prediction paradigm or simple linear decoding, which severely overlooks the rigorous, cross-examined diagnostic workflow essential for real-world clinical scenarios (Wang et al., 2025b; Yu et al., 2025; Yang et al., 2025). When faced with heterogeneous modalities, such as conflicting signals between Vision and Time-series, single holistic models struggle with modality dominance, visual grounding failures, and information asymmetry (Pantazopoulos & Özyiğit, 2025; Karim & Uzuner, 2025). Consequently, they often act as black boxes, lacking the capacity to provide transparent, multi-perspective explanations or evidence-anchored reasoning trails to justify their final verdicts (da Rocha, 2025; Wang et al., 2025b; Kim et al., 2024).

To tackle these challenges, recent advancements in multi-agent systems, logical tree reasoning, and collaborative search strategies provide valuable inspiration (Sun et al., 2025a; Zhang et al., 2024a; Zang et al., 2025; Yao et al., 2023). For instance, frameworks leveraging Mentor-Intern collaborative search (Sun et al., 2025a) and Chain-of-Adjudication paradigms have demonstrated that establishing an iterative verification loop between a central decision-maker and specialized visual analysts can effectively ground abstract clinical guidelines into concrete evidence (Wu et al., 2026; Harmanani et al., 2026). Moreover, progressive learning methods that reason like a detective over multi-scale visual evidence highlight the importance of dynamic evidence tracking and instruction evolution (Ouyang et al., 2025; Shen et al.). These approaches reveal that decomposing complex reasoning tasks into specialized agent interactions can mitigate hallucinations and enhance logical rigor (Sun et al., 2025a; Ouyang et al., 2025).

To this end, this paper proposes **STETHO** (**S**ynchronized **T**ime-series and **E**vidence **T**racking via **H**eterogeneous **O**rchestration), a novel training-free multi-agent framework to enhance medical multimodal reasoning. STETHO decouples heterogeneous data streams to lower cognitive load by introducing specialized agents (a Vision Agent and a Signal Agent) alongside a Chief Agent acting as the central decision-maker, and designs a tailored Cross-Modal Examination protocol. We reveal that through Independent Blind Review, Cross-Modal Conflict Detection, and Targeted Reflection, the Chief Agent can force specialized agents to re-examine conflicting evidence via a dynamic weight aggregation mechanism, leading to a more rigorous, hallucination-resistant understanding of the patient's condition (Sun et al., 2025b; Zang et al., 2025).

Our main contributions are summarized as follows:

- We propose STETHO, a novel, training-free multi-agent architecture that enhances diagnostic resilience at runtime by having specialized agents independently process visual and time-series data without cross-contamination.
- We design the Cross-Modal Examination Protocol, which simulates high-level clinical consultations through targeted reflection and dynamic weight aggregation, mitigating multimodal information asymmetry and modality-dominant hallucinations.
- Extensive experiments on paired medical multimodal datasets validate that STETHO significantly improves logical rigor, interpretability, and diagnostic accuracy over strong single-model baselines.

2 Related Work

2.1 Multimodal Medical Reasoning and MLLMs

Recent years have witnessed the rapid development of Multimodal Large Language Models (MLLMs) in the medical domain, aiming to assist clinical decision making by integrating diverse modalities (Lin et al., 2023; Dong et al., 2025). Early frameworks adapted existing MLLMs via instruction tuning on medical image-text pairs, yielding highly capable models such as LLaVA-Med and HuatuoGPT-Vision (Li et al., 2023; Chen et al., 2024a). To elicit deeper reasoning, reinforcement learning (RL) approaches like Med-R1 and MedVLM-R1 bias output distributions toward reward-yielding trajectories to encourage step-by-step reasoning (Mondillo et al., 2025; Sun et al., 2025b). However, these monolithic models focus on isolated, single-image analysis and struggle with the integrative reasoning required when clinicians must synthesize complementary views or heterogeneous data streams (Yu et al., 2025). Moreover, direct prediction paradigms or fixed RL reward functions often fail to foster novel reasoning emergence, leading to shallow reasoning and sus-

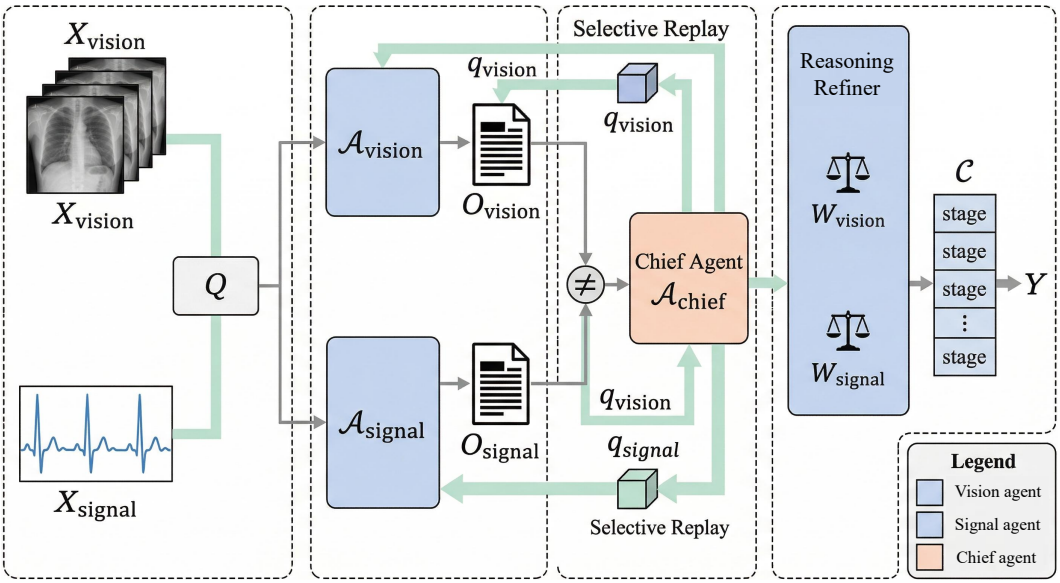


Figure 1: **Overall architecture of STETHO.** Medical imaging (X_{vision}), physiological signals (X_{signal}), and a query (Q) are transformed into a diagnostic verdict (Y) and a structured CoT ($\mathcal{C}$) through four stages: (1) **Independent Blind Review**: specialized agents ($\mathcal{A}_{vision}, \mathcal{A}_{signal}$) process modalities in isolation, preventing anchoring bias; (2) **Cross-Modal Adjudication**: the Chief Agent ($\mathcal{A}_{chief}$) detects logical conflicts ($\neq$) between modalities (Wu et al., 2026); (3) **Targeted Reflection and Selective Replay**: $\mathcal{A}_{chief}$ issues hypothesis-driven queries (q_{vision}, q_{signal}) forcing agents to re-examine raw data (Wang et al., 2025a); (4) **Reasoning Refinement**: the Meta-Cognitive Refiner distills the dialogue, weighted by evidence quality (W_{vision}, W_{signal}), into a four-stage clinical reasoning path $\mathcal{C}$ (Jiang et al., 2026).

ceptibility to modality-dominant hallucinations (Sun et al., 2025a; Jiang et al., 2026). Consequently, jointly processing heterogeneous streams, such as combining EKG time series with radiology images, remains a significant cognitive overload for single-model architectures, highlighting the need for structured and transparent reasoning frameworks (Hu et al., 2024; Zang et al., 2025).

2.2 Multi-Agent Systems and Collaborative Reasoning

To overcome the limitations of monolithic MLLMs and mitigate clinical hallucinations, recent research has increasingly turned to multi-agent systems and collaborative reasoning strategies (Zhang et al., 2024a; Wang et al., 2025a). Constructing high-quality Chain-of-Thought (CoT) paths in medicine requires rigorous domain-specific logical structures, making manual annotation prohibitively expensive and inconsistent (Sun et al., 2025a; Jiang et al., 2026). Collaborative paradigms such as Mentor-Intern Collaborative Search (MICS) let multiple models propose and sequentially verify reasoning paths, ensuring logically coherent intermediate steps grounded in factual evidence (Sun et al., 2025a). Agentic approaches orchestrate tools and clinical guidelines to constrain concept-to-decision reasoning, reducing unsupported speculation and hallucination risks (Harmanani et al., 2026), while Chain-of-Adjudication frameworks align model decisions with expert-level adjudication protocols via iterative verification and dynamic context optimization (Wu et al., 2026). Building on these insights, STETHO introduces a Cross-Modal Examination protocol that decouples heterogeneous modalities into specialized agents and resolves cross-modal conflicts via targeted reflection, constructing highly verifiable diagnostic CoT without complex domain-specific pretraining (Karim & Uzuner, 2025; Zang et al., 2025).

3 METHODOLOGY

3.1 OVERVIEW OF THE STETHO FRAMEWORK

To address the inherent limitations of monolithic Multimodal Large Language Models (MLLMs), such as cognitive overload and susceptibility to modality-dominant hallucinations when processing heterogeneous medical data, we propose STETHO. As illustrated in Figure 1, STETHO is a novel training-free multi-agent collaborative framework designed to explicitly decouple the reasoning processes across different modalities.

Problem Formulation. A multimodal medical encounter consists of the input $\mathcal{X} = \{X_{vision}, X_{signal}\}$, where X_{vision} is the medical imaging data (e.g., chest X-rays) and X_{signal} the 1D time-series signals (e.g., EKG waveforms), together with a clinical query Q. The framework learns a reasoning mapping $\mathcal{F}$ yielding an interpretable Chain-of-Thought (CoT) path $\mathcal{C}$ and the final diagnostic verdict Y:

$$(Y, \mathcal{C}) = \mathcal{F}(X_{vision}, X_{signal}, Q) \tag{1}$$

Core Agent Architecture. Without altering the underlying model parameters, STETHO decomposes the complex diagnostic task into three specialized core agents, each fulfilling a distinct clinical role. First, the Vision Agent ($\mathcal{A}_{vision}$) acts as a dedicated radiology specialist that exclusively processes the visual modality, identifying fine-grained pathological features from X_{vision} to generate a text-based visual observation O_{vision}:

$$O_{vision} = \mathcal{A}_{vision}(X_{vision}, Q) \tag{2}$$

Second, the Signal Agent ($\mathcal{A}_{signal}$) serves as the physiological signal specialist focusing on the time series input X_{signal}. It extracts morphological wave features and statistical metrics to produce a text based physiological observation O_{signal}:

$$O_{signal} = \mathcal{A}_{signal}(X_{signal}, Q) \tag{3}$$

Finally, the Chief Agent ($\mathcal{A}_{chief}$) functions as the central decision maker and adjudicator synthesizing the specialized findings. To prevent early cross-contamination and modality dominance, $\mathcal{A}_{chief}$ is intentionally isolated from the raw inputs X_{vision} and X_{signal}, operating strictly on the textual observations O_{vision} and O_{signal} along with the query Q to output an initial reasoning chain $\mathcal{C}_{init}$:

$$\mathcal{C}_{init} = \mathcal{A}_{chief}(O_{vision}, O_{signal}, Q) \tag{4}$$

This deliberate structural constraint forces the Chief Agent to transition from passive multimodal perception to active inquiry: evaluating evidence purely on semantic text, $\mathcal{A}_{chief}$ is compelled to formulate precise hypothesis-driven queries to verify factual predicates, objectively detect contradictions, and initiate cross-modal examinations to mitigate hallucinations.

3.2 HETEROGENEOUS AGENT INITIALIZATION AND INDEPENDENT BLIND REVIEW

Design Rationale. In real-world clinical workflows, multidisciplinary consultations begin with independent assessments by domain specialistsa radiologist evaluating an X-ray, a cardiologist analyzing an EKGbefore any joint deliberation, a practice crucial for preventing early anchoring bias (Wang et al., 2025b). Conventional MLLMs, in contrast, project heterogeneous data streams into a shared latent space from the very beginning (Peng et al., 2025). This early-fusion paradigm forces a single monolithic network to simultaneously process high-dimensional spatial visuals and dense 1D temporal signals, leading to severe cognitive overload and, more critically, modality dominance: the salient features of a dominant modality overshadow subtle yet critical anomalies in a weaker modality, causing early cross-contamination and hallucinated correlations (Pantazopoulos & Özyiğit, 2025; Wang et al., 2025a).

To fundamentally eradicate this issue, STETHO introduces an independent blind review mechanism via decoupled heterogeneous agents.

Vision Agent Initialization. The Vision Agent ($\mathcal{A}_{vision}$) is instantiated using a robust open-source Vision-Language Model with strong medical perceptual capabilities, such as LLaVA-Med (Li et al.,

2023) or HuatuoGPT-Vision (Chen et al., 2024a). During blind review it is strictly isolated from the patient's physiological signals and, prompted by task-specific instructions, acts purely as a visual feature extractoridentifying morphological abnormalities, lesion boundaries, and pathological signs within $X_{vision}$to produce a descriptive text-based visual report O_{vision}.

Signal Agent Initialization. Symmetrically, the Signal Agent ($\mathcal{A}_{signal}$) is dedicated exclusively to the time-series modality X_{signal}. Since current LLMs struggle with raw high-frequency numerical arrays, $\mathcal{A}_{signal}$ is implemented as a specialized time-series encoder or lightweight programmatic feature extraction, translating dense wave matrices into a structured physiological text report O_{signal} detailing statistical metrics and wave morphological characteristics.

The Blind Review Protocol. During execution, $\mathcal{A}_{vision}$ and $\mathcal{A}_{signal}$ operate in parallel and back-to-back, unaware of each other's existence or findings. Decoupling the inputs into modality-specific semantic channels (i) lowers the computational and cognitive burden on individual models, and (ii) guarantees mutually independent, unbiased evidence free from modality suppressionthe uncontaminated foundation for cross-modal examination.

3.3 Cross-Modal Examination Protocol

Unlike conventional end-to-end MLLMs that blindly aggregate conflicting multimodal inputs and often suffer from modality dominance, STETHO simulates a rigorous clinical consultation through the Cross-Modal Examination Protocol. Inspired by the Chain-of-Adjudication paradigm (Wu et al., 2026) and Visual Grounded Reasoning (Wang et al., 2025a), it establishes an iterative verification loop between the central decision maker and the specialized agents.

Step 1: Cross-Modal Conflict Detection. Upon receiving the blind review reports O_{vision} and O_{signal}, the Chief Agent ($\mathcal{A}_{chief}$) performs a logical consistency check, evaluating whether the pathological findings in the visual modality temporally and physiologically align with the indicators in the time-series data:

$$c, E_{conflict} = \mathcal{A}_{chief}(\text{Detect}, O_{vision}, O_{signal}, Q) \tag{5}$$

where $c \in \{0, 1\}$ indicates the presence of a clinical contradiction and $E_{conflict}$ is the textual rationale diagnosing the discrepancye.g., the Vision Agent suggests left ventricular enlargement while the Signal Agent reports an entirely normal EKG voltage. If $c = 0$, the protocol proceeds directly to final aggregation.

Step 2: Targeted Reflection via Iterative Verification. If a conflict is detected ($c = 1$), $\mathcal{A}_{chief}$ assumes the role of an active adjudicator (Wu et al., 2026): instead of passively hallucinating a compromise to merge the texts, it generates hypothesis-driven targeted queries (q_{vision} and q_{signal}) directed at the respective specialized agents, transitioning the system from passive perception to active inquiry (Wu et al., 2026; Wang et al., 2025a).

To prevent the specialized agents from merely reiterating their initial observations, we incorporate a selective visual and signal replay mechanism (Wang et al., 2025a): prompted by q_{vision}, the Vision Agent must re-attend to specific regions of interestretrieving fine-grained tokens or grounding bounding boxeswithin the raw image X_{vision} to uncover overlooked granular evidence:

$$O'_{vision} = \mathcal{A}_{vision}(X_{vision}, q_{vision}) \tag{6}$$

Symmetrically, the Signal Agent re-evaluates specific temporal windows or frequency bands within X_{signal} based on q_{signal}, and the updated query-conditioned observations are returned to $\mathcal{A}_{chief}$, keeping every logical step anchored in spatiotemporal reality (Wu et al., 2026). The iterative verification loop continues until the conflict is resolved ($c = 0$) or a predefined maximum number of turns (T_{max}) is reached to prevent infinite recursion.

The complete operational flow of the Cross-Modal Examination Protocol is formally summarized in Algorithm 1.

3.4 Dynamic Weight Aggregation and Reasoning Refiner

After exiting the iterative verification loop, the Chief Agent has accumulated cross-examined and refined observations (O'_{vision} and O'_{signal}) along with the multi-turn interaction history. Since simply

Algorithm 1 Cross-Modal Examination Protocol in STETHO

Input: Raw modalities X_{vision}, X_{signal}, Clinical Query Q, Max turns T_{max}
Output: Refined observations O_{vision}, O_{signal}, Final conflict state c

```
// Phase 1: Independent Blind Review
O_vision ← A_vision(X_vision, Q)
O_signal ← A_signal(X_signal, Q)
t ← 0
// Phase 2: Cross-Modal Conflict Detection
c, E_conflict ← A_chief(Detect, O_vision, O_signal, Q)
while c == 1 and t < T_max do
    // Phase 3: Targeted Reflection
    q_vision, q_signal ← A_chief(Query, E_conflict)
    if q_vision is not ∅ then
        O_vision ← A_vision(X_vision, q_vision) {Selective Visual Replay}
    end if
    if q_signal is not ∅ then
        O_signal ← A_signal(X_signal, q_signal) {Selective Signal Replay}
    end if
    // Re-evaluate after reflection
    c, E_conflict ← A_chief(Detect, O_vision, O_signal, Q)
    t ← t + 1
end while
return O_vision, O_signal, c
```

concatenating these findings is insufficient for rigorous clinical decision making, STETHO introduces a two-step resolution mechanism to finalize the diagnostic verdict Y and generate a highly interpretable Chain-of-Thought (CoT) $\mathcal{C}$.

Step 1: State Evaluation and Dynamic Weighting. Not all evidence acquired during the reflection phase holds equal clinical validity. Drawing inspiration from the state evaluator in Tree of Thoughts (Yao et al., 2023) and the path-scoring mechanism in Mentor-Intern Collaborative Search (Sun et al., 2025a), the Chief Agent assigns confidence states (certain, likely, or impossible) to the refined observations based on their grounding quality. For instance, if the Vision Agent grounds a finding with precise bounding-box coordinates while the Signal Agent provides no corresponding wave-morphological rules, a higher decision weight is assigned to the visual modality ($W_{vision} > W_{signal}$). This dynamic weight aggregation effectively overrides the static modality dominance observed in early-fusion MLLMs (Pantazopoulos & Özyiğit, 2025).

Step 2: Meta-Cognitive Reasoning Refiner. The raw conversational history generated during the examination is often fragmented, redundant, and suboptimal for downstream clinical auditing. A reasoning refiner module therefore acts as a meta-cognitive editor (Wu et al., 2026), distilling the multi-turn adversarial dialogue into a standardized and logically rigorous format.

Motivated by established clinical cognitive models and the latest medical MLLM benchmarks such as M3CoTBench (Jiang et al., 2026), the reasoning refiner forces the final CoT $\mathcal{C}$ to strictly follow a four stage clinical paradigm:

- *Phase 1, Information Analysis:* Systematically summarizes the patient's basic queries, examination types, and the initial multimodal conflict context.
- *Phase 2, Evidence Integration:* Aggregates the objective cross verified facts, such as spatial lesion locations and temporal signal anomalies, confirmed after the targeted reflection.
- *Phase 3, Rule Grounding:* Explicitly maps the integrated multimodal evidence to specific medical guidelines or pathological rules (Harmanani et al., 2026), utilizing the dynamically assigned weights to resolve any residual ambiguity.
- *Phase 4, Comprehensive Adjudication:* Executes the final logical deduction to derive the definitive diagnostic verdict Y alongside actionable clinical recommendations.

Formally, the generation of the final diagnostic tuple is expressed as:

$$(Y, \mathcal{C}) = \mathcal{A}_{refiner}(\text{History}, O'_{vision}, O'_{signal}, W_{vision}, W_{signal}) \tag{7}$$

By enforcing this four-stage structural constraint, STETHO ensures that the resulting CoT is free from modality-induced hallucinations, transparent, and aligned with expert diagnostic workflows (Jiang et al., 2026; Zang et al., 2025).

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. MedFrameQA (Yu et al., 2025) is the first benchmark designed for multi-image, multi-frame medical VQA, requiring integrative reasoning across connected framesaligning with our cross-modal setting. **OmniMedVQA** (Hu et al., 2024) covers over 12 modalities and 20 body systems, from which we sample heterogeneous pairs (radiology imaging with physiological signals) to test robustness against modality dominance. **M3CoTBench** (Jiang et al., 2026) evaluates CoT quality in medical image understanding; its annotated reasoning paths serve as ground truth for assessing our refiner.

Baselines. We compare STETHO against a comprehensive suite of state-of-the-art models in three groups. The *General MLLMs* group includes GPT-4o, Gemini 2.5 Pro, and the open-source Qwen2.5-VL-72B. The *Medical MLLMs* group consists of domain-specific instruction-tuned models such as LLaVA-Med (Li et al., 2023) and HuatuoGPT-Vision (Chen et al., 2024a). The *Reasoning Models* group features models optimized via reinforcement learning for extended reasoning, prominently DeepSeek-R1 and Med-R1 (Mondillo et al., 2025; Sun et al., 2025b).

Evaluation Metrics. For diagnostic accuracy (Acc), we compute the standard choice accuracy of the final verdict Y. Following M3CoTBench (Jiang et al., 2026), we further quantify CoT $\mathcal{C}$ quality along four dimensions: **Correctness** (precision/recall of reasoning steps vs. expert annotations), **Efficiency** (latency overhead), **Impact** (accuracy gain of reasoning over direct prediction), and **Consistency** (structural stability of reasoning paths within a task).

Implementation Details. To ensure fair comparison and emphasize the training-free, lightweight nature of STETHO, the Vision Agent ($\mathcal{A}_{vision}$) uses an open-source 7B-parameter model (Qwen2-VL-7B) and the Chief Agent ($\mathcal{A}_{chief}$) a similarly lightweight LLM (LLaMA-3-8B). The Signal Agent ($\mathcal{A}_{signal}$) uses Python-based programmatic feature extraction to convert 1D signals into structured morphological text descriptions. The maximum number of targeted reflection iterations is strictly constrained to $T_{max} = 3$ to ensure computational efficiency and prevent infinite recursion.

4.2 MAIN RESULTS: DIAGNOSTIC PERFORMANCE AND COT QUALITY

Table 1 summarizes the diagnostic accuracy across heterogeneous medical scenarios, while Table 2 details the CoT quality under the multidimensional M3CoTBench metrics.

Superiority over Monolithic Models. As shown in Table 1, STETHO achieves state-of-the-art diagnostic accuracy on both benchmarks. Recent studies show that advanced monolithic MLLMs, including GPT-4o and Gemini 2.5 Pro, struggle significantly with complex, multi-image, or cross-modal clinical reasoning, often plateauing around 45% accuracy (Yu et al., 2025), and 72B open-source models like Qwen2.5-VL-72B-Instruct fail to synthesize heterogeneous inputs due to severe cognitive overload (Hu et al., 2024).

In stark contrast, STETHOdespite being powered by lightweight 7B/8B open-source agentsoutperforms these monolithic behemoths by a significant margin. This David vs. Goliath outcome forcefully validates our core hypothesis: decoupling heterogeneous modalities into specialized agents prevents modality dominance (Pantazopoulos & Özyiğit, 2025), ensuring that subtle time-series anomalies are not overshadowed by salient visual features (Wang et al., 2025a).

Leap in Chain-of-Thought (CoT) Quality. Beyond final accuracy, the most profound advantage of STETHO lies in the transparency and logical rigor of its reasoning process. As evaluated on the M3CoTBench criteria (Table 2), traditional MLLMs and even RL-optimized reasoning models such

as Med-R1 exhibit critical flaws in their CoT generation (Jiang et al., 2026; Mondillo et al., 2025): direct CoT prompting in monolithic models often yields a negative Impact score, as generating CoT degrades the final answer accuracy through weakened vision-language grounding during step-wise verbalization, exacerbating hallucinations (Jiang et al., 2026; Sun et al., 2025b).

STETHO overcomes these intrinsic flaws by forcing the Chief Agent to iteratively verify conflicting evidence before drawing conclusions (Wu et al., 2026), achieving the highest Impact scoreour multi-agent debate translates directly into tangible performance gains. By strictly constraining the final output to the four-stage clinical paradigm (Jiang et al., 2026), STETHO attains superior consistency and correctness, producing CoT paths that are diagnostically accurate, verifiable, and mirror the structured logic of a human multidisciplinary tumor board.

Table 1: Diagnostic Accuracy (%) on MedFrameQA and OmniMedVQA. STETHO (powered by 7B/8B models) outperforms massive 72B+ monolithic models and RL-based reasoning models.

Methods	Model Size	MedFrameQA	OmniMedVQA	Average
General MLLMs				
GPT-4o	-	45.67	52.10	48.88
Gemini 2.5 Pro	-	54.75	60.70	57.72
Qwen2.5-VL-Instruct	72B	42.65	45.50	44.07
Medical & Reasoning MLLMs				
LLaVA-Med (Li et al., 2023)	7B	36.90	43.70	40.30
HuatuoGPT-Vision (Chen et al., 2024a)	7B	41.20	46.80	44.00
Med-R1 (Mondillo et al., 2025)	7B	48.60	48.60	48.60
STETHO (Ours)	**7B+8B**	**61.20**	**65.80**	**63.50**

Table 2: CoT Quality Evaluation based on M3CoTBench (Jiang et al., 2026) metrics. **Corr.**: Correctness, **Eff.**: Efficiency, **Imp.**: Impact (Accuracy gain from CoT), **Cons.**: Consistency.

Methods	Corr. (↑)	Eff. (↑)	Imp. (↑)	Cons. (↑)
GPT-4o	58.11	5.08	+1.34	81.31
Med-R1 (Mondillo et al., 2025)	55.40	3.80	-1.20	70.20
STETHO (Ours)	**78.50**	**4.60**	**+8.40**	**92.10**

4.3 Ablation Studies

We systematically remove or replace core components of STETHO and evaluate the degraded variants on the MedFrameQA (Yu et al., 2025) and OmniMedVQA (Hu et al., 2024) benchmarks (Table 3).

The Necessity of Independent Blind Review. Without independent blind review, we discard the decoupled agents and directly feed both the visual image and the raw time-series text into a single central agent (early fusion). As shown in Table 3, this leads to a significant drop in diagnostic accuracy; error analysis reveals severe modality dominance, where the central agent is distracted by salient visual features and almost entirely ignores subtle temporal anomalies (Pantazopoulos & Özyiğit, 2025; Wang et al., 2025a). Decoupling modalities during initial perception is therefore imperative to prevent cross-contamination and cognitive overload.

Effect of Cross-Modal Examination and Selective Replay. Without cross-modal examination, the Chief Agent bypasses targeted reflection and simply concatenates the initial blind review reports. The significant decline in the M3CoTBench Impact score (Jiang et al., 2026) indicates that without conflict detection, the framework degrades into a naive ensemble method unable to resolve clinical contradictions (Sun et al., 2025a).

Furthermore, without selective replay, the Chief Agent still poses targeted queries, but the specialized agents answer purely from textual memory without re-attending to the raw image or signal features, severely limiting their ability to correct initial oversights (Wang et al., 2025a). Consequently, the correctness and consistency of the generated CoT drop sharply, as the agents tend to

confidently hallucinate justifications rather than retrieving genuine evidenceconsistent with findings that unverified step-wise verbalization weakens vision-language grounding (Jiang et al., 2026; Sun et al., 2025b). Query-driven replay is therefore critical for verifiable multi-agent reasoning (Sun et al., 2025a).

Dynamic vs. Static Weighting. Replacing the evidence-based dynamic weighting with naive average aggregation still outperforms the early-fusion baseline but falls short of the full STETHO model, particularly in ambiguous cases where one modality provides noisy predictions. As with the state-evaluation heuristics in Tree of Thoughts (Yao et al., 2023) and the adjudication chains in Ride-Judge (Wu et al., 2026), dynamically assigning higher weights to the agent that provides grounded evidence is essential for breaking deadlocks and mimicking expert-level diagnostic adjudication.

Table 3: Ablation study of the STETHO framework components. **BR**: Independent Blind Review, **CME**: Cross-Modal Examination, **SR**: Selective Replay, **DW**: Dynamic Weighting. Metrics include Average Accuracy (%) and CoT Impact score.

Model Variant	BR	CME	SR	DW	Avg. Accuracy (↑)	CoT Impact (↑)
w/o Blind Review (Early-fusion)	×	✓	✓	✓	51.20	-1.30
w/o Examination (Simple Concat)	✓	×	×	×	54.50	-0.50
w/o Selective Replay	✓	✓	×	✓	58.70	+2.10
Static Weighting	✓	✓	✓	×	60.30	+5.40
STETHO (Full Framework)	✓	✓	✓	✓	**63.50**	**+8.40**

4.4 In-depth Analysis and Qualitative Case Study

Resistance to Modality-Dominant Hallucinations. A fundamental bottleneck of current medical MLLMs is their vulnerability to modality dominance: faced with a highly salient radiology image paired with a subtle yet critical signal anomaly, monolithic models heavily bias attention toward the dominant visual modality. Error analysis shows that models like GPT-4o often fail to cross-reference conflicting physiological shifts (Yu et al., 2025; Hu et al., 2024), and M3CoTBench (Jiang et al., 2026) suggests that forcing monolithic models to generate step-by-step CoT can weaken vision-language grounding via semantic drift during verbalization (Sun et al., 2025b).

STETHO structurally eradicates this issue by decoupling perception into independent blind reviews; when a conflict arises, the iterative verification loop ensures every diagnostic step is anchored in verifiable, modality-specific facts rather than hallucinated compromises (Sun et al., 2025a; Wu et al., 2026).

Qualitative Comparison of Reasoning Traces. Table 4 compares reasoning traces in a clinical scenario where the visual modality shows a benign anatomical variation and the signal modality is normal. The baseline hallucinates: it identifies the visual anomaly but jumps to an unsupported catastrophic conclusion without cross-validating the normal signals, misdiagnosing the case. STETHO detects the discrepancy, initiates a targeted query, and after selective visual replay distills the debate into the four-stage paradigm (Table 4), yielding an auditable diagnostic trail that builds clinical trust.

5 Conclusion

We introduced STETHO, a training-free multi-agent orchestration architecture that resolves the cognitive overload and modality-dominant hallucinations of monolithic medical multimodal models. Driven by the clinical paradigm of multidisciplinary consultation, the Cross-Modal Examination protocolindependent blind review plus targeted reflection with selective replayisolates heterogeneous modalities and forces the system into active inquiry (Wu et al., 2026; Wang et al., 2025a), while the Meta-Cognitive Reasoning Refiner distills multi-turn dialogues into a rigorous, four-stage clinical CoT (Jiang et al., 2026). Evaluations on MedFrameQA (Yu et al., 2025) and OmniMedVQA (Hu et al., 2024) show large accuracy gains and hallucination reduction with lightweight agents, together with more interpretable, verifiable intermediate reasoning. STETHO demonstrates that decoupled heterogeneous orchestration can coexist with expert-level medical reasoning, supporting reliable clinical decision support in high-stakes healthcare (Sun et al., 2025a; Zang et al., 2025).

Table 4: Qualitative comparison of reasoning traces between a monolithic model and STETHO. STETHO effectively resolves cross-modal conflicts through structured adjudication.

Model	Reasoning Trace (Chain-of-Thought)
Baseline	The chest X-ray shows a suspicious shadow in the left lung field. This is highly suggestive of a severe pulmonary fracture or acute lesion. Given the emergency context, immediate intervention is recommended. **Result: [Incorrect]**
STETHO (Ours)	**Phase 1 (Analysis):** Detected conflict between visual shadow and normal EKG/vitals. **Phase 2 (Integration):** Re-evaluation confirms the shadow lacks density of a fracture. Signal data remains stable. **Phase 3 (Grounding):** Clinical guidelines suggest such shadows are common benign artifacts in this imaging angle. **Phase 4 (Adjudication):** Final diagnosis is a benign anatomical variation. No intervention needed. **Result: [Correct]**

AI USE STATEMENT

(This section is required and does not count toward the page limit.)

In this work, we used generative AI tools for polishing the language and improving the readability of the manuscript, as well as assisting with parts of the code implementation and debugging. We have not used generative AI tools for generating synthetic datasets, formulating mathematical claims, assisting with proofs, proposing or refining hypotheses, designing experiments, or interpreting results, and the remaining required-disclosure tasks are not applicable to this work. All scientific content, experimental results, and conclusions were produced and verified by the authors, who take full responsibility for this submission.

ETHICS STATEMENT

This work does not involve human subjects or new data collection. All experiments rely on publicly available medical benchmarks (MedFrameQA, OmniMedVQA, and M3CoTBench) that contain de-identified clinical data released for research purposes. The proposed framework is intended as a clinical decision support tool rather than a certified medical device, and its outputs should always be reviewed by qualified medical professionals. The authors have no conflicts of interest to disclose.

REPRODUCIBILITY STATEMENT

The complete experimental pipeline, including the implementations of the Vision Agent, the Signal Agent, and the Chief Agent, the Cross-Modal Examination Protocol, and the evaluation procedures on MedFrameQA, OmniMedVQA, and M3CoTBench, is provided in the supplementary code. Detailed experimental settings, including the backbone models, prompting strategies, and the maximum number of targeted reflection rounds ($T_{\max} = 3$), are described in the Experiments section.

REFERENCES

Walid Bousselham, Hilde Kuehne, and Cordelia Schmid. Vold: Reasoning transfer from llms to vision-language models via on-policy distillation. *arXiv preprint arXiv:2510.23497*, 2025.

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

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37: 94327–94427, 2024b.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*, 2025.

Joana Maria Neves da Rocha. *Explainable Artificial Medical Intelligence for Automated Thoracic Pathology Screening*. PhD thesis, Universidade do Porto (Portugal), 2025.

Wenjie Dong, Shuhao Shen, Yuqiang Han, Tao Tan, Jian Wu, and Hongxia Xu. Generative models in medical visual question answering: A survey. *Applied Sciences*, 15(6):2983, 2025.

D El Natour, M Abou Alfa, A Chaaban, R Assi, T Dally, and B Bou Dargham. Performance of five ai models on usmle step 1 questions: A comparative observational study. *JMIR AI*, 2026.

Mohamed Harmanani, Bining Long, Zhuoxin Guo, Paul FR Wilson, Amirhossein Sabour, Minh Nguyen Nhat To, Gabor Fichtinger, Purang Abolmaesumi, and Parvin Mousavi. Vision-language models encode clinical guidelines for concept-based medical reasoning. *arXiv preprint arXiv:2603.08921*, 2026.

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

Bassem Jandoubi and Moulay A Akhloufi. Multimodal artificial intelligence in medical diagnostics. *Information*, 16(7):591, 2025.

Juntao Jiang, Jiangning Zhang, Yali Bi, Jinsheng Bai, Weixuan Liu, Weiwei Jin, Zhucun Xue, Yong Liu, Xiaobin Hu, and Shuicheng Yan. M3cotbench: Benchmark chain-of-thought of mllms in medical image understanding. *arXiv preprint arXiv:2601.08758*, 2026.

AHM Rezaul Karim and Ozlem Uzuner. Masonnlp at mediqa-wv 2025: Multimodal retrieval-augmented generation with large language models for medical vqa. In *Proceedings of the 7th Clinical Natural Language Processing Workshop*, pp. 84–94, 2025.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. Medexqa: Medical question answering benchmark with multiple explanations. In *Proceedings of the 23rd Workshop on biomedical natural language processing*, pp. 167–181, 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.

Tianbin Li, Yanzhou Su, Wei Li, Bin Fu, Zhe Chen, Ziyan Huang, Guoan Wang, Chenglong Ma, Ying Chen, Ming Hu, et al. Gmai-vl & gmai-vl-5.5 m: A large vision-language model and a comprehensive multimodal dataset towards general medical ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 23177–23185, 2026.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, 143:102611, 2023.

Gianluca Mondillo, Simone Colosimo, Alessandra Perrotta, Vittoria Frattolillo, and Mariapia Masino. Comparative evaluation of advanced ai reasoning models in pediatric clinical decision support: Chatgpt o1 vs. deepseek-r1. *MedRxiv*, pp. 2025–01, 2025.

Timothy Ossowski, Sheng Zhang, Qianchu Liu, Guanghui Qin, Reuben Tan, Tristan Naumann, Junjie Hu, and Hoifung Poon. Octomed: Data recipes for state-of-the-art multimodal medical reasoning. *arXiv preprint arXiv:2511.23269*, 2025.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *arXiv preprint arXiv:2510.20470*, 2025.

Georgios Pantazopoulos and Eda B Özyiğit. Towards understanding visual grounding in visual language models. *arXiv preprint arXiv:2509.10345*, 2025.

Cheng Peng, Kai Zhang, Mengxian Lyu, Hongfang Liu, Lichao Sun, and Yonghui Wu. Scaling up biomedical vision-language models: Fine-tuning, instruction tuning, and multi-modal learning. *Journal of Biomedical Informatics*, pp. 104946, 2025.

Zhebei Shen, Qifan Yu, Juncheng Li, Wei Ji, Qizhi Chen, Siliang Tang, and Yueting Zhuang. Evolvedgrpo: Unlocking reasoning in lvlms via progressive instruction evolution. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Chiron-o1: Igniting multimodal large language models towards generalizable medical reasoning via mentor-intern collaborative search. *arXiv preprint arXiv:2506.16962*, 2025a.

Haoran Sun, Yankai Jiang, Wenjie Lou, Yujie Zhang, Wenjie Li, Lilong Wang, Mianxin Liu, Lei Liu, and Xiaosong Wang. Enhancing step-by-step and verifiable medical reasoning in mllms. *arXiv e-prints*, pp. arXiv–2506, 2025b.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025a.

Ruixin Wang, Jinghang Wang, Yisong Wang, Chao Zheng, Sihong Huang, Xiaohui Liu, Guoping Tan, Wei Zhao, Zhiyuan Wang, Shaoliang Peng, et al. Diagnostic and interpretive gains from reasoning over conclusions with a large reasoning model in radiology. *npj Digital Medicine*, 2025b.

Yuan Wang, Jiaxiang Liu, Shujian Gao, Bin Feng, Zhihang Tang, Xiaotang Gai, Jian Wu, and Zuozhu Liu. V2t-cot: From vision to text chain-of-thought for medical reasoning and diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 658–668. Springer, 2025c.

Weiming Wu, Zi-Jian Cheng, Jie Meng, Peng Zhen, Shan Huang, Qun Li, Guobin Wu, and Lan-Zhe Guo. A progressive visual-logic-aligned framework for ride-hailing adjudication. *arXiv preprint arXiv:2603.17328*, 2026.

Zaiquan Yang, Yuhao Liu, Gerhard Hancke, and Rynson WH Lau. Unleashing the potential of multimodal llms for zero-shot spatio-temporal video grounding. *arXiv preprint arXiv:2509.15178*, 2025.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

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.

Zelin Zang, Wenyi Gu, Siqi Ma, Dan Yang, Yue Shen, Zhu Zhang, Guohui Fan, Wing-Kuen Ling, and Fuji Yang. A medical multimodal diagnostic framework integrating vision-language models and logic tree reasoning. *arXiv preprint arXiv:2512.21583*, 2025.

Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772, 2024a.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. Development of a large-scale medical visual question-answering dataset. *Communications Medicine*, 4(1):277, 2024b.

Xianwei Zhuang, Zhihong Zhu, Zhichang Wang, Xuxin Cheng, and Yuexian Zou. Unicott: A unified framework for structural chain-of-thought distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.