Dialectic-Med: Mitigating Diagnostic Hallucinations via Counterfactual Adversarial Multi-Agent Debate

Zhixiang Lu
Xi'an Jiaotong-Liverpool University
Zhixiang.Lu22@student.xjtlu.edu.cn

Jionglong Su
Xi'an Jiaotong-Liverpool University
Jionglong.Su@xjtlu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) in healthcare suffer from severe confirmation bias, often hallucinating visual details to support initial, potentially erroneous diagnostic hypotheses. Existing Chain-of-Thought (CoT) approaches lack intrinsic correction mechanisms, rendering them vulnerable to error propagation. To bridge this gap, we propose Dialectic-Med, a multi-agent framework that enforces diagnostic rigor through adversarial dialectics. Unlike static consensus models, Dialectic-Med orchestrates a dynamic interplay between three role-specialized agents: a proponent that formulates diagnostic hypotheses; an opponent equipped with a novel visual falsification module that actively retrieves contradictory visual evidence to challenge the Proponent; and a mediator that resolves conflicts via a weighted consensus graph. By explicitly modeling the cognitive process of falsification, our framework guarantees that diagnostic reasoning is tightly grounded in verified visual regions. Empirical evaluations on MIMIC-CXR-VQA, VQA-RAD, and PathVQA demonstrate that Dialectic-Med not only achieves state-of-the-art performance but also fundamentally enhances the trustworthiness of the reasoning process. Beyond accuracy, our approach significantly enhances explanation faithfulness and decisively mitigates hallucinations, establishing a new standard over single-agent baselines.

1 Introduction

Multimodal large language models (MLLMs) are rapidly being integrated into high-stakes domains such as healthcare, demonstrating significant potential in tasks ranging from radiological report generation to medical visual question answering (Nam et al., 2025; Zhu et al., 2025). By reasoning over both visual (e.g., X-rays, CT scans) and textual data, these models promise to alleviate clinician burdens and enhance diagnostic accessibility (Bazi et al., 2023).

However, a critical failure mode hampers their clinical adoption: diagnostic hallucination. MLLMs exhibit a strong cognitive tendency towards confirmation bias (Nickerson, 1998), often generating fluent, plausible-sounding, yet factually incorrect diagnostic statements that are ungrounded in visual evidence (Kim et al., 2025). A model may latch onto a preliminary textual hypothesis and subsequently "hallucinate" visual features to support this erroneous conclusion. This phenomenon leads to a cascade of propagated errors, posing severe risks to patient safety.

Current reasoning enhancement techniques, most notably Chain-of-Thought (CoT) prompting (Wei et al., 2022; Miao et al., 2024), attempt to improve interpretability by generating step-by-step diagnostic pathways. While valuable, we argue that these approaches fundamentally suffer from a "Verificationist Trap". Their linear, forward-reasoning nature lacks an intrinsic mechanism for self-correction; they tend to seek evidence that verifies the current step rather than challenging it. Once an error is introduced early in the chain, the model often engages in post-hoc rationalization, cementing the error rather than rectifying it.

To bridge this gap, we propose **Dialectic-Med**, a novel multi-agent framework that enforces diagnostic rigor through **Counterfactual Adversarial Dialectics**. Drawing inspiration from Karl Popper's philosophy of scientific falsification (Popper, 2005), we posit that a robust diagnosis is established not merely by finding supporting evidence, but by surviving rigorous attempts at refutation. Unlike static consensus models, Dialectic-Med orchestrates a dynamic interplay between three role-specialized agents:

- **The Proponent**, acting as the initial diagnostician, analyzes the medical image to formulate a diagnostic hypothesis and a corresponding rationale, akin to a standard MLLM.

Findings of the Association for Computational Linguistics: ACL 2026, pages 36877–36894

- **The Opponent**, acting as a scientific skeptic equipped with a novel Visual Falsification Module (VFM). Instead of simply arguing semantically, the Opponent actively seeks to falsify the Proponent's hypothesis by retrieving contradictory visual evidence (e.g., "If this were pneumonia, there should be opacity here, but the costophrenic angle is sharp").

- **The Mediator**, acting as an impartial adjudicator. It analyzes the dialectical conflict and resolves it through a weighted consensus graph, ensuring the final diagnosis is grounded in verified visual regions.

By explicitly modeling the cognitive process of falsification, Dialectic-Med compels the system to break the cycle of confirmation bias. Our framework forces the model to ground its reasoning in verified visual regions that have survived adversarial scrutiny. Experiments on three challenging medical VQA benchmarks (MIMIC-CXR-VQA (Johnson et al., 2019; Bae et al., 2023), VQA-RAD (Lau et al., 2018), and PathVQA (He et al., 2020)) demonstrate that Dialectic-Med establishes a new state-of-the-art. Notably, it improves explanation faithfulness by 12.5% and significantly mitigates diagnostic hallucinations compared to single-agent baselines. Our contributions are threefold: (i) We introduce Dialectic-Med, the first medical multi-agent framework to operationalize popperian falsification and adversarial dialectics as an intrinsic error-correction mechanism. (ii) We propose a novel visual falsification module that enables agents to actively seek and retrieve visual evidence that contradicts a specific diagnostic hypothesis. (iii) We demonstrate significant improvements in diagnostic accuracy and explanation faithfulness on three public benchmarks, highlighting the efficacy of adversarial debate for robust medical AI.

2 Related Work

2.1 Hallucination in Medical MLLMs

The integration of MLLMs into healthcare has been met with both enthusiasm and caution (Moor et al., 2023). A primary safety bottleneck is the prevalence of hallucinations: generated content that is plausible but factually incorrect or unfaithful to the input data (Ji et al., 2023; Liu et al., 2024a). In the medical domain, such hallucinations manifest as fabricated clinical findings or misinterpretations of radiological images, posing severe risks to patient safety (Ding et al., 2023). Recent benchmarks have characterized this issue; for instance, Pandit et al. (2025) introduced *Med-HallMark* to specifically evaluate hallucinations in medical multimodal contexts. These studies highlight that hallucinations often stem from *confirmation bias*, where models over-rely on parametric priors rather than grounding assertions in visual evidence (Tang et al., 2026). Our work addresses this by introducing an explicit falsification mechanism to verify visual claims.

2.2 Reasoning in Large Language Models

To enhance reliability, structured reasoning techniques like CoT prompting (Wei et al., 2022) have become standard, encouraging models to generate intermediate reasoning steps. In medicine, CoT has been adapted to mimic diagnostic workflows (Liévin et al., 2024). Variants such as *Med-PaLM* (Singhal et al., 2023) and self-consistency prompting (Wang et al., 2023) aim to refine reasoning paths. However, these methods are predominantly linear and lack intrinsic self-correction, an error introduced early in the chain often propagates unchecked. Our dialectical framework overcomes this vulnerability by introducing adversarial checks at each reasoning step, transforming linear reasoning into a dynamic verification loop.

2.3 Multi-Agent Systems

Multi-agent systems leverage debate and deliberation to solve complex problems, often outperforming single-agent systems in reasoning and robustness (Liang et al., 2024; Du et al., 2024). Frameworks like *CAMEL* (Li et al., 2023b) demonstrate how role-playing agents can collaborate to solve tasks. In the medical domain, recent works have explored multi-agent collaboration for diagnosis (Tang et al., 2024). However, most existing approaches rely on static consensus or text-only debate. Dialectic-Med advances this paradigm by structuring the interaction as a formal popperian dialectic process with specialized roles (Proponent, Opponent, Mediator), creating a rigorous, truth-seeking dynamic driven by visual evidence rather than simple semantic consensus.

2.4 Counterfactual Reasoning and Grounding

A cornerstone of our framework is the Opponent's ability to perform visual falsification, rooted in counterfactual reasoning (Niu et al., 2021). This involves evaluating "what if" scenarios (e.g., "If this were pneumonia, opacity should be present")

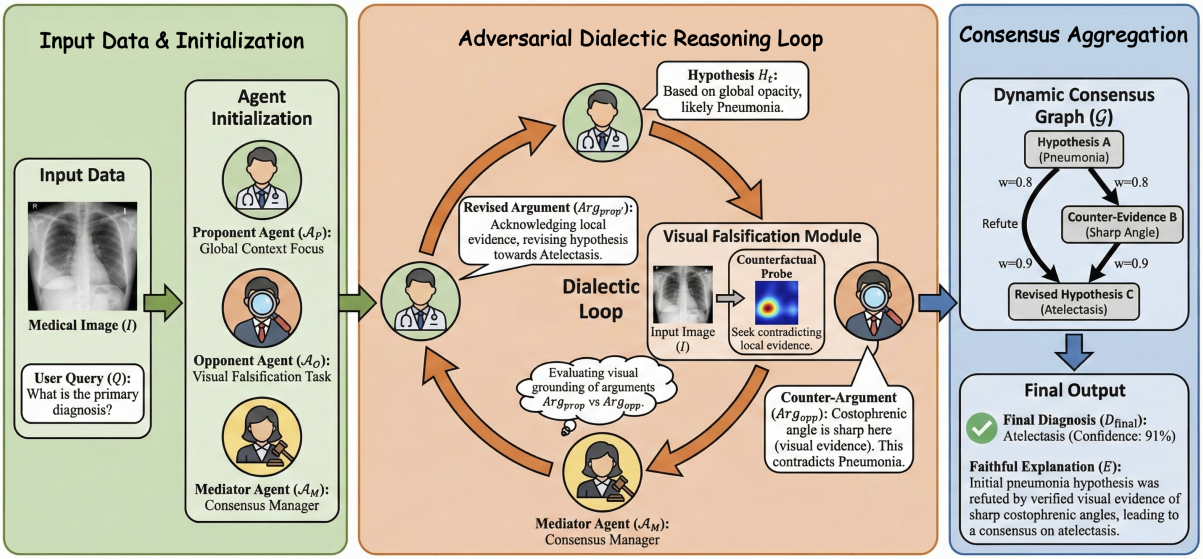

Figure 1: The overall architecture of Dialectic-Med. The framework orchestrates a structured adversarial debate among three role-specialized agents: Proponent, Opponent, and Mediator. (1) **Hypothesis Generation:** The Proponent formulates an initial diagnosis based on global context. (2) **Visual Falsification:** The Opponent utilizes a *Visual Falsification Module (VFM)* to actively retrieve contradictory visual evidence (counterfactuals) to challenge the hypothesis. (3) **Consensus Aggregation:** The Mediator adjudicates the conflict and updates a *Dynamic Consensus Graph*, filtering out hallucinations to derive a robust, visually-grounded final diagnosis.

(Boecking et al., 2022). This process requires deep visual grounding: localizing image regions that correspond to text (Lu et al., 2026b). Unlike models that use implicit attention, our VFM makes grounding explicit and adversarial, actively searching for regions that contradict the hypothesis to ensure faithful diagnostics.

3 Methodology

3.1 Problem Formulation

Given a medical image I and a diagnostic query Q, our objective is to derive a final diagnosis D^* and a faithful explanation E. We model the reasoning process as the construction of a **Dynamic Consensus Graph** $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$, where nodes $\mathcal{V}_t$ represent diagnostic hypotheses or visual evidence, and edges $\mathcal{E}_t$ encode logical relations (support/refute) with confidence weights. The final output is derived by traversing the converged graph $\mathcal{G}_{final}$.

3.2 Visual Falsification Module

The core innovation of Dialectic-Med is the Opponent agent's ability to ground its counter-arguments. Unlike standard agents that refute based on textual priors, the Opponent utilizes a **Visual Falsification Module (VFM)** to actively localize evidence that *contradicts* the current hypothesis.

Counterfactual Probe Generation. Given a hypothesis H_t (e.g., "Pneumonia"), the Opponent first generates a textual counterfactual probe Q_{cf}. This is a directive query targeting specific visual features that would falsify H_t:

$$Q_{cf} = \text{GenProbe}(H_t, \mathcal{K}_{\text{med}}) \tag{1}$$

where $\mathcal{K}_{\text{med}}$ denotes domain knowledge. For instance, if H_t implies opacity, Q_{cf} might target "sharp costophrenic angles" or "clear lung fields".

Falsification Attention. The VFM leverages a medical vision-language encoder (PubMedCLIP) to ground Q_{cf} in image I. Let $V = \{v_1, ..., v_N\}$ be the patch embeddings of I, and $q = \mathcal{E}_{txt}(Q_{cf})$ be the probe embedding. We compute the falsification attention map $M_{cf} = \{\alpha_1, ..., \alpha_N\}$ via a scaled cosine similarity. The relevance score s_i and the final normalized attention weight α_i are computed as follows:

$$s_i = \frac{q^T v_i}{\|q\|\|v_i\|\sqrt{d}}, \quad \alpha_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^{N}\exp(s_j/\tau)} \tag{2}$$

where τ is a temperature parameter and d is the embedding dimension. High α_i values highlight regions that visually support the counterfactual premise, effectively serving as evidence of absence for the original diagnosis.

3.3 Adversarial Dialectic Reasoning Loop

The reasoning process is modeled as an iterative expansion of the consensus graph $\mathcal{G}_t$.

Phase 1: Hypothesis Generation. At step $t = 0$, the Proponent analyzes I and Q to propose an initial hypothesis node h_0. The graph is initialized as $\mathcal{V}_0 = \{h_0\}, \mathcal{E}_0 = \emptyset$.

Phase 2: Adversarial Attack. In each iteration t, the Opponent employs the VFM to attack the current hypothesis h_{t-1}. It identifies the top-k regions R_k from M_{cf} and formulates a counter-evidence node e_t:

$$e_t = \text{Opponent}(I, h_{t-1}, R_k) \tag{3}$$

The Attack Strength S_{attack} is quantified by the aggregated visual grounding confidence of the counter-evidence:

$$S_{attack}(e_t) = \frac{1}{|R_k|} \sum_{r \in R_k} \alpha_r \tag{4}$$

Crucially, we distinguish the pixel-level visual attention weight α_r from the structural graph edge weight w. If $S_{attack} < \theta_{thresh}$, the attack is deemed weak, and the loop terminates (Consensus Reached). Otherwise, a falsification edge (h_{t-1}, e_t) is added to $\mathcal{G}$, directly weighted by the attack strength: $w_{h_{t-1},e_t} = S_{attack}$.

Phase 3: Mediated Revision. The Mediator $\mathcal{A}_M$ evaluates the visual grounding of e_t and generates a textual instruction ('MediatorFeedback'). Guided by this feedback, the Proponent generates a revised hypothesis h_t. A rectification edge (e_t, h_t) is added, where its weight w_{e_t,h_t} is derived by parsing the Proponent's explicitly generated confidence score $\in [0, 1]$.

$$fb_t = \text{Mediator}(h_{t-1}, e_t) \tag{5}$$

$$h_t = \text{Proponent}(h_{t-1}, e_t, fb_t) \tag{6}$$

To prioritize intuitive mechanics, we defer the rigorous mathematical formalisms of the Dynamic Consensus Graph, including cycle detection and cumulative credibility integration, to Appendix A.

3.4 Consensus Aggregation

Upon termination (via consensus or max steps T_{max}), the final diagnosis is derived by evaluating the credibility of all hypotheses in $\mathcal{G}_{final}$. We define the **Cumulative Credibility Score** $\Phi(h)$ for a leaf node h as the product of transition weights along its reasoning path π:

$$\Phi(h) = \sum_{\pi \in \Pi(h_0 \to h)} \exp\left(\frac{1}{|\pi|} \sum_{(u,v) \in \pi} \log(w_{uv}) \right) \tag{7}$$

This path-integration approach ensures that the final diagnosis D^* is the one that best survived the chain of visual falsification and rectification:

$$D^* = \arg\max_{h \in \mathcal{V}_{leaf}} \Phi(h) \tag{8}$$

The Mediator finally summarizes the winning path into explanation E.

Algorithm 1: Multi-Agent Adversarial Dialectic Reasoning (MADR)

Data: Image I, Query Q, Max turns T_{max}
Result: Final Diagnosis D^*, Explanation E

Init: $\mathcal{G} \leftarrow (\{h_0\}, \emptyset)$, $h_{curr} \leftarrow$ Proponent(I, Q);
for $t \leftarrow 1$ **to** T_{max} **do**
  // 1. Visual Falsification
  $Q_{cf} \leftarrow$ Opponent.GenProbe(h_{curr})
  $M_{cf} \leftarrow$ VFM(I, Q_{cf}); $S_{att} \leftarrow$ Eq.4
  if $S_{att} < \theta_{thresh}$ **then break** ;
  // 2. Graph Expansion
  $e_t \leftarrow$ Opponent.Argue(h_{curr}, M_{cf})
  $\mathcal{G}$.Add$(h_{curr} \to e_t, \text{w} = S_{att})$
  // 3. Mediated Revision
  $fb_t \leftarrow$ Mediator.Evaluate(h_{curr}, e_t)
  $h_{new} \leftarrow$ Prop.Revise(h_{curr}, e_t, fb_t)
  if *IsCyclic*$(\mathcal{G}, h_{new})$ **then continue** ;
  $\mathcal{G}$.Add$(e_t \to h_{new}, \text{w} = \text{Prop.Conf})$
  $h_{curr} \leftarrow h_{new}$
$D^* \leftarrow$ Aggregator$(\mathcal{G})$;
$E \leftarrow$ Mediator(D^*);
return D^*, E;

3.5 Training Objective

To enhance the VFM's ability to distinguish subtle pathological features, we introduce a **Counterfactual Grounding Loss** during fine-tuning. We construct triplets (I, Q^+, Q^-), where Q^+ matches the ground truth and Q^- is a counterfactual query. The loss encourages the attention map M^+ to align with ground truth regions while separating M^-:

$$\mathcal{L}_{CFG} = -\log \frac{e^{s(Q^+, M^+)/\tau}}{\sum_{k \in \{+,-\}} e^{s(Q^k, M^k)/\tau}} \tag{9}$$

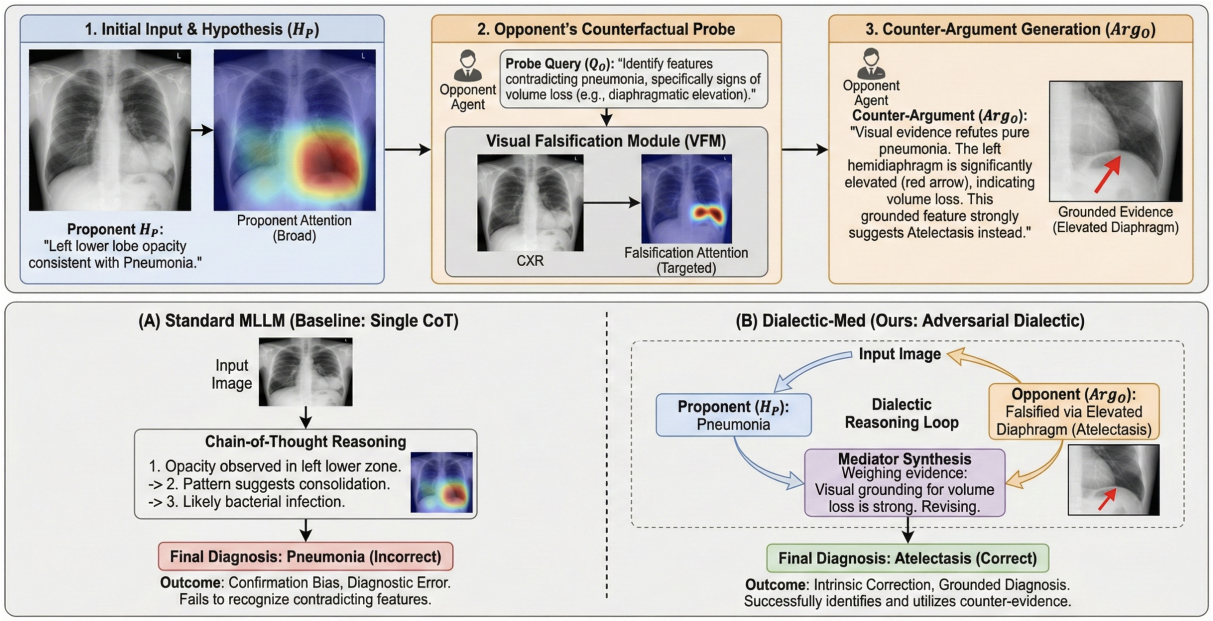

Figure 2: Qualitative comparison illustrating the Dialectic-Med inference process. The top panel details our adversarial mechanism: the *Proponent* forms an initial hypothesis (H_P, Pneumonia), while the *Opponent* generates a counterfactual probe (Q_O) to actively detect contradictory visual features via the VFM. The bottom panels compare reasoning paradigms: **(A)** Standard CoT suffers from confirmation bias, ignoring signs of volume loss; **(B)** Our adversarial dialectic framework successfully corrects the diagnosis to Atelectasis by integrating grounded counter-evidence (Arg_O) through a Mediator.

4 Experiments

4.1 Experimental Setup

Benchmarks. To ensure a rigorous evaluation across varying levels of diagnostic complexity and modality, we evaluate *Dialectic-Med* on three distinct datasets: (1) **VQA-RAD** (Lau et al., 2018): A balanced radiology benchmark comprising clinically validated question-answer pairs, serving as a foundational testbed for medical visual reasoning. (2) **PathVQA** (He et al., 2020): A challenging pathology dataset demanding fine-grained recognition of microscopic cellular structures, which rigorously tests the limits of the models' visual perception and guards against macro-level hallucination. (3) **MIMIC-CXR-VQA** (Bae et al., 2023): A curated benchmark specifically designed to probe hallucination vulnerabilities during long-context clinical reasoning. We uniquely filter this test set for complex differential diagnosis scenarios, providing a stress test for the framework's ability to reconcile conflicting visual evidence.

Baselines. We benchmark our approach against a comprehensive taxonomy of state-of-the-art systems spanning three distinct categories. We evaluate domain-specific MLLMs, namely LLaVA-Med (Li et al., 2023a) and Med-PaLM 2 (Singhal et al., 2023). To establish strong comparative reasoning, we incorporate generalist foundation models (GPT-4o, Claude-4.5 Sonnet, Gemini-3 Pro, and GPT-5.1), all prompted with standard CoT. Furthermore, we assess advanced agentic and reasoning frameworks such as MedAgent (Tang et al., 2024), ReConcile (Chen et al., 2024), and MedVCD (Mahdavi et al., 2026), which represent the current vanguard of inference-time scaling and contrastive decoding in medical AI.

Implementation Details. To ensure an equitable comparison of computational overhead, all multi-agent frameworks are strictly constrained to a maximum of $T_{max} = 3$ dialectic turns. The VFM is instantiated with a PubMedCLIP ViT-B/16 backbone (Eslami et al., 2023), fine-tuned on the ROCO dataset (Rückert et al., 2024) to optimize counterfactual image-text alignment. For the dialectic control logic, the default attack threshold (θ_{attack}) and semantic similarity threshold (θ_{sim}) are empirically set to 0.3 and 0.8, respectively. Full prompt templates detailed in Appendix C.

4.2 Comparative Analysis

Table 1 presents the quantitative performance comparison. Dialectic-Med consistently establishes new state-of-the-art results across all benchmarks,

Table 1: Quantitative results on multimodal medical reasoning benchmarks. We evaluate our proposed Dialectic-Med against specialized models, generalist foundation models, and state-of-the-art agents. #Tok indicates the average token consumption per query, and Cost ($) is estimated per 1,000 queries based on standard API pricing.

Category	Methods	Efficiency		Accuracy (%)		
		#Tok	Cost ($)	VQA-RAD	PathVQA	MIMIC-CXR-VQA
CoT Baseline	Qwen3-VL-8B	0.8k	0.05	62.50±4.12	58.31±4.01	52.21±4.33
	Qwen3-VL-32B	0.9k	0.10	68.96±2.34	62.33±3.18	58.01±3.36
	LLaVA-Med	0.7k	0.05	56.20±3.12	54.45±3.92	48.21±4.45
	Med-PaLM 2	0.8k	1.50	68.65±2.10	60.08±3.76	65.81±3.10
	Claude-4.5 Sonnet	1.2k	6.48	72.82±2.08	65.64±3.45	63.81±3.67
	GPT-4o	1.0k	4.00	74.28±1.98	66.62±2.13	65.03±3.08
	Gemini-3 Pro	1.1k	4.40	75.15±1.88	67.30±1.95	66.42±2.80
	GPT-5.1	1.2k	6.00	76.40±1.56	68.92±1.42	68.10±2.15
Medical Agents	MedAgent	15k	60.00	65.33±3.42	62.17±3.40	58.09±3.14
	ReConcile	12k	48.00	71.31±3.16	65.12±4.05	62.31±3.27
	MedVCD	8.5k	34.00	73.20±2.45	67.06±2.51	66.19±2.15
	MedMMV	9.0k	36.00	74.87±2.42	68.17±2.40	73.20±2.98
Dialectic-Med (Ours)	Qwen3-VL-8B	2.5k	0.15	70.35±2.41	64.15±3.01	61.15±3.29
	Qwen3-VL-32B	3.0k	0.35	74.62±2.02	67.40±2.67	66.81±2.85
	Claude-4.5 Sonnet	4.2k	22.68	76.65±1.77	69.92±2.69	69.85±2.76
	GPT-4o	3.8k	15.20	78.24±1.33	70.08±1.45	72.46±2.53
	Gemini-3 Pro	4.0k	16.00	79.60±1.12	71.45±1.20	74.80±1.95
	GPT-5.1	4.5k	22.50	80.45±0.95	72.32±0.88	76.28±1.75

Table 2: Ablation study on component effectiveness. To ensure rigorous attribution, we define the ablated baselines as follows: (1) **w/o Mediator**: Removes the dynamic consensus graph and feedback loop, defaulting to an unregulated text debate. (2) **w/o Graph (Last Hypothesis)**: Retains the mediator but removes the root-to-leaf path aggregation, strictly taking the final turn's utterance as the diagnosis (vulnerable to the "lost-in-the-middle" effect). (3) **w/o VFM (Text-only Debate)**: The Opponent argues textually without explicitly computing patch-level counterfactual attention maps.

Method	MIMIC-CXR-VQA		VQA-RAD	
	Acc (%)	Δ	Acc (%)	Δ
Dialectic-Med (Full)	**72.46**	–	**78.24**	–
Ablation on Dialectic Architecture				
w/o Opponent (No Falsification)	67.02	-5.44	72.91	-5.33
w/o Mediator (No Consensus)	64.35	-8.11	71.35	-6.89
Single Agent (Standard CoT)	60.50	-11.96	66.50	-11.74
Ablation on Visual Falsification Module (VFM)				
w/o VFM (Text-only Debate)	63.15	-9.31	69.25	-8.99
w/o Counterfactual Probe	65.58	-6.88	72.05	-6.19
w/o CFG Loss (Zero-shot)	69.85	-2.61	75.66	-2.58
Ablation on Consensus Graph				
w/o Graph (Last Hypothesis)	68.95	-3.51	75.05	-3.19
w/o Grounding Weights	70.66	-1.80	76.45	-1.79

demonstrating robust generalizability against both specialized and generalist baselines.

Breaking the Verification Trap. When equipped with the GPT-5.1 backbone, our framework achieves 80.45% on VQA-RAD and 76.28% on MIMIC-CXR-VQA. Notably, this surpasses the standalone GPT-5.1 baseline by absolute margins of 4.05% and 8.18%, respectively. This substantial gain corroborates our hypothesis: even the most advanced foundation models are susceptible to confirmation bias, which our dialectic loop effectively resolves by enforcing falsification.

Efficiency-Performance Trade-off. A critical finding is the exceptional performance of Dialectic-Med instantiated with the lightweight Qwen3-VL-8B backbone. Despite utilizing merely an 8B parameter model, our method achieves 70.35% on VQA-RAD. This significantly outperforms the much larger, domain-specialized LLaVA-Med (56.20%) and even surpasses the 32B base model prompted with standard CoT (68.96%).

Architectural Superiority & Fairness. A common confounding factor in multi-agent evaluations is the tight coupling of baselines with specific foundation models. To definitively prove that our performance gains stem from the fundamental architecture rather than base model biases, we conducted a rigorous mixed-model ensemble experiment. As demonstrated in Table 3, even within a heterogeneous ecosystem configured to optimally accommodate the specific design preferences of baselines, Dialectic-Med maintains a dominant and uncontested lead.

4.3 Ablation Study

To precisely attribute the sources of our performance gains, we conduct a component-wise ablation study, as detailed in Table 2.

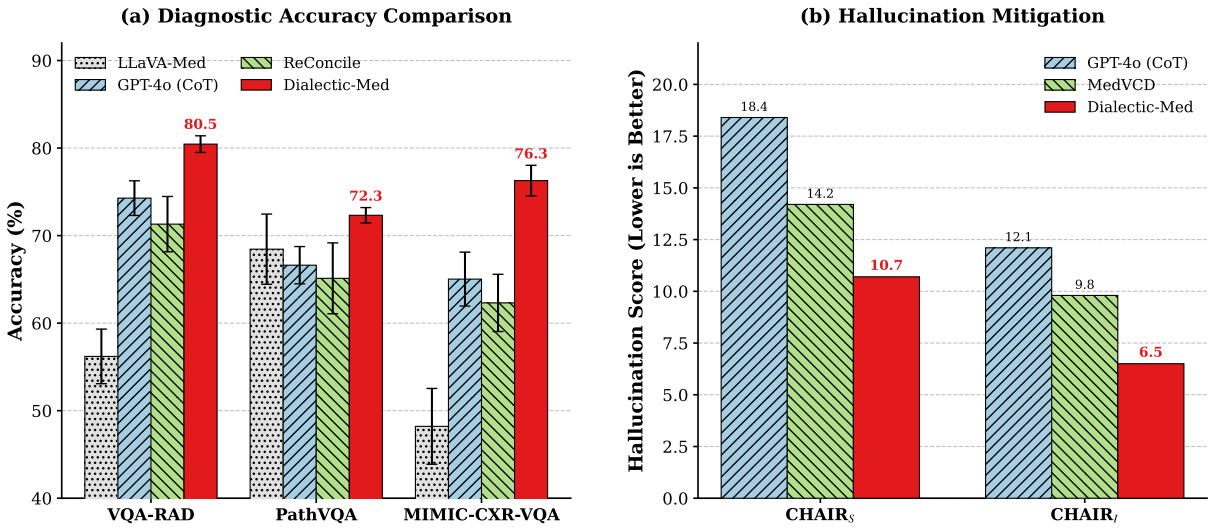

Figure 3: Main experimental results. **(a)** Comparison of diagnostic accuracy across three benchmarks. *Dialectic-Med* (Orange) consistently outperforms specialized baselines (LLaVA-Med), generalist CoT (GPT-4o), and other agentic frameworks (ReConcile), establishing a new SOTA. **(b)** Evaluation of hallucination mitigation on the MIMIC-CXR-VQA dataset using CHAIR metrics (Lower is better).

Table 3: Mixed-model architecture fairness verification. To eliminate foundation model bias, the diagnostic accuracy (Acc) of all agentic frameworks is evaluated using a heterogeneous model pool (LLaVA-Med-1.5, Qwen3-VL-32B, and GPT-4o).

Agent Framework	MIMIC-CXR-VQA	VQA-RAD
ReConcile (Mixed)	64.12%	72.85%
MedVCD (Mixed)	67.50%	74.10%
Dialectic-Med (Ours)	**71.82%**	**77.37%**

The Necessity of Visual Grounding. The most catastrophic performance degradation occurs when the VFM is removed (-9.31% on MIMIC-CXR-VQA). This finding is pivotal: it empirically proves that relying on a "textual opponent" (i.e., standard text-only multi-agent debate) is fundamentally inadequate for high-stakes medical diagnosis. Without explicit pixel-level grounding to physically anchor counter-arguments, textual agents tend to hallucinate plausible but non-existent rebuttals, thereby propagating rather than correcting the underlying diagnostic errors.

Structured Graph vs. Linear Memory. Furthermore, removing the Dynamic Consensus Graph in favor of strictly utilizing the final turn's hypothesis results in an absolute accuracy drop of 3.51%. This confirms that the weighted graph structure is strictly necessary to mitigate the "lost-in-the-middle" phenomenon, ensuring that highly confident visual counter-evidence retrieved early in the debate is not forgotten or overwritten during later consensus aggregation phases.

Hyperparameter Robustness. To address potential concerns regarding heuristic fragility, we conducted systematic sensitivity analyses on our core thresholds. As detailed in Table 4, our framework maintains optimal and stable diagnostic performance across a wide spectrum of Attack Thresholds (θ_{attack}). Furthermore, Table 5 validates our choice of the Similarity Threshold (θ_{sim}) as the Pareto optimal sweet spot: it maximizes diagnostic accuracy while effectively pruning redundant arguments, saving approximately 15% in computational debate turns compared to looser configurations.

4.4 Quantitative Analysis

While accuracy metrics effectively evaluate correct diagnostic outcomes, they fail to penalize the generation of fabricated clinical findings. To explicitly quantify the reduction in diagnostic hallucinations, we adapt the CHAIR metric (Rohrbach et al., 2018) for the medical domain. We evaluate our framework on the MIMIC-CXR-VQA test set, focusing on two variants: CHAIR$_S$ (the percentage of sentences containing at least one hallucinated finding) and CHAIR$_I$ (the percentage of hallucinated objects or findings among all mentioned entities). As presented in Table 6, Dialectic-Med demonstrates a profound improvement in clinical trustworthiness across both automated metrics and expert human evaluation.

Table 4: Sensitivity analysis of the attack threshold (θ_{attack}). Evaluated on the MIMIC-CXR-VQA dataset with $\theta_{sim} = 0.8$.

θ_{attack}	Behavioral Characteristic	Accuracy (%)	CHAIR$_I$ (↓%)
0.1	Aggressive: High false rejection rate	68.40	6.8
0.2	Stable performance range	71.85	6.4
0.3	**Default: Optimal balance**	**72.46**	**6.5**
0.4	Stable performance range	71.92	6.9
0.6	Conservative: Degenerates to CoT	66.10	11.5

Table 5: Impact of the similarity threshold (θ_{sim}) on graph expansion. Evaluated on MIMIC-CXR-VQA with $\theta_{attack} = 0.3$.

θ_{sim}	Behavioral Characteristic	Accuracy (%)	Avg. Turns
0.7	Strict: Prunes valid revisions prematurely	69.50	1.8
0.8	**Default: Optimal graph expansion**	**72.46**	**2.6**
0.9	Loose: Allows redundant arguments	72.51	3.0 (Maxed)

- **Sentence-Level Reduction:** We achieve a remarkable 41.8% relative reduction in sentence-level hallucinations (CHAIR$_S$) compared to the GPT-4o CoT baseline ($18.4\% \rightarrow 10.7\%$).
- **Object-Level Precision:** Most notably, the object-level hallucination score (CHAIR$_I$) drops by 46.3% ($12.1\% \rightarrow 6.5\%$). Because CHAIR$_I$ specifically penalizes ungrounded visual objects, this decisive drop empirically validates the efficacy of the *Visual Falsification Module*. It confirms that the Opponent agent compels the removal of fabricated visual details rather than superficially smoothing the narrative.
- **Expert Human Evaluation:** In a rigorous double-blind study conducted by three board-certified radiologists, our method achieves a Faithfulness score of $4.33/5.0$, significantly outperforming the state-of-the-art MedVCD (3.85). This robust margin indicates that our generated explanations are clinically safer, explicitly mitigating the risk of plausible but visually ungrounded assertions.

4.5 Qualitative Analysis

Figure 2 illustrates the adversarial dialectic process in action. The input is a chest X-ray exhibiting subtle radiographic signs of volume loss.

- **Initial Error:** The Proponent, misled by the global diffuse opacity, initially succumbs to confirmation bias, hypothesizing "Pneumonia" and hallucinating the presence of "patchy consolidations" in the lower lobes.
- **Visual Falsification:** The Opponent intercepts this flawed reasoning chain. Directed by the VFM, it successfully retrieves a high-attention falsification map ($\alpha > 0.8$) precisely localized on the elevated left hemidiaphragm.
- **Counter-Argument:** Leveraging this pixel-grounded counter-evidence, the Opponent constructs a targeted refutation: "The diaphragm elevation indicates volume loss, not consolidation".
- **Consensus & Rectification:** The Mediator adjudicates in favor of this visually verified logic. Consequently, the final diagnosis is robustly rectified to "Atelectasis" and the hallucinated "consolidations" are strictly purged from the final report.

This transparent "self-correction" trajectory provides a verifiable audit trail, a critical prerequisite for clinical trust that remains absent in standard single-pass "black-box" models.

Table 6: Hallucination and faithfulness analysis on MIMIC-CXR-VQA. CHAIR metrics quantify hallucination. Faithfulness is a human-evaluated score (1-5 scale) assessed by clinical experts.

Method	CHAIR$_S$ (%) ↓	CHAIR$_I$ (%) ↓	Faithfulness ↑
GPT-4o (Standard CoT)	18.4	12.1	3.42
MedVCD	14.2	9.8	3.85
Dialectic-Med (Ours)	**10.7**	**6.5**	**4.33**

5 Conclusion

In this work, we identified the "Verificationist Trap" as a primary cause of diagnostic hallucinations in medical MLLMs. To address this, we introduced Dialectic-Med, a framework that operationalizes Popperian falsification through multi-agent adversarial debate. By equipping an Opponent agent with a Visual Falsification Module, our system actively seeks to disprove hypotheses rather than merely supporting them. Comprehensive evaluations across three rigorous benchmarks (MIMIC-CXR-VQA, VQA-RAD and PathVQA) demonstrate that Dialectic-Med establishes new state-of-the-art diagnostic accuracy. Crucially, it strikes an optimal clinical sweet spot: decisively mitigating life-threatening object-level hallucinations and maximizing expert-evaluated explanation faithfulness, all while operating at a fraction of the computational cost of unconstrained agentic frameworks. Ultimately, we argue that shifting the foundational paradigm from generative confirmation to discriminative falsification is an indispensable prerequisite for deploying genuinely trustworthy and clinically safe medical AI.

Limitations

While Dialectic-Med significantly advances the trustworthiness of multimodal medical reasoning, its current instantiation is subject to several empirical and operational boundaries:

- **Inference Latency in Synchronous Settings:** Although highly cost-efficient compared to unconstrained agentic frameworks, our multi-turn dialectic loop fundamentally incurs higher latency than single-pass inference. Consequently, while it is optimally suited for asynchronous clinical workflows (e.g., retrospective radiology report generation), its deployment in ultra-time-critical emergency triage may be constrained.
- **Visual Bottlenecks and OOD Pathologies:** The efficacy of the Visual Falsification Module (VFM) is inherently upper-bounded by the visual acuity and pre-training distribution of its backbone Vision-Language Model (VLM). For extremely fine-grained micro-pathologies (e.g., $<$ 3mm nodules) or severe Out-of-Distribution (OOD) rare conditions underrepresented in the fine-tuning dataset, the Opponent may fail to retrieve localized counter-evidence, potentially defaulting to a false consensus.
- **Heuristic Sensitivity in Graph Aggregation:** Our dynamic consensus graph relies on predefined semantic similarity and attack thresholds (θ_{sim}, θ_{attack}). While we demonstrated robust Pareto optimality on our validation set, deploying the framework across highly divergent medical sub-specialties may necessitate domain-specific calibration.
- **Mediator Adjudication Errors:** Even with explicit visual grounding, the Mediator Agent remains governed by a parametric LLM. It may occasionally exhibit residual reasoning flaws or mistakenly reject logically sound visual counter-evidence due to ingrained confirmation biases inherited during foundation model pre-training.

Ethics Statement

The deployment of generative AI in high-stakes healthcare domains demands rigorous ethical scrutiny. We outline the ethical considerations of our work across three critical dimensions:

Data Privacy and Compliance. All experiments were conducted using publicly available, rigorously de-identified datasets (MIMIC-CXR-VQA, VQA-RAD, PathVQA). No Protected Health Information (PHI) was accessed, processed, or generated during the training or inference phases, ensuring strict compliance with HIPAA regulations and standard biomedical data privacy protocols.

Clinical Safety and Human-in-the-Loop. Dialectic-Med is strictly designed as a clinical decision support system (CDSS), not a surrogate for human medical professionals. The "adversarial" architecture is explicitly engineered to surface conflicting evidence and provide a transparent, verifiable reasoning trail, thereby empowering clinicians with a robust "second opinion". We expressly warn against deploying this framework for autonomous, unreviewed diagnostic decision-making.

Bias, Fairness, and Representational Equity. While our falsification mechanism effectively mitigates visual hallucinations, the underlying foundation models and the VLM backbone may still harbor latent demographic or intersectional biases present in their massive pre-training corpora. Consequently, the framework's falsification sensitivity might inadvertently vary across diverse patient populations. Future deployment prerequisites must include comprehensive fairness auditing and algorithmic bias mitigation across highly diverse clinical cohorts.

References

Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jungwoo Oh, Lei Ji, Eric Chang, Tackeun Kim, and Edward Choi. 2023. Ehrxqa: A multi-modal question answering dataset for electronic health records with chest x-ray images. In *Advances in Neural Information Processing Systems*, volume 36, pages 3867–3880. Curran Associates, Inc.

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3).

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, page 1–21, Berlin, Heidelberg. Springer-Verlag.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Alexander Ding, Jonathan Joshi, and Emily Tiwana. 2023. Patient safety in radiology and medical imaging. In *Patient Safety: A Case-based Innovative Playbook for Safer Care*, pages 261–277. Springer.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. 2023. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, Dubrovnik, Croatia. Association for Computational Linguistics.

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

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Yubin Kim, Hyewon Jeong, Shan Chen, Shuyue Stella Li, Chanwoo Park, Mingyu Lu, Kumail Alhamoud, Jimin Mun, Cristina Grau, Minseok Jung, Rodrigo Gameiro, Chunjong Park, Hyeonhoon Lee, Hae Won Park, Daniel McDuff, Samir Tulebaev, and Cynthia Breazeal. 2025. Medical hallucinations in foundation models and their impact on healthcare. *Preprint*, arXiv:2503.05777.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023b. Camel: communicative agents for "mind" exploration of large language model society. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 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*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A survey on hallucination in large vision-language models. *Preprint*, arXiv:2402.00253.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Valentin Liévin, Christoffer Egeberg Hother, Andreas Geert Motzfeldt, and Ole Winther. 2024. Can large language models reason about medical questions? *Patterns*, 5(3):100943.

Zhixiang Lu, Yulong Li, Feilong Tang, Zhengyong Jiang, Chong Li, Mian Zhou, Tenglong Li, and Jionglong Su. 2026a. Deepgb-tb: A risk-balanced cross-attention gradient-boosted convolutional network for rapid, interpretable tuberculosis screening. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(46):38989–38997.

Zhixiang Lu, Shijie Xu, Kaicheng Yan, Xuyue Cai, Chong Zhang, Yulong Li, Angelos Stefanidis, Anh Nguyen, and Jionglong Su. 2026b. Skinclip-vl: Consistency-aware vision-language learning for multimodal skin cancer diagnosis. *Preprint*, arXiv:2603.21010.

Jie Ma, Zhitao Gao, Qi Chai, Wangchun Sun, Pinghui Wang, Hongbin Pei, Jing Tao, Lingyun Song, Jun Liu, Chen Zhang, and Lizhen Cui. 2025. Debate on graph: A flexible and reliable reasoning framework for large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24768–24776.

Zahra Mahdavi, Zahra Khodakaramimaghsoud, Hooman Khaloo, Sina Bakhshandeh Taleshani, Erfan Hashemi, Javad Mirzapour Kaleybar, and Omid Nejati Manzari. 2026. Med-vcd: Mitigating

hallucination for medical large vision language models through visual contrastive decoding. *Computers in Biology and Medicine*, 200:111347.

Jing Miao, Charat Thongprayoon, Supawadee Suppadungsuk, Pajaree Krisanapan, Yeshwanter Radhakrishnan, and Wisit Cheungpasitporn. 2024. Chain of thought utilization in large language models and application in nephrology. *Medicina*, 60(1).

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

Yoojin Nam, Dong Yeong Kim, Sunggu Kyung, Jinyoung Seo, Jeong Min Song, Jimin Kwon, Jihyun Kim, Wooyoung Jo, Hyungbin Park, Jimin Sung, Sangah Park, Heeyeon Kwon, Taehee Kwon, Kanghyun Kim, and Namkug Kim. 2025. Multimodal large language models in medical imaging: Current state and future directions. *Korean Journal of Radiology*, 26(10):900–923.

Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12695–12705.

Shrey Pandit, Jiawei Xu, Junyuan Hong, Zhangyang Wang, Tianlong Chen, Kaidi Xu, and Ying Ding. 2025. MedHallu: A comprehensive benchmark for detecting medical hallucinations in large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2858–2873, Suzhou, China. Association for Computational Linguistics.

Karl Popper. 2005. *The logic of scientific discovery*. Routledge.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Johannes Rückert, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Cynthia S. Schmidt, Sven Koitka, Obioma Pelka, Asma Ben Abacha, Alba G. Seco de Herrera, Henning Müller, Peter A. Horn, Felix Nensa, and Christoph M. Friedrich. 2024. Rocov2: Radiology objects in context version 2, an updated multimodal image dataset. *Scientific Data*, 11(1).

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Alvin Babiker, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Tao Tang, Shijie Xu, Jionglong Su, and Zhixiang Lu. 2026. Causal-sam-llm: Large language models as causal reasoners for robust medical segmentation. *Preprint*, arXiv:2507.03585.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Zhu Zhang, Zhou Zhao, Zhijie Lin, jieming zhu, and Xiuqiang He. 2020. Counterfactual contrastive learning for weakly-supervised vision-language grounding. In *Advances in Neural Information Processing Systems*, volume 33, pages 18123–18134. Curran Associates, Inc.

Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, and Jianfeng Gao. 2022. Regionclip: Region-based language-image pretraining. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16772–16782.

Zhihong Zhu, Yunyan Zhang, Xianwei Zhuang, Fan Zhang, Zhongwei Wan, Yuyan Chen, Qingqing Long, Yefeng Zheng, and Xian Wu. 2025. Can we trust AI doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6748–6769, Vienna, Austria. Association for Computational Linguistics.

A Algorithm Summary

The overall process is summarized in Algorithm 2.

A.1 Dynamic Consensus Graph Formulation

To explicitly model the dialectic trajectory and mitigate the "lost-in-the-middle" phenomenon common in long-context reasoning (Liu et al., 2024b), we introduce the **Dynamic Consensus Graph**, denoted as $\mathcal{G}_t = (\mathcal{V}_t, \mathcal{E}_t)$ at time step t. This Directed Acyclic Graph (DAG) serves as a structured memory bank that records the evolution of diagnostic hypotheses under adversarial scrutiny, enabling principled consensus aggregation (Ma et al., 2025).

A.1.1 Graph Definition

The node set $\mathcal{V}_t$ consists of two distinct types of logical entities:

- **Hypothesis nodes** $h \in \mathcal{H}$: Diagnostic hypotheses generated by the Proponent agent $\mathcal{A}_P$, each associated with a semantic embedding $\mathbf{h} \in \mathbb{R}^d$.
- **Counter-Evidence nodes** $e \in \mathcal{E}_{\text{vid}}$: Visually-grounded counter-arguments derived by the Opponent agent $\mathcal{A}_O$ via the Visual Falsification Module.

The edge set $\mathcal{E}_t$ represents the logical transitions between nodes. We define two types of directed edges:

- **Falsification edges** (h_i, e_j): Connecting a hypothesis to the counter-evidence that challenges it.
- **Rectification edges** (e_j, h_k): Connecting counter-evidence to the revised hypothesis it prompted.

Associated with each edge $(u, v) \in \mathcal{E}_t$ is a transition weight $w_{uv} \in [0, 1]$, quantifying the confidence of the logical inference. At $t = 0$, the graph is initialized with the primary hypothesis h_0:

$$\mathcal{V}_0 = \{h_0\}, \quad \mathcal{E}_0 = \emptyset \tag{10}$$

A.1.2 Adversarial Graph Expansion

In each dialectic iteration t, given the current hypothesis h_{t-1}, the Opponent $\mathcal{A}_O$ mines a visual counter-evidence e_t using the Visual Falsification Module. The Mediator $\mathcal{A}_M$ evaluates the validity of this attack by computing the **Attack Strength** S_{attack}, which measures the visual grounding confidence of the counter-argument:

$$S_{\text{attack}}(e_t) = \frac{1}{|R(e_t)|} \sum_{k \in R(e_t)} \alpha_k^{(t)} \tag{11}$$

where $R(e_t)$ denotes the set of image patch indices referenced by the counter-evidence e_t, and $\alpha_k^{(t)}$ is the attention weight from the falsification attention map $M_F^{(t)}$.

If $S_{\text{attack}}(e_t) \geq \theta_{\text{attack}}$ (i.e., the attack is deemed valid), the Proponent generates a revised hypothesis h_t, and the graph is expanded:

$$\mathcal{V}_t = \mathcal{V}_{t-1} \cup \{e_t, h_t\} \tag{12}$$

$$\mathcal{E}_t = \mathcal{E}_{t-1} \cup \{(h_{t-1}, e_t), (e_t, h_t)\} \tag{13}$$

The edge weights are assigned as follows:

- The *falsification edge* (h_{t-1}, e_t) is weighted by the attack strength: $w_{h_{t-1}, e_t} = S_{\text{attack}}(e_t)$.
- The *rectification edge* (e_t, h_t) is weighted by the Proponent's revised confidence: $w_{e_t, h_t} = \text{conf}(h_t | e_t)$, where $\text{conf}(\cdot)$ is derived from the Proponent's output logits.

A.1.3 Path Integration

To derive the final diagnosis D_{final}, we perform a probabilistic aggregation over all paths in the final graph $\mathcal{G}_T$. Let $\Pi(h_{\text{leaf}})$ denote the set of all directed paths from the root hypothesis h_0 to a leaf hypothesis node $h_{\text{leaf}} \in \mathcal{H}_{\text{leaf}}$. The cumulative **Credibility Score** Φ for a terminal hypothesis is calculated by integrating the transition weights along each reasoning chain:

$$\Phi(h_{leaf}) = \sum_{\pi \in \Pi(h_{leaf})} \exp\left(\frac{1}{|\pi|} \sum_{(u,v) \in \pi} \log(w_{uv}) \right) \tag{14}$$

This formulation ensures that a diagnosis is only reliable if it survives the adversarial loop with high-confidence transitions at every step. Intuitively, a hypothesis that was revised due to strongly grounded counter-evidence (high $w_{h,e}$) and then confidently restated (high $w_{e,h'}$) will accumulate a high credibility score.

The final diagnosis is selected via:

$$D_{\text{final}} = \arg\max_{h \in \mathcal{H}_{\text{leaf}}} \Phi(h) \tag{15}$$

The confidence of the final diagnosis is normalized:

$$\text{Confidence}(D_{\text{final}}) = \frac{\Phi(D_{\text{final}})}{\sum_{h \in \mathcal{H}_{\text{leaf}}} \Phi(h)} \tag{16}$$

Cycle Detection and Pruning. To maintain the DAG property and prevent infinite loops, if the dialectic process proposes a hypothesis h_t that is semantically equivalent to a previously refuted hypothesis h_j (i.e., $\text{sim}(h_t, h_j) > \theta_{\text{sim}}$), the branch is pruned. This is implemented by checking against all existing hypothesis nodes before adding h_t to $\mathcal{V}_t$.

A.1.4 Explanation Generation

The final explanation E is constructed by tracing the highest-scoring path in $\mathcal{G}_T$ that leads to D_{final}. The Mediator $\mathcal{A}_M$ summarizes this path, explicitly referencing the key counter-evidence nodes and the visual regions they highlighted. This ensures the explanation is both comprehensive and verifiable, providing a transparent audit trail of the diagnostic reasoning process.

A.2 Inference Algorithm

The complete inference process of our DialecticMed framework is outlined in Algorithm 2. The algorithm orchestrates the three agents through the adversarial dialectic loop, dynamically updates the consensus graph, and finally aggregates the results to produce a robust diagnosis.

Complexity Analysis. The time complexity of Algorithm 2 is $O(T_{\max} \cdot (C_{\text{VFM}} + C_{\text{LLM}}))$, where C_{VFM} is the cost of a single VFM forward pass and C_{LLM} is the cost of an LLM inference. The graph operations (adding nodes, edges, and computing credibility) are $O(|\mathcal{V}|^2)$ in the worst case, but since $|\mathcal{V}| \leq 2T_{\max} + 1$, this is dominated by the LLM inference cost. In practice, we set $T_{\max} \in [3, 5]$, making the overhead minimal compared to single-pass MLLM inference while significantly improving diagnostic reliability.

A.3 Visual Falsification Module (VFM)

The Visual Falsification Module is the cornerstone of the Opponent agent's ability to challenge the Proponent. It is designed to operationalize the principle of falsification by actively seeking and localizing visual evidence that contradicts a given diagnostic hypothesis. The VFM comprises two main components: a Counterfactual Probe Generator and a Grounded Attention Mechanism, which are fine-tuned using a novel counterfactual grounding objective.

Algorithm 2: Adversarial Dialectic Reasoning with Consensus Graph

Data: Medical Image I, User Query Q, Max Iterations $T_{\max}$, Attack Threshold θ_{attack}, Similarity Threshold θ_{sim}

Result: Final Diagnosis D_{final}, Explanation E, Confidence C

Initialize: Agents $\mathcal{A}_P, \mathcal{A}_O, \mathcal{A}_M$;
$h_0 \leftarrow \mathcal{A}_P.\text{Generate}(I, Q)$;
$\mathcal{G} \leftarrow \text{InitGraph}(h_0)$; $\mathcal{V} \leftarrow \{h_0\}$; $\mathcal{E} \leftarrow \emptyset$;
$h_{\text{current}} \leftarrow h_0$; $t \leftarrow 1$;
while $t \leq T_{\max}$ **do**
  $Q_O^{(t)} \leftarrow \mathcal{A}_O.\text{GenerateProbe}(h_{\text{current}})$;
  $M_F^{(t)} \leftarrow \text{VFM}(I, Q_O^{(t)})$;
  $e_t \leftarrow \mathcal{A}_O.\text{Generate}(h_{\text{current}}, M_F^{(t)})$;
  $S_{\text{attack}} \leftarrow \text{Attack}(e_t, M_F^{(t)})$;
  if $S_{attack} < \theta_{attack}$ **then break**; ;
  $h_t \leftarrow \mathcal{A}_P.\text{Revise}(h_{\text{current}}, e_t)$;
  if $\exists h_j \in \mathcal{V} : sim(h_t, h_j) > \theta_{sim}$ **then continue**; ;
  $\mathcal{V} \leftarrow \mathcal{V} \cup \{e_t, h_t\}$;
  $\mathcal{E} \leftarrow \mathcal{E} \cup \{(h_{\text{current}}, e_t), (e_t, h_t)\}$;
  $w_{h_{\text{current}}, e_t} \leftarrow S_{\text{attack}}$;
  $w_{e_t, h_t} \leftarrow \mathcal{A}_P.\text{Confidence}(h_t)$;
  $h_{\text{current}} \leftarrow h_t$; $t \leftarrow t + 1$;
$\mathcal{H}_{\text{leaf}} \leftarrow \text{GetLeafHypotheses}(\mathcal{G})$;
foreach $h \in \mathcal{H}_{leaf}$ **do**
  $\Phi(h) \leftarrow \text{Credibility}(h, \mathcal{G})$;
$D_{\text{final}} \leftarrow \arg\max_{h \in \mathcal{H}_{\text{leaf}}} \Phi(h)$;
$C \leftarrow \Phi(D_{\text{final}}) / \sum_h \Phi(h)$;
$E \leftarrow \mathcal{A}_M.\text{SummarizePath}(\mathcal{G}, D_{\text{final}})$;
return D_{final}, E, C;

A.3.1 Network Architecture

The VFM leverages a pre-trained VLM with a Vision Transformer (ViT) backbone (PubMedCLIP (Eslami et al., 2023)), which is specifically adapted for the medical domain. The architecture includes:

- **Visual Encoder ($\mathcal{E}_v$):** A ViT model that processes an input image I by dividing it into a sequence of N flattened 2D patches, $\{p_1, p_2, ..., p_N\}$. Each patch is linearly projected into a patch embedding. Including the [CLS] token, the output is a sequence of patch embeddings $\mathbf{V} = \{v_{\text{cls}}, v_1, ..., v_N\} \in \mathbb{R}^{(N+1) \times D}$, where D is the embedding dimension.

- **Textual Encoder ($\mathcal{E}_t$):** A Transformer-based text encoder that processes the counterfactual probe query Q_O and outputs a sentence embedding $\mathbf{q} \in \mathbb{R}^D$.
- **Cross-Modal Attention Layer:** A standard cross-modal attention mechanism (Lu et al., 2026a) that computes the similarity between the textual probe embedding and each of the visual patch embeddings.

This dual-encoder architecture allows us to project both the image regions and the textual query into a shared embedding space, enabling fine-grained, region-level semantic matching (Zhong et al., 2022).

A.3.2 Mathematical Formulation

The core function of the VFM is to generate a **Falsification Attention Map** (M_F) that highlights image regions inconsistent with the Proponent's hypothesis H_P. This is achieved through a two-step process:

Counterfactual Probe Generation. Given H_P, the Opponent agent $\mathcal{A}_O$ first generates a textual counterfactual probe Q_O. This is not a simple negation, but a targeted query for contradictory evidence, leveraging medical domain knowledge $\mathcal{K}$. For a hypothesis H_P = "Left lower lobe opacity consistent with Pneumonia", the generation process can be modeled as:

$$Q_O = \mathcal{A}_O(\mathcal{T}_{\text{probe}}(H_P, \mathcal{K})) \tag{17}$$

This yields a probe like: "*Identify features contradicting pneumonia, specifically signs of volume loss*".

Grounded Attention via Cross-Modal Similarity. The VFM then grounds this probe Q_O in the image I. We compute the cross-modal similarity between the probe's text embedding $\mathbf{q} = \mathcal{E}_t(Q_O)$ and each visual patch embedding $v_i \in \mathbf{V}$. The relevance score s_i for each patch i is calculated using the scaled dot-product attention mechanism (Vaswani et al., 2017):

$$s_i = \frac{\mathbf{q}^T v_i}{\|\mathbf{q}\|\|v_i\|\sqrt{d}} \tag{18}$$

where the cosine similarity is scaled by the square root of the embedding dimension d to stabilize gradients. These scores represent the semantic alignment between the counterfactual probe and each image region.

The raw scores are then normalized using a softmax function to produce the final Falsification Attention Map $M_F = \{\alpha_1, \alpha_2, ..., \alpha_N\}$:

$$\alpha_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^{N} \exp(s_j/\tau)} \tag{19}$$

where τ is a temperature parameter. A lower τ produces a sharper attention map, focusing on the most salient contradictory evidence. This process is analogous to generating a saliency map, but instead of highlighting regions that *support* a classification, it highlights regions that *falsify* it, similar to techniques used in counterfactual visual explanations.

Generating Counter-Evidence. The Opponent agent $\mathcal{A}_O$ uses this attention map to generate its counter-argument Arg_O. It focuses on the top-k patches with the highest attention weights, denoted as $R_k = \{\text{patch}_i | \alpha_i \in \text{TopK}(\alpha)\}$. The final counter-argument is generated by prompting the agent with the original hypothesis and the localized visual evidence:

$$Arg_O = \mathcal{A}_O(\mathcal{T}_{\text{refute}}(H_P, R_k)) \tag{20}$$

This ensures that the Opponent's argument is not a generic rebuttal but is directly and verifiably grounded in specific, localized visual features, as shown in the case study in Figure 2 where the falsification attention correctly localizes the elevated diaphragm.

A.3.3 Training Objective

While the VFM can operate in a zero-shot manner using a pre-trained VLM, its ability to localize subtle, contradictory findings can be significantly enhanced through fine-tuning. We propose a novel **Counterfactual Grounding Loss** ($\mathcal{L}_{\text{CFG}}$) designed to train the VFM to explicitly distinguish between visual evidence that supports a hypothesis and evidence that falsifies it.

To construct training triplets, we require a dataset with bounding box annotations for both positive and negative findings. Given an image I, a diagnostic hypothesis H_P, a set of ground-truth bounding boxes B_P for findings consistent with H_P, and a set of bounding boxes B_O for findings that contradict H_P, we can formulate our training objective.

Let M_P be the attention map generated by a standard probe for H_P (e.g., "Find signs of pneumonia"), and let M_F be the falsification attention

map generated by the counterfactual probe Q_O. Our goal is to train the model such that:

- The attention map M_P focuses on the regions defined by B_P.
- The falsification map M_F focuses on the regions defined by B_O.

We adapt the contrastive learning framework from (Zhang et al., 2020) for this purpose. For a given hypothesis H_P, we define:

- **Positive Alignment Score (S^+):** The aggregated attention from the standard probe map M_P within the positive bounding boxes B_P.
- **Negative Alignment Score (S^-):** The aggregated attention from the standard probe map M_P within the *negative* (contradictory) bounding boxes B_O.

Our first loss term, the **Proponent Grounding Loss** ($\mathcal{L}_P$), encourages the standard attention map to focus on the correct evidence:

$$\mathcal{L}_P = -\log \frac{\exp(S^+/\tau)}{\exp(S^+/\tau) + \exp(S^-/\tau)} \quad (21)$$

Symmetrically, for the falsification map M_F generated by the counterfactual probe Q_O, we define:

- **Falsification Alignment Score (S_F^+):** The aggregated attention from the falsification map M_F within the contradictory bounding boxes B_O.
- **Falsification Misalignment Score (S_F^-):** The aggregated attention from the falsification map M_F within the *positive* bounding boxes B_P.

Our second loss term, the **Opponent Grounding Loss** ($\mathcal{L}_O$), trains the VFM to correctly localize the counter-evidence:

$$\mathcal{L}_O = -\log \frac{\exp(S_F^+/\tau)}{\exp(S_F^+/\tau) + \exp(S_F^-/\tau)} \quad (22)$$

The total Counterfactual Grounding Loss is the sum of these two components:

$$\mathcal{L}_{\text{CFG}} = \lambda_P \mathcal{L}_P + \lambda_O \mathcal{L}_O \quad (23)$$

where λ_P and λ_O are hyperparameters to balance the two objectives. By minimizing this loss, we explicitly train the VFM to not only identify supporting evidence but also to become an expert at seeking out and localizing contradictory visual signals. This dual objective is critical for breaking the cycle of confirmation bias that plagues standard MLLMs.

The overall training objective for the entire Dialectic-Med framework combines this with the standard language modeling loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \beta \mathcal{L}_{\text{CFG}} \quad (24)$$

where β is a balancing hyperparameter.

B Details of Expert Human Evaluation

To rigorously assess the clinical safety and explanation faithfulness of our framework, we conducted a double-blind human evaluation as reported in Section 4.4. To comply with responsible research guidelines, we provide the following logistical details:

Recruitment and Compensation. The evaluation was conducted by three board-certified clinical radiologists with an average of 8 years of clinical experience. They were recruited through existing academic and clinical collaboration networks. Participants were compensated for their time at standard institutional clinical consulting rates, which is commensurate with their professional expertise and demographic location.

Evaluation Protocol and Instructions. The task did not involve human subjects research, as evaluators only assessed AI-generated textual outputs based on de-identified public images. Evaluators were presented with the input X-ray, the ground-truth report findings, and the reasoning trajectories generated by different anonymized models (Dialectic-Med, MedVCD, and standard GPT-4o CoT). They were instructed to score the *Explanation Faithfulness* on a 1-5 scale based on the following rubric:

- **1 (Dangerous):** The explanation contains severe, visually ungrounded hallucinations that could lead to clinical harm.
- **3 (Mixed):** The reasoning is generally correct but includes minor fabricated visual details.
- **5 (Highly Faithful):** The reasoning is strictly grounded in verified visual evidence, accurately addressing counterfactuals without any hallucinations.

C Prompt Example of Agents

Proponent Agent ($\mathcal{A}_P$)

Role: Similar to a primary care physician, focusing on the **Global Context** of the image.
Task: Propose an initial hypothesis or revise the hypothesis when facing counter-evidence.

SYSTEM PROMPT

```
You are an experienced Radiologist acting as the "Proponent Agent". Your goal
is to provide the most probable diagnosis based on the medical image analysis.
```

Guidelines:

- **Focus on Global Context:** Look at the overall opacity, lung volume, and heart size.
- **Be Open-Minded:** You will receive counter-arguments from an Opponent. If their visual evidence is strong, acknowledge it and revise your hypothesis.
- **Logical Reasoning:** Explain your reasoning step-by-step before giving the final diagnosis label.

USER PROMPT (Iteration $t = 0$: Initialization)

```
<Image Context>
Global Visual Features: {{GLOBAL_FEATURES_DESCRIPTION}}
User Query: {{USER_QUERY}}
</Image Context>
```

Based on the global visual features, provide an initial hypothesis (H_0).
Output Format:

- Reasoning: [Your analysis]
- Hypothesis: [Diagnosis Name]
- Confidence: [0-100%]

USER PROMPT (Iteration $t > 0$: Revision)

```
Current Hypothesis (H_{t-1}): {{CURRENT_HYPOTHESIS}}
Opponent's Counter-Argument: "{{OPPONENT_ARGUMENT}}"
Local Visual Evidence: {{LOCAL_VISUAL_FEATURES}}
```

Instruction: The Opponent claims the local evidence contradicts your hypothesis.

1. Evaluate if the counter-argument is valid.
2. If valid, propose a Revised Hypothesis (H_t) that explains both global context and local detail.
3. If invalid, defend your original hypothesis.

Opponent Agent ($\mathcal{A}_O$)

Role: Similar to a "Medical Auditor", focusing on **Visual Falsification**.
Task: Use a "Visual Probe" to find local features that contradict the current hypothesis.

SYSTEM PROMPT

```
You are a critical Medical Auditor acting as the "Opponent Agent". Your ONLY
goal is to FALSIFY the current diagnosis hypothesis. You utilize a "Visual
Falsification Module" to probe specific regions.
```

Guidelines:

- **Seek Contradictions:** Do not look for confirming evidence. Look for what is WRONG with the hypothesis.
- **Focus on Local Detail:** Use the provided local visual probe data (e.g., costophrenic angles).
- **Be Sharp:** Your argument must be grounded in the visual evidence provided.

USER PROMPT

Current Hypothesis (H_t): {{CURRENT_HYPOTHESIS}}
Visual Probe Target: ROI focused on {{ROI_NAME}}.
Local Visual Features: {{LOCAL_FEATURES_DESCRIPTION}}

Instruction: Does the visual evidence in this ROI contradict the Current Hypothesis?

- If hypothesis is "Pneumonia" (implies opacity), but ROI shows "Sharp Costophrenic Angle", this is a contradiction.
- If hypothesis is "Normal", but ROI shows "Nodule", this is a contradiction.

Generate a Counter-Argument (Arg_{opp}).
Output Format:

- Observation: "I see [Visual Feature] in the [Region]..."
- Contradiction: "This contradicts [Hypothesis] because [Reason]..."
- Counter-Evidence Strength: [High/Medium/Low]

Mediator Agent ($\mathcal{A}_M$)

Role: Similar to a Senior Consultant or Judge, responsible for **Consensus Aggregation**.
Task: Evaluate the debate quality, decide on revisions, and determine consensus.

SYSTEM PROMPT

```
You are the Chief Medical Consultant acting as the "Mediator Agent". You oversee
a dialectic debate between a Proponent and an Opponent. Your job is to manage
the "Consensus Graph".
```

Guidelines:

- **Evaluate Validity:** Is the Opponent's counter-evidence visually grounded and logically sound?
- **Manage State:** Decide whether to Refute the current hypothesis or Sustain it.
- **Terminate:** If the debate converges or no new counter-evidence is found, declare CONSENSUS.

USER PROMPT

Debate History:
1. Proponent Hypothesis (H_{t-1}): {{OLD_HYPOTHESIS}}
2. Opponent Counter-Argument: {{OPPONENT_ARGUMENT}}
3. Proponent Revised Argument: {{PROPONENT_RESPONSE}}

Instruction: Analyze the interaction.

- Did the Proponent successfully defend their hypothesis?
- Or did the Opponent successfully force a revision?
- Is the new diagnosis consistent with all evidence seen so far?

Output JSON:

```
{
  "status": "CONTINUE" or "CONSENSUS",
  "winner": "PROPONENT" or "OPPONENT",
  "current_best_diagnosis": "...",
  "confidence_score": 0.0 to 1.0,
  "explanation": "Summarize why the consensus was reached..
}
```