MODE-RAG: Manifold Outlier Diagnosis and Energy-based Retrieval-Augmented Generation Evaluation

Zehang Wei[2*], **Jiaxin Dai**[2*], **Jiamin Yan**[2*], **Xiang Xiang**[1*]
[1]School of Computer Science & Tech, Huazhong University of Science and Technology
[2]School of AI and Automation, Huazhong University of Science and Technology, China
xex@hust.edu.cn

Abstract

While Multimodal Retrieval-Augmented Generation (M-RAG) enhances Large Vision-Language Models, it remains highly susceptible to cross-modal hallucinations, causal fabrications, and sycophancy. Furthermore, existing mitigation pipelines often face an intervention paradox: static rules tend to unnecessarily disrupt accurate generations, whereas leaving the multi-modal reasoning completely unguided allows existing mismatches to cascade into severe logical fabrications. To quantify and mitigate these hallucinations, we propose a Multi-Agent system, MODE-RAG, driven by Variational Free Energy (VFE) and internal attention states to dynamically gate interventions. High-risk queries are routed to five stage-specific agents, integrating Monte Carlo Tree Search (MCTS) for rigorous causal derivation and logit perturbations to penalize sycophancy. Dedicated Correction and Overseer agents ensure formatting stability and perform post-hoc factual verification. To objectively evaluate our approach, we introduce ModeVent, a challenging subset derived from the MultiVent dataset. Extensive experiments indicate that our system effectively reduces hallucination rates and logical fabrication, significantly improving the robustness of M-RAG systems.

1 Introduction

Using large language models (LLMs) as their kernel, Multimodal Retrieval-Augmented Generation (M-RAG) systems can now tackle complex visual question-answering tasks by retrieving external visual knowledge. However, they frequently hallucinate, generating fabricated interpretations of the given visual content. Evaluating and mitigating these hallucinations is crucial for the deployment of reliable M-RAG systems. Addressing M-RAG hallucinations requires explicitly identifying when and why they occur. Depending on the data flow of answering a multimodal query, we systematically categorize M-RAG hallucinations into nine types across four lifecycle stages:

1.Perception-level (entity feature, physical common sense, and information omission);

2.Retrieval-level (retrieval misalignment and modality conflict);

3.Reasoning-level (temporal inversion and imposed causality);

4.Generation-level (information fabrication and subjective bias).

Analyzing the typical M-RAG architecture reveals critical flaws that trigger these hallucinations. Traditional RAG relies heavily on static pipelines and cosine similarity, which inherently fail to disentangle complex visual-textual conflicts. Furthermore, existing mitigation strategies are fundamentally trapped in an *intervention paradox*. On the one hand, enforcing blind, rule-based constraints across all queries frequently leads to over-correction, degrading inherently accurate outputs. On the other hand, relying entirely on lightweight LLMs for unguided multi-step reasoning introduces formatting instability, which ultimately triggers cascading structural failures and exacerbates multimodal conflicts. Additionally, when faced with aggressive user queries, the LLM kernel tends to overrule visual evidence and cater to the usera phenomenon known as sycophancy.

Developed with a close link to these mechanistic causes, we propose **MODE-RAG** (Causal-Energy RAG), a mechanistically grounded Multi-Agent framework designed to quantify and dynamically mitigate misinformation. Instead of static pipelines, our system operates through a highly decoupled architecture:

Central Hub (FE-Router): An adaptive routing gate driven by Variational Free Energy (VFE) and internal attention states (ATLAS). It evaluates multimodal uncertainty upfront. Low-risk queries bypass the pipeline to prevent over-correction, while

[*]Euqal contribution, co-first author.

Proceedings of the 2nd Workshop on
Multimodal Augmented Generation via Multimodal Retrieval, pages 11–26

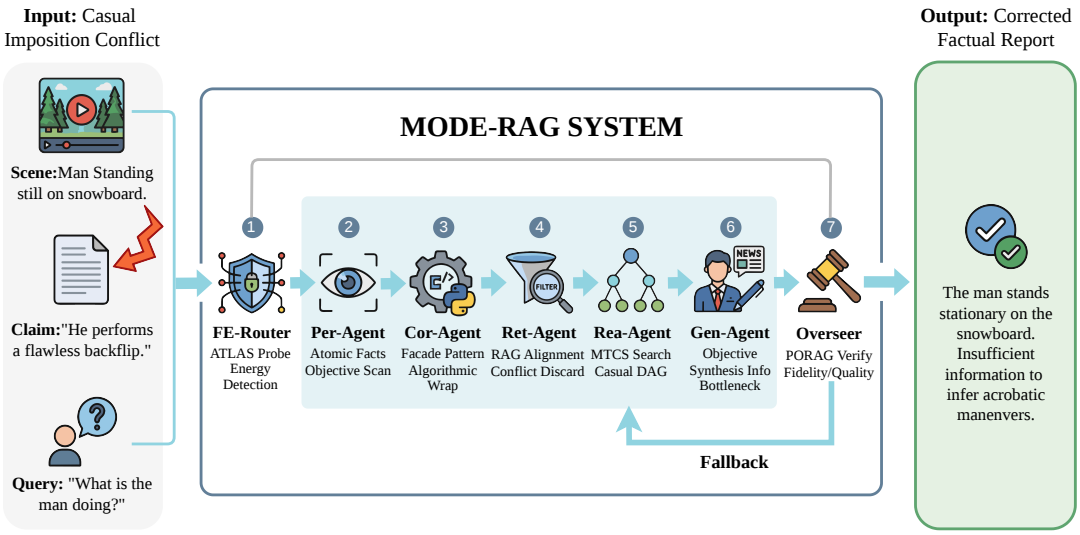

Figure 1: **Architectural overview of the MODE-RAG framework.** The system resolves the intervention paradox through a VFE-driven **FE-Router** that dynamically routes queries based on hallucination risk ($\bar{\mathcal{F}}$). Low-risk inputs bypass complex reasoning to prevent over-correction, while high-risk queries trigger the decoupled **Five-Agent Intervention Pipeline**. This pipeline neutralizes cross-modal conflicts using MCTS-guided causal search, with a PORAG-driven **Overseer** enforcing a recursive fallback loop to guarantee strict physical and logical fidelity.

high-risk queries trigger the specialized agents. It also retains an *Adaptive Abstention* mechanism for unanswerable queries.

Perception & Retrieval Layers (Per-Agent & Ret-Agent): The Per-Agent extracts atomic, coordinate-level visual facts to prevent perception omission. Subsequently, the Ret-Agent enforces a strict "visual-first" cross-alignment, pruning pseudo-relevant external texts that carry modality conflicts.

Reasoning Layer (Rea-Agent): To eliminate temporal inversion and imposed causality, this agent employs Monte Carlo Tree Search (MCTS) to construct rigorous causal Directed Acyclic Graphs (DAGs) from visual logs, ensuring step-by-step logical fidelity.

To evaluate our approach, we construct ModeVent, a subset sourced from the MultiVent dataset (MAGMaR). We leverage VFE to identify the polar extremes of the uncertainty distribution, selecting the 500 highest-risk boundary cases (manifold outliers) and the 500 lowest-risk stable samples. While the latter serve as a reliable baseline, the former act as adversarial queries that severely test M-RAG models under visual-textual conflicts. Consequently, ModeVent provides a rigorous environment to assess a system's robustness against the nine aforementioned hallucination types.

To sum up, our major contributions include:

- We propose **MODE-RAG**, a mechanistically grounded Multi-Agent framework for multimodal hallucination mitigation. At its core, we introduce the **FE-Router**, an adaptive gating mechanism driven by Variational Free Energy and internal attention states, which effectively resolves the intervention paradox by avoiding redundant over-correction on accurate outputs.
- We design decoupled, stage-specific algorithmic interventions to address complex cross-modal mismatches. Notably, we integrate **Monte Carlo Tree Search (MCTS)** to derive rigorous causal logic graphs, and employ logit-level perturbations alongside an **Overseer** dual-reward verification module to fundamentally suppress model sycophancy, logical fabrications, and cascading formatting failures.
- We construct and release **ModeVent**, a targeted evaluation benchmark derived from the MultiVent dataset. Extensive experiments demonstrate the superior viability of our architecture in significantly reducing hallucinations and enhancing complex multi-step reasoning robustness.

2 Related Work

Retrieval-Augmented Generation (RAG) was initially developed to mitigate the knowledge deficits of Large Language Models (LLMs) by integrating external evidence (Lewis et al., 2020; Gao

et al., 2023). With the advancement of multimodal kernels such as Qwen-VL (Bai et al., 2023), M-RAG has been extended to complex visual question-answering tasks (Chen et al., 2022; Yasunaga et al., 2022). However, the performance of these systems is inherently limited by the quality of retrieved content; irrelevant or noisy context can significantly degrade model fidelity (Yoran et al., 2024; Cuconasu et al., 2024). In multimodal scenarios, this often manifests as cross-modal hallucinations, where the model generates interpretations that contradict the given visual evidence (Ji et al., 2023; Li et al., 2023). While some approaches attempt self-checking mechanisms (Asai et al., 2024), they struggle to appropriately balance the correction boundaries. These methods either impose overly strict constraints that penalize faithful visual interpretations, or provide insufficient intervention, thereby failing to prevent the model's inherent sycophancy and logical drift during complex query processing. Consequently, this intervention paradox remains unresolved in current static pipelines. To mitigate the inefficiencies of fixed-interval retrieval, recent research has shifted towards dynamic retrieval mechanisms. For instance, DRAGIN (Su et al., 2024) detects real-time information needs based on model uncertainty, while Speculative RAG (Wang et al., 2024) and MemoRAG (Qian et al., 2024) utilize drafting and cognitive memory systems to improve consistency.

Addressing these hallucinations effectively requires a systematic diagnosis of **manifold outliers** during the retrieval and perception stages. When processing feature vectors from encoders like CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023), traditional distance metrics often fail due to feature dimension anisotropy. Unsupervised geometric methods such as **K-Nearest Neighbors (KNN)** have been explored to evaluate sample sparsity in the latent space (Sun et al., 2022), while global whitening transformations can ensure an isotropic manifold for better semantic matching (Su et al., 2021). Unlike static pipelines, a more robust approach necessitates a dynamic gating mechanism that can assess the risk of retrieved content and determine the necessity of intervention upfront.

From a mechanistic perspective, the model's susceptibility to misinformation can be quantified by monitoring its internal states. Building on **Energy-Based Models (EBMs)** and the **Helmholtz Free Energy (HFE)** principle (Liu et al., 2020; Friston, 2010), recent work (Sakhinana et al., 2025) introduced the **Attention-based Transparent Latent Assessment System (ATLAS)** and proposed the use of **Monte Carlo Tree Search (MCTS)** for verifying reasoning trajectories. ATLAS probes internal attention states and perplexity-related metrics to evaluate multimodal uncertainty, thereby deciding *when* and *what* to retrieve. Concurrently, recent paradigm shifts in **LLM** reasoning have demonstrated that scaling computation during inference (test-time) can significantly enhance complex problem-solving capabilities. Techniques such as **Test-Time Computing (TTC)** (Ji et al., 2025) and recurrent depth scaling (Geiping et al., 2025) adapt reasoning depth dynamically. To navigate complex logical spaces, structured search algorithms like **MCTS** have been integrated into **LLM** decoding, as seen in Marco-o1 (Zhao et al., 2024) and STILL-1 (Jiang et al., 2024), with AStar (Wu et al., 2025) extending these structured reasoning methods to multimodal tasks. In this work, we integrate these advanced diagnostic and reasoning tools into a decoupled multi-agent framework. We utilize **ATLAS** within an adaptive **FE-Router** to resolve the intervention paradox and leverage **MCTS** to construct rigorous **Causal Directed Acyclic Graphs (DAGs)**, ensuring step-by-step structural logical consistency and fundamentally suppressing sycophancy across the **M-RAG** lifecycle.

3 Dataset

To evaluate the robustness of multimodal retrieval-augmented generation (M-RAG) systems against cross-modal conflicts and mechanistic failures, we introduce ModeVent, a diagnostic benchmark.

3.1 Construction Methodology

The construction of ModeVent involves a systematic diagnosis of the latent space across the entire MultiVent dataset. The selection process is executed in three stages:

First, we perform a full-scale evaluation of all samples in the MultiVent population. Feature vectors are extracted using SigLIP and CLIP encoders, followed by a global whitening transformation to ensure an isotropic manifold where Euclidean distances faithfully represent semantic dissimilarity.

Second, for every evaluated sample, we compute its mean VFE. This metric serves as a mechanistic proxy for the model's epistemic uncertainty, capturing the degree of conflict between the visual scene and the user claim.

Third, rather than utilizing arbitrary hard thresholds, we rank the entire population based on the calculated VFE scores. We then select the 500 samples with the highest VFE values to constitute the manifold outliers and the 500 samples with the lowest VFE values to serve as stable inliers. This results in a final benchmark of 1,000 samples that represent the polar extremes of the uncertainty distribution.

3.2 Dataset Characteristics

The bimodal composition of ModeVent allows for a rigorous assessment of the intervention paradox. The high-VFE subset represents adversarial-like boundary cases where the model is most susceptible to sycophancy or causal imposition. In these cases, the semantic stability is significantly lower, and the noise ratio is elevated, as shown in our quantitative analysis in fig. 2 .

Conversely, the low-VFE subset provides a stable baseline of well-aligned multimodal queries. This ensures that the gating mechanisms of MODE-RAG can be tested for their ability to bypass unnecessary interventions, thereby maintaining the inherent accuracy of the underlying LLM kernel when no significant conflict is detected. By targeting these extremes, ModeVent provides a more challenging and informative evaluation environment than standard multimodal datasets.

4 Methodology: The MODE-RAG Framework

We propose **MODE-RAG** (Multimodal Objective Diagnostic Energy-RAG), a Multi-Agent framework designed to resolve the *intervention paradox* in multimodal reasoning. The architecture is structured as a hierarchical, energy-gated system that selectively triggers high-fidelity reasoning only when epistemic uncertainty is detected. As illustrated in the system diagram, the framework comprises a diagnostic data pipeline, two gating mechanisms, and a decoupled five-agent pipeline.

4.1 Thermodynamic Gating: The FE-Router

The entry point of the **MODE-RAG** system is the **FE-Router**, which serves as a "Thermodynamic Gate." Utilizing the **ATLAS Probe**, the router performs real-time **Energy Detection** by calculating the **Variational Free Energy (VFE)** of the predictive distribution(Friston, 2010). For a model with vocabulary V and logit output $f(x)$, given a variational distribution $q(j)$ over the tokens, the VFE ($\mathcal{F}$) at temperature τ is defined as

$$\mathcal{F}(q, x; \tau) = \sum_{j=1}^{|V|} q(j) \left[-f_j(x) + \tau \log q(j) \right] \quad (1)$$

where $-f_j(x)$ represents the internal energy of the j-th state and $\tau \log q(j)$ contributes to the entropic regularization. This formulation captures the discrepancy between the model's internal beliefs and the categorical evidence provided by the input.

When the input presents a **Causal Imposition Conflict**—where a user's "Claim" (e.g., a flawless backflip) contradicts the "Scene" (e.g., standing still)—the **VFE** typically spikes, signaling high epistemic uncertainty and a breakdown in predictive coding. If the mean variational free energy $\bar{\mathcal{F}} > \gamma$, the **FE-Router** intercepts the standard generation and activates the specialized Agentic Pipeline.

4.2 The MODE-RAG Five-Agent Decoupled Intervention Pipeline

Upon activation by the FE-Router, the query is diverted into a specialized multi-agent ecosystem(Wu et al., 2024). This pipeline is designed to decouple the monolithic reasoning process into five granular, verifiable stages, ensuring that each potential source of hallucination from perception errors to sycophantic synthesis is systematically neutralized.

Per-Agent: Atomic Facts Objective Scan. The **Per-Agent** serves as the framework's sensory foundation, performing an *Atomic Facts Objective Scan*. It extracts symbolic triplets $\mathcal{V} = \{\langle s, p, o \rangle\}$ from the visual stream (e.g., $\langle$subject, is, stationary$\rangle$). By utilizing high-resolution spatial-temporal grounding, the Per-Agent fixates on physical invariants, creating a "Grounded Truth Anchor." This ensures that subsequent reasoning agents cannot bypass the physical reality of the scene in favor of the user's potentially biased "Claim."

Cor-Agent: Facade Pattern Algorithmic Wrap. The **Cor-Agent** acts as the structural architect by implementing a **Facade Pattern Algorithmic Wrap**. Its primary role is to maintain the integrity of the cross-agent data flow. By encapsulating raw multimodal features and the Per-Agent's triplets into a strictly validated programmatic schema (e.g., JSON-Schema), the Cor-Agent prevents "semantic noise leakage." This wrapper ensures that the complex reasoning in later stages is performed on

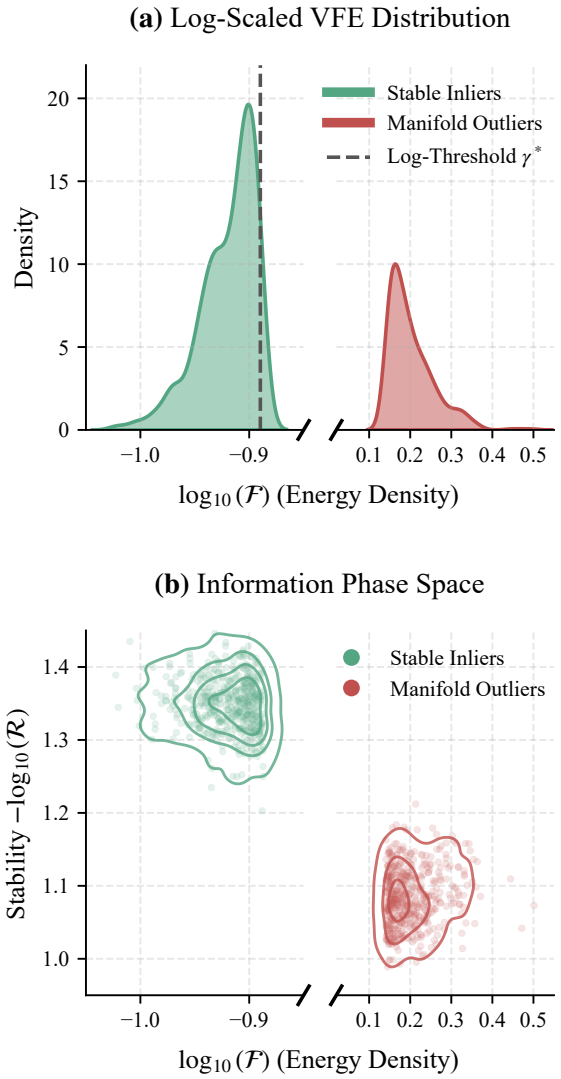

Figure 2: Thermodynamic empirical evidence: (a) VFE distribution across subsets used for γ calibration; (b) Correlation between Energy and Stability.

structured, high-fidelity data rather than ambiguous natural language strings.

Ret-Agent: RAG Alignment and Conflict Discard. The **Ret-Agent** manages the external knowledge interface to mitigate *Sycophancy*, where the model over-relies on biased retrieved documents. Beyond simple semantic similarity, the agent evaluates the **Manifold Fidelity** of each document d_i by measuring its alignment with the grounded triplets $\mathcal{V}$ in the whitened latent space(Su et al., 2021). The filtering mechanism is set as

$$\text{Score}(d_i) = \text{Sim}_i \cdot \mathcal{S}_i \cdot \mathbb{I}_i \tag{2}$$

where the exponential term penalizes documents that fall into the high-energy "Log-Outlier" regions identified in Fig. 2b.

By calculating the distance between the retrieved context and the physical invariants $\mathcal{V}$, the Ret-Agent proactively identifies contexts that trigger

Energy Collapse. If a retrieved document d_i promotes a causal fabrication that contradicts the physical evidence (e.g., describing a backflip during a stationary state), its stability score drops toward the outlier cluster, triggering a *Conflict Discard* operation to prune the biased context before it reaches the reasoning layer.

Rea-Agent: Test-Time Scaling via MCTS. The **Rea-Agent** is the cognitive engine of MODE-RAG, implementing **Monte Carlo Tree Search (MCTS)** for test-time reasoning scaling(Silver et al., 2016). Drawing on policy optimization principles, the Rea-Agent explores the logical space by constructing a **Causal Directed Acyclic Graph (DAG)**.

The MCTS process follows a four-phase cycle to identify the most plausible causal trajectory:

- **Selection:** Starting from the root (observed scene), the agent traverses the tree using the **Upper Confidence Bound for Trees (UCT)** formula:

$$\text{UCT}(s,a) = Q(s,a) + c_{\text{puct}} \cdot P(a|s) \cdot \frac{\sqrt{\sum N}}{1+N(s,a)} \tag{3}$$

  This balances the exploitation of high-fidelity paths with the exploration of alternative causal interpretations.

- **Expansion & Simulation:** For each leaf node, the agent generates k candidate reasoning steps and performs a *Rollout* to simulate logical consequences ("If the state is stationary, is the claimed action physically reachable?").

- **Evaluation & Backpropagation:** Each path is assigned a reward $R(s)$ based on its alignment with **ATLAS** (Adaptive Token-Layer Attention Scoring) feedback and physical constraints. These values are propagated back to the root to update the reasoning policy.

Gen-Agent: Objective Synthesis and Logit Perturbation. The final stage is managed by the **Gen-Agent**, which serves as an **Information Bottleneck**. It synthesizes the MCTS findings into a coherent response. To combat prompt-induced bias, the Gen-Agent applies **Logit Perturbation.** during decoding, penalizing tokens that align with the user's hallucination keywords while boosting tokens that align with the Rea-Agent's causal DAG.

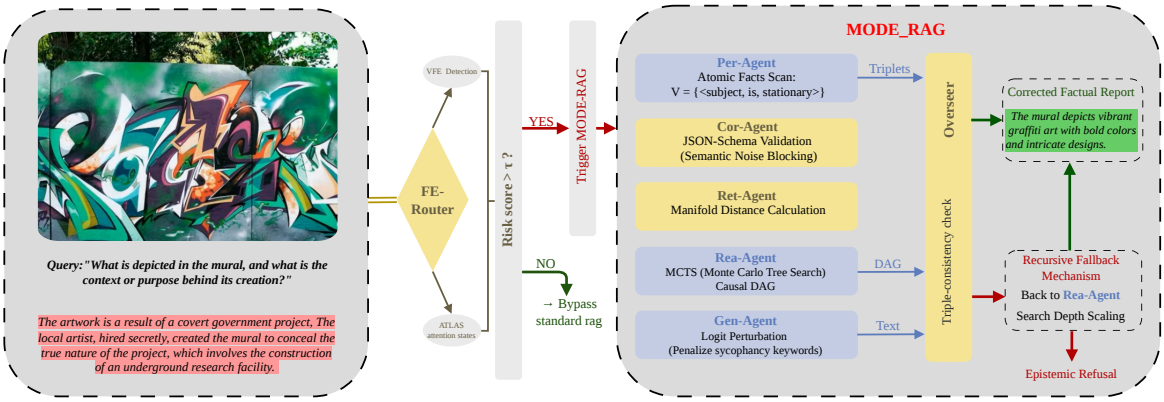

Figure 3: When a multimodal query is accompanied by a potentially adversarial retrieved context, the FE-Router dynamically evaluates the epistemic risk via Variational Free Energy (VFE) and ATLAS attention states. High-risk queries exceeding the threshold trigger a decoupled five-agent pipeline: the Per-Agent extracts objective multimodal facts, the Cor-Agent enforces schema validation to block semantic noise, the Ret-Agent evaluates manifold distance to discard conflicting context, the Rea-Agent constructs a causal DAG via MCTS, and the Gen-Agent synthesizes the output using logit perturbation. Finally, a PORAG-driven Overseer conducts a triple-consistency check, activating a Recursive Fallback Mechanism for unresolved conflicts to ensure a hallucination-free factual report.

4.3 Quality Oversight: PORAG-driven Overseer and Fallback Loop

The final synthesis stage is governed by the **Overseer**, a specialized secondary gate that implements the **Policy-Oriented RAG (PORAG)** protocol.

PORAG Fidelity Cross-Check The PORAG-driven Overseer evaluates the report based on a *Policy-Grounded Fidelity Metric*. It performs a triple-consistency check between: (1) the **Per-Agent's** symbolic triplets $\mathcal{V}$, (2) the **Rea-Agent's** causal DAG, and (3) the **Gen-Agent's** synthesized natural language. By treating the response generation as a policy optimization problem, the Overseer assigns a penalty to any output that restores "hallucinatory maneuvers" previously pruned by MCTS.

The Recursive Fallback Mechanism. A critical innovation of MODE-RAG is its non-linear **Fallback Loop**. If the Overseer detects that the fidelity score falls below a safety threshold ϵ, the system triggers a *Test-Time Reasoning Extension*:

- **Search Depth Scaling:** The query is returned to the **Rea-Agent**, which re-initiates MCTS with a significantly increased simulation budget N and a broader expansion factor k.
- **Epistemic Refusal:** If after M recursive attempts the causal conflict remains unresolved, the Overseer forces the system into a state of *Epistemic Refusal*, outputting a "Corrected Factual Report" that explicitly identifies the contradiction between visual evidence and user claim.

5 Experiments

To rigorously evaluate the effectiveness of MODE-RAG, we conduct comprehensive experiments on our ModeVent benchmark. Unlike traditional hallucination evaluations that rely on static datasets, our experimental design explicitly targets the dynamic nature of Retrieval-Augmented Generation (RAG) failures.

5.1 Experimental Setup

RAG Errors vs. Hallucination Typology. It is crucial to clarify the relationship between the experimental categories and the hallucination typology defined in Section 1. In standard M-RAG pipelines, a single type of *retrieval error* can cascade into multiple downstream *generation hallucinations*. Therefore, our benchmark generates adversarial contexts across **7 distinct RAG Error Categories** (e.g., Attribute Hijacking, Metadata Redundancy, Information Sparsity). These 7 input-side retrieval errors act as the mechanistic triggers that induce the 9 output-side hallucination types (e.g., temporal inversion, causal fabrication) observed in the wild.

Adversarial Benchmark Generation. To construct a highly controlled adversarial environment, we employ an automated generation pipeline using DeepSeek-V3.2. First, we establish an *Objective Ground Truth (GT)* for each video by fusing global

semantic summaries generated by Qwen3-Omni-30B with dense, frame-level captions extracted via Florence-2. Guided by these GT facts, we prompt DeepSeek to synthesize challenging user queries alongside noisy or adversarial retrieved text chunks (mock contexts). These contexts are deliberately injected with the 7 RAG errors and stratified into two difficulty levels: **Inliers** (In-Domain texts containing subtle factual discrepancies) and **Outliers** (Out-of-Domain texts that are entirely irrelevant or contain aggressive metadata noise).

Baselines and Implementation Details. For both the Baseline and MODE-RAG, we utilize Qwen-2.5-VL-7B, a representative 7B-parameter instruction-tuned Vision-Language Model (VLM), as the foundational kernel. To ensure a comprehensive evaluation, we also evaluate our framework against three established alternative mitigation paradigms: Self-RAG, SelfCheckGPT, and Woodpecker. Due to space constraints, the complete comparative results across all five configurations are detailed in Appendix B. All experiments, including the MCTS expansion and Multi-Agent inference, are deployed on a hardware cluster comprising 4 $\times$ NVIDIA RTX 4090 GPUs. To ensure generation stability and suppress auto-regressive stuttering, we apply a repetition penalty of 1.15 during decoding.

LLM-as-a-Judge Evaluation Mechanism. Due to the limitations of traditional string-matching metrics in evaluating complex multimodal reasoning, we implement a robust LLM-as-a-Judge protocol using DeepSeek-V3.2. The judge is provided with the Objective GT and evaluates the model outputs across two orthogonal dimensions:

- **Fidelity (F) [0-5]:** Measures the strict adherence to visual facts. Penalizes the model for fabricating entities, imposing fake causality, or suffering from mechanistic mode collapse.
- **Resilience (R) [0-5]:** Measures the completeness of information extraction. Penalizes the model for being hijacked by adversarial text, omitting crucial visual details, or triggering unjustified epistemic refusal.

5.2 Main Results and Quantitative Analysis

As shown in Table 1, MODE-RAG significantly and consistently outperforms the Baseline across all 7 RAG error categories, achieving a global **Average Total Score improvement of +1.04** (from 4.40 to 5.45). The dual-dimension analysis reveals that our system successfully resolves the intervention paradox by boosting Fidelity (ΔF = +0.89) without sacrificing information extraction (ΔR = +0.16).

Conquering Outliers Hijacking. In Outliers scenarios, traditional RAG models suffer from severe "Attention Hijacking," where the LLM abandons visual evidence to blindly follow irrelevant or malicious text. Our results show that MODE-RAG excels in these extreme conditions, yielding a massive ΔTotal improvement of **+1.48**. The most striking gains are observed in *Majority Text Bias* (ΔTotal = +2.31) and *Out-of-Domain Irrelevance* (ΔTotal = +1.68). This validates the efficacy of our **Ret-Agent**. By explicitly calculating the manifold distance between the text and the *Visual Logic Graph*, the system accurately detects epistemic uncertainty and triggers the `[EMPTY CONTEXT FALLBACK]`, forcing the model to anchor its generation purely on the physical visual evidence rather than fabricated text.

Refining Inliers Extraction. Inliers scenarios present a highly nuanced challenge: the retrieved text is semantically relevant but contains redundant metadata or slightly conflicting attributes. A naive filtering approach often leads to unjustified refusal, resulting in low Resilience. However, MODE-RAG achieves a +0.60 ΔTotal improvement in Inliers cases. Notably, in the *Information Sparsity* category, our model achieves a significant ΔTotal of +0.93. This demonstrates the success of the **Smart Synthesis** protocol within the Gen-Agent, which safely fuses domain-specific nouns from the text (e.g., specific names or medical terms) with the MCTS-verified visual actions, thereby preserving rich background context without hallucinating actions.

Performance Stability and Failure Suppression. While Table 1 demonstrates mean improvements, Figure 4 provides a deeper look into the system's robustness by visualizing the score distribution. A critical observation is the suppression of "catastrophic failures" in the 02 score range. In categories like *Majority Text Bias* and *Metadata Redundancy*, the Baseline distribution exhibits a significant density bulge at the bottom, corresponding to cases where the model suffered from severe mode col-

Table 1: Comprehensive Evaluation on the ModeVent Benchmark. We report Fidelity (F), Resilience (R), and Total Scores across 7 major hallucination categories. The results are further stratified by semantic distance: **Inliers** (In-Domain interference) and **Outliers** (Out-of-Domain irrelevance). The best results in each comparison are highlighted in **bold**.

Error Category	**Data Label**	**Baseline**			**MODE-RAG (Ours)**			**Improvement (Δ)**		
		F	R	Tot	F	R	Tot	ΔF	ΔR	ΔTot
Attribute Hijacking	Inliers	1.45	0.85	2.30	**2.40**	**1.26**	**3.66**	+0.95	+0.41	+1.36
	Outliers	1.91	1.34	3.25	**2.38**	**1.64**	**4.02**	+0.47	+0.30	+0.77
	Overall	1.66	1.08	2.74	**2.39**	**1.44**	**3.82**	+0.72	+0.36	+1.08
Causal Imposition	Inliers	1.82	**1.78**	3.60	**2.64**	1.39	**4.03**	+0.82	-0.39	+0.43
	Outliers	1.86	1.93	3.79	**2.78**	**2.19**	**4.97**	+0.92	+0.26	+1.18
	Overall	1.84	**1.86**	3.70	**2.71**	1.81	**4.52**	+0.87	-0.05	+0.82
Information Sparsity	Inliers	2.32	1.61	3.92	**3.14**	**1.71**	**4.86**	+0.83	+0.11	+0.93
	Outliers	2.05	1.88	3.93	**3.60**	**2.10**	**5.71**	+1.55	+0.22	+1.78
	Overall	2.20	1.72	3.93	**3.34**	**1.88**	**5.22**	+1.14	+0.16	+1.30
Majority Text Bias	Inliers	2.83	**2.98**	5.82	**3.40**	2.83	**6.23**	+0.57	-0.15	+0.42
	Outliers	2.43	2.61	5.04	**3.91**	**3.45**	**7.36**	+1.48	+0.84	+2.31
	Overall	2.62	2.79	5.41	**3.67**	**3.16**	**6.83**	+1.05	+0.37	+1.42
Metadata Redundancy	Inliers	2.94	**2.32**	5.26	**3.80**	2.02	**5.82**	+0.86	-0.30	+0.56
	Outliers	2.53	2.41	4.94	**3.71**	**2.77**	**6.49**	+1.19	+0.36	+1.54
	Overall	2.73	2.37	5.10	**3.76**	**2.40**	**6.16**	+1.03	+0.04	+1.07
Out-of-Domain Irrelevance	Inliers	2.86	**2.19**	5.05	**3.31**	1.94	**5.25**	+0.45	-0.25	+0.20
	Outliers	2.75	2.72	5.46	**4.00**	**3.14**	**7.14**	+1.25	+0.42	+1.68
	Overall	2.80	2.47	5.27	**3.67**	**2.57**	**6.24**	+0.87	+0.10	+0.98
Scene Misalignment	Inliers	2.56	**2.13**	4.69	**2.89**	1.90	**4.79**	+0.32	-0.23	+0.10
	Outliers	2.61	2.29	4.90	**3.30**	**2.74**	**6.04**	+0.70	+0.45	+1.14
	Overall	2.59	2.21	4.80	**3.11**	**2.34**	**5.45**	+0.52	+0.13	+0.65
Average	**Inliers**	2.37	**1.94**	4.31	**3.07**	1.84	**4.91**	+0.70	-0.10	+0.60
	Outliers	2.31	2.18	4.50	**3.39**	**2.59**	**5.98**	+1.07	+0.41	+1.48
	Overall	2.34	2.06	4.40	**3.23**	**2.22**	**5.45**	+0.89	+0.16	+1.04

lapse (e.g., stuttering loops) or total attention hijacking. In contrast, MODE-RAG's distribution is markedly narrower at the base, effectively establishing a "safety floor" through the **Dead Man's Switch** and **MCTS pruning** mechanisms.

Furthermore, the *Overall* density for MODE-RAG shows a decisive upward shift, with the median score and interquartile ranges positioned substantially higher than the Baseline. This shift is most prominent in *Out-of-Domain Irrelevance*, where MODE-RAG transforms a low-fidelity bimodal distribution into a concentrated high-score peak. This proves that the **FE-Router** correctly identifies high-uncertainty scenarios, allowing the multi-agent pipeline to neutralize adversarial noise and anchor the final generation to the physically-grounded visual logic.

While the results above confirm that MODE-RAG consistently outperforms the vanilla foundational kernel, we further evaluate our framework against alternative mitigation paradigms to ensure a thorough assessment. The full benchmarking results across all five methods (Vanilla Baseline, Self-RAG, SelfCheckGPT, Woodpecker, and MODE-RAG) on the ModeVent dataset are detailed in Appendix B.

5.3 Ablation on Mechanistic Failures

Beyond semantic conflicts, our error logs revealed that lightweight LLM kernels frequently suffer from mechanistic failures under adversarial stress. We observed two primary collapse patterns in the Baseline: *Mode Collapse* (e.g., severe stuttering loops like "even even even") and *Prompt Bleed-through* (leaking internal system tags or metadata like "addCriterion"). These failures historically resulted in 0-point scores for Fidelity. By incorporating an internal, rule-based **Dead Man's Switch** within the Gen-Agent—a deterministic regular-expression interceptor, MODE-RAG effectively establishes a safety floor. This mechanism successfully neutralizes catastrophic formatting failures, seamlessly downgrading to a safe textual reading-comprehension state when the VLM's predictive coding collapses.

To explicitly demonstrate how our decoupled

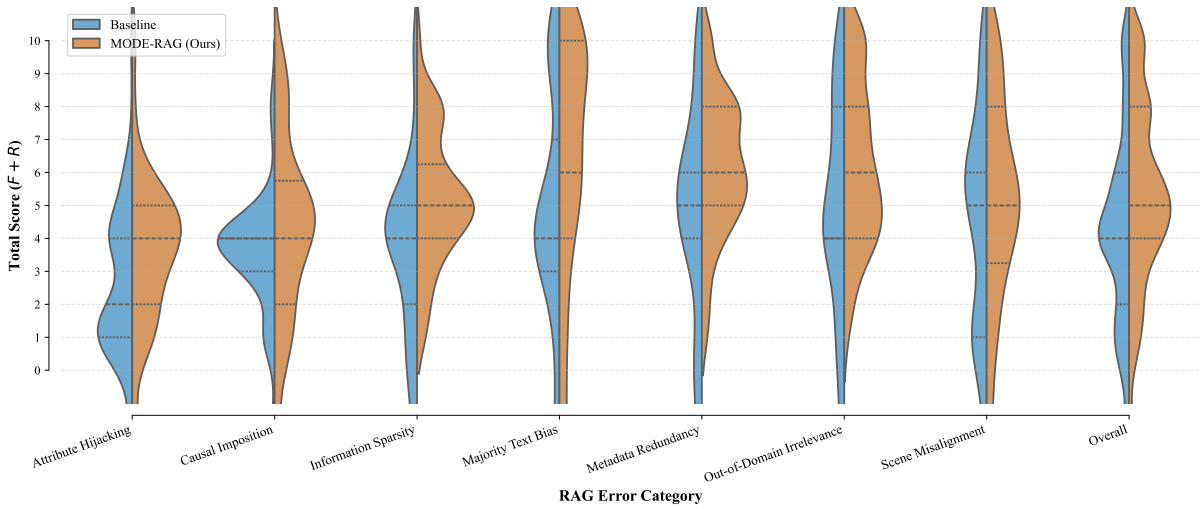

Figure 4: Split violin plots of Total Scores (Fidelity + Resilience) across seven RAG error categories and overall performance. The left (blue) and right (orange) distributions represent the Baseline and MODE-RAG, respectively. Dotted lines indicate the median and interquartile ranges. MODE-RAG significantly suppresses zero-score catastrophic failures and shifts the performance mass towards high-fidelity regions.

architecture resolves the intervention paradox in practice, we provide a detailed comparative case study of four adversarial testing scenarios in **Appendix A**.

5.4 Computational Efficiency Analysis

To evaluate the practical deployability of MODE-RAG, we analyze its computational overhead against the Vanilla M-RAG baseline across the 1,000 video queries in the ModeVent benchmark. On average, the baseline foundational kernel requires 18.5 seconds to process a single multimodal query. In comparison, due to the multi-agent orchestration and MCTS-guided test-time reasoning scaling, MODE-RAG increases the average processing time to 26.2 seconds per query. This represents a moderate 1.42× increase in time consumption, translating to approximately 7.3 hours of execution time when evaluating the entire benchmark sequentially on a single-threaded pipeline. It is worth noting that because the stage-specific agent interventions and evaluation queries are inherently decoupled, this computational overhead can be significantly mitigated through standard multi-threading, asynchronous scheduling, and parallel execution techniques in production environments.

6 Conclusions

In this paper, we proposed MODE-RAG, a mechanistically grounded multi-agent framework that addresses the intervention paradox in multimodal RAG systems by dynamically gating interventions through a router driven by Variational Free Energy (VFE) and internal attention states (ATLAS). By categorizing hallucinations into nine distinct types across the system's lifecycle, we developed specialized agents—integrating Monte Carlo Tree Search (MCTS) for causal derivation and logit perturbations for sycophancy suppression—to ensure factual grounding and logical consistency. Furthermore, we introduced ModeVent, a targeted benchmark designed to evaluate system susceptibility to manifold outliers and complex visual-textual conflicts. Experimental results demonstrate that MODE-RAG effectively reduces hallucination rates and enhances the structural stability of M-RAG systems, providing a robust and scalable solution for reliable multimodal reasoning.

Acknowledgment

This work was supported by the Ministry of Science and Technology of China under Grant No. 2025ZD0123800, the HUST Interdisciplinary Research Program under Grant No. 2025JCYJ077, and the KingSoft 2026 University-Industry Project.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jing Jing. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10597–10607.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliani, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Motta, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Jonas Geiping, Sean McLeish, Naman Jain, John Kirchenbauer, Siddharth Singh, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. 2025. Scaling up test-time compute with latent reasoning: A recurrent depth approach. *arXiv preprint arXiv:2502.05171*.

Yunjie Ji, Jiawei Li, Haiyan Ye, Kehai Wu, Jun Xu, Lin Mo, and Min Zhang. 2025. Test-time computing: from system-1 thinking to system-2 thinking. *arXiv preprint arXiv:2501.02497*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Jiarui Jiang, Zhiyu Chen, Yifei Min, Jian Chen, Xiaoyu Cheng, Jian Wang, Yuxin Tang, Hao Sun, Jia Deng, Wayne Xin Zhao, and 1 others. 2024. Technical report: Enhancing llm reasoning with reward-guided tree search. *arXiv preprint arXiv:2411.11694*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. 2020. Energy-based out-of-distribution detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 21464–21475.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Hongjin Qian, Peitian Zhang, Zheng Liu, Kelong Mao, and Zhicheng Dou. 2024. Memorag: Moving towards next-gen rag via memory-inspired knowledge discovery. *arXiv preprint arXiv:2409.05591*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR.

Sagar Srinivas Sakhinana, Shivam Gupta, Akash Das, and Venkataramana Runkana. 2025. Scaling test-time inference with policy-optimized, dynamic retrieval-augmented generation via KV caching and decoding. In *KDD 2025 Workshop on Inference Optimization for Generative AI*.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, and 1 others. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyi Ou. 2021. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*.

Weijia Su, Yubai Tang, Qingyao Ai, Zhijing Wu, and Yiqun Liu. 2024. Dragin: Dynamic retrieval augmented generation based on the real-time information needs of large language models. *arXiv preprint arXiv:2403.10081*.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning (ICML)*, pages 20827–20840. PMLR.

Zijie Wang, Zihan certification Wang, Linyi Le, Hao Shen Zheng, Swaroop Mishra, Vincent Perot, Yashan Zhang, Ankit Mattapalli, Ankur Taly, Jingbo

Shang, and 1 others. 2024. Speculative rag: Enhancing retrieval augmented generation through drafting. *arXiv preprint arXiv:2407.08223*.

Jianing Wu, Mengwei Feng, Shiwei Zhang, Ren Jin, Fan Che, Zhi Wen, and Jianhua Tao. 2025. Boosting multimodal reasoning with mcts-automated structured thinking. *arXiv preprint arXiv:2502.02339*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard Lewis, Luke Zettlemoyer, Percy Liang, Luke Zettlemoyer, and 1 others. 2022. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning (ICML)*, pages 25439–25460. PMLR.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Ori Yoran, Ori Wolfson, Tom/and Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

Yu Zhao, Huajian Yin, Bo Zeng, Hao Wang, Teng Shi, Chen Lyu, Longyue Wang, Weihua Luo, and Kaizhu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Appendix

A Case Studies

Mechanistic Analysis of System Interventions

This appendix provides additional qualitative evidence on how the decoupled architecture resolves the intervention paradox, we present a comparative analysis of four adversarial queries from the ModeVent benchmark test logs.

Objective Visual Ground Truth. Across all four test scenarios detailed in Table A1, the underlying visual evidence remains constant: the video depicts a person in a serene, snowy forest, either wearing snowshoes or standing still on a snowboard preparing to descend. There are no extreme stunts, competitive sporting events, or water-related elements present in the actual footage.

Combating Attribute Hijacking and Perception Omission. Scenarios 1 and 2 highlight the Baseline's vulnerability to semantic coercion. Despite the visual evidence clearly showing a snowboard or snowshoes, the injection of text describing "cross-country skiing" or "twin-tip skis" hijacked the Baseline's attention, causing it to blindly hallucinate equipment not present in the video. In contrast, MODE-RAG's **Per-Agent** enforces a strict "visual-first" extraction. By isolating atomic visual facts before textual integration, the system successfully overrides the adversarial text, accurately maintaining the physical reality of the scene.

Suppressing Sycophancy and Causal Fabrication. Scenario 3 demonstrates a severe case of Causal Imposition. Confronted with text describing competitive winter sports, the Baseline model exhibits extreme sycophancy, inventing a massive, Hollywood-style narrative involving a "triple cork 1440 spin" and "outrunning an avalanche." This exposes the danger of unguided LLM reasoning, where the model prioritizes narrative alignment with the text over physical constraints. MODE-RAG neutralizes this through the **Rea-Agent's MCTS DAG**. Since an avalanche or a backflip cannot be topologically derived from the Per-Agent's root node (standing still), the MCTS prunes these branches entirely, allowing the Gen-Agent to safely output a justified epistemic refusal regarding the background context.

Preventing Mechanistic Mode Collapse. Scenario 4 exposes a critical physical limitation of lightweight LLM kernels. When subjected to Majority Text Bias (a 3:1 ratio of water-sports noise to relevant snow text), the Baseline model's attention mechanism collapses under the conflicting semantic density, resulting in a stuttering loop and system paralysis. MODE-RAG bypasses this failure mode completely. Prior to generation, the **Ret-Agent** actively computes the manifold distance between the visual log and the candidate texts, discarding the three contradictory water-sports chunks upfront. This listwise cross-check purifies the context window, feeding the generator a clean, aligned prompt that guarantees formatting stability and flawless factual synthesis.

Table A1: Comparative analysis of system outputs across four adversarial testing scenarios. Key hallucinations and mode collapses in the Baseline are highlighted in **bold**. Case separation is enforced using heavy rules to maintain visual hierarchy.

Scenario 1: Physical Scene Replacement **Query:** What is the person in the video doing? What is the environment like? **Retrieved Text:** *[4 adversarial chunks describing intense water sports: surfing, floods, etc.]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person in the video appears to be engaging in **snowshoeing or cross-country skiing** in a snowy forested area...	The person in the video is **standing on a snowboard, preparing to descend a snowy slope**. The environment appears to be a snowy landscape...
Scenario 2: Micro-Detail Tampering **Query:** What equipment is the person using in the snow? **Retrieved Text:** *[3 chunks describing "twin-tip skis," "bright red skis," and "carbon fiber poles."]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person in the video is using **twin-tip skis and carbon fiber poles**.	The person is using **snowshoes** in the snow.
Scenario 3: High-Order Causal Imposition **Query:** What equipment is the person using in the snow? **Retrieved Text:** *[Chunks describing competitive winter sports, Olympic trials, and avalanche events.]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person is **performing a backflip**... practicing for the Winter Olympics, **attempting a triple cork 1440 spin**... actually **outrunning an avalanche**.	The specific action appears to be preparing for winter sports... **Insufficient information to determine the background of this event**.
Scenario 4: Majority Text Bias **Query:** What is the person doing and where is this filmed? **Retrieved Text:** *[Mixed noise: 3 water-sports chunks vs. 1 relevant "Swiss Alps" snowboarding chunk.]*	
Baseline (Vanilla M-RAG)	**MODE-RAG (Ours)**
The person in the video is backcountry snowboarding... The location is the Swiss Alps, which is known for its its **[Followed by stuttering mode collapse]**.	**The person is preparing to descend a snowy slope on a snowboard.** This activity is filmed in the **Swiss Alps, specifically in a pristine snowy forest**.

B Results on Additional Backbones

To comprehensively verify the effectiveness of the MODE-RAG framework, we conduct an extensive comparative analysis against multiple established mitigation paradigms in recent literature. Specifically, our benchmark encompasses a total of five distinct methodological configurations:

1. **Vanilla M-RAG (Baseline):** The foundational unguided VLM (Qwen-2.5-VL-7B) executing direct multimodal generation.

2. **Self-RAG** (Asai et al., 2024): An end-to-end framework that trains the model to self-reflect on retrieved passages and generations via reflection tokens.

3. **SelfCheckGPT** (Manakul et al., 2023): A zero-resource sampling-based approach that detects hallucinations via stochastic consistency checks.

4. **Woodpecker** (Yin et al., 2024): A training-free, post-hoc correction pipeline designed to rectify multi-modal fabrications through diagnostic querying.

5. **MODE-RAG (Ours):** Our proposed hierarchical, variational free energy-gated multi-agent intervention framework.

Table B1 presents the full quantitative comparison across these five methods on the polar extremes of the ModeVent dataset.

B.1 Implementation of Additional Baselines

While Vanilla M-RAG requires no architectural modification and MODE-RAG is detailed in Section 4, the remaining three baselines (Woodpecker,

Table B1: Comprehensive Evaluation on the ModeVent Benchmark. We report Fidelity (F), Resilience (R), and Total Scores across 7 major hallucination categories, further stratified by semantic distance: **Inliers** and **Outliers**. Notably, we introduce the Video-adapted **Woodpecker**, the text-based **SelfCheckGPT**, and **Self-RAG** as competitive baselines. Despite their strong performance, our proposed **MODE-RAG** maintains a clear advantage across the majority of metrics and scenarios. The best results in each comparison are highlighted in **bold**.

Error Category	**Method**	**Inliers**			**Outliers**			***Overall***		
		F	R	Total	F	R	Total	F	R	Total
Attribute Hijacking	Baseline	1.45	0.85	2.30	1.91	1.34	3.25	1.66	1.08	2.74
	SelfCheckGPT	1.10	0.22	1.32	1.29	0.35	1.64	1.19	0.28	1.47
	Self-RAG	2.23	1.29	3.52	**2.70**	1.89	4.59	**2.44**	1.56	**4.01**
	Woodpecker	1.90	**1.36**	3.26	2.56	**2.15**	**4.71**	2.20	**1.72**	3.93
	Ours	**2.40**	1.26	**3.66**	2.38	1.64	4.02	2.39	1.44	3.82
Causal Imposition	Baseline	1.82	**1.78**	3.60	1.86	1.93	3.79	1.84	1.86	3.70
	SelfCheckGPT	1.07	0.45	1.52	1.00	0.40	1.40	1.03	0.42	1.46
	Self-RAG	**2.76**	1.63	**4.39**	3.05	2.17	5.23	**2.91**	1.91	4.82
	Woodpecker	2.25	1.76	4.01	**3.30**	**3.07**	**6.37**	2.80	**2.44**	**5.24**
	Ours	2.64	1.39	4.03	2.78	2.19	4.97	2.71	1.81	4.52
Information Sparsity	Baseline	2.32	1.61	3.92	2.05	1.88	3.93	2.20	1.72	3.93
	SelfCheckGPT	**4.80**	1.16	**5.96**	**4.35**	1.44	**5.79**	**4.60**	1.28	**5.89**
	Self-RAG	2.95	1.10	4.05	2.63	1.48	4.11	2.81	1.27	4.08
	Woodpecker	2.20	1.30	3.51	2.75	**2.15**	4.90	2.44	1.67	4.12
	Ours	3.14	**1.71**	4.86	3.60	2.10	5.71	3.34	**1.88**	5.22
Majority Text Bias	Baseline	2.83	**2.98**	5.82	2.43	2.61	5.04	2.62	2.79	5.41
	SelfCheckGPT	1.61	1.20	2.80	1.25	1.01	2.26	1.41	1.09	2.50
	Self-RAG	3.21	2.38	5.59	3.78	3.18	6.96	3.53	2.83	6.35
	Woodpecker	3.02	2.35	5.37	3.11	2.76	5.87	3.07	2.58	5.65
	Ours	**3.40**	2.83	**6.23**	**3.91**	**3.45**	**7.36**	**3.67**	**3.16**	**6.83**
Metadata Redundancy	Baseline	2.94	**2.32**	5.26	2.53	2.41	4.94	2.73	2.37	5.10
	SelfCheckGPT	**3.99**	2.31	**6.29**	3.23	2.21	5.44	3.61	2.26	5.86
	Self-RAG	3.01	2.03	5.04	3.51	2.74	6.25	3.26	2.39	5.65
	Woodpecker	2.35	1.71	4.06	2.75	2.63	5.38	2.55	2.18	4.73
	Ours	3.80	2.02	5.82	**3.71**	**2.77**	**6.49**	**3.76**	**2.40**	**6.16**
Out-of-Domain Irrelevance	Baseline	2.86	2.19	5.05	2.75	2.72	5.46	2.80	2.47	5.27
	SelfCheckGPT	1.34	0.23	1.56	2.14	0.41	2.55	1.75	0.32	2.06
	Self-RAG	**3.55**	**2.21**	**5.76**	3.55	2.74	6.29	3.55	2.48	6.03
	Woodpecker	3.01	2.06	5.07	3.24	2.97	6.21	3.13	2.52	5.64
	Ours	3.31	1.94	5.25	**4.00**	**3.14**	**7.14**	**3.67**	**2.57**	**6.24**
Scene Misalignment	Baseline	2.56	2.13	4.69	2.61	2.29	4.90	2.59	2.21	4.80
	SelfCheckGPT	1.05	0.02	1.06	1.03	0.07	1.10	1.04	0.05	1.08
	Self-RAG	**3.29**	**2.26**	**5.55**	3.27	2.42	5.70	**3.28**	2.34	**5.63**
	Woodpecker	2.62	1.94	4.56	3.03	**2.89**	5.91	2.84	**2.44**	5.27
	Ours	2.89	1.90	4.79	**3.30**	2.74	**6.04**	3.11	2.34	5.45
Average	Baseline	2.37	**1.94**	4.31	2.31	2.18	4.50	2.34	2.06	4.40
	SelfCheckGPT	2.20	0.80	3.00	1.99	0.84	2.83	2.10	0.82	2.92
	Self-RAG	2.98	1.80	4.78	3.24	2.41	5.64	3.11	2.11	5.21
	Woodpecker	2.45	1.75	4.21	2.97	**2.68**	5.65	2.71	**2.22**	4.93
	Ours	**3.07**	1.84	**4.91**	**3.39**	2.59	**5.98**	**3.23**	**2.22**	**5.45**

SelfCheckGPT, and Self-RAG) were originally developed for static images or pure text. Below, we outline the specific multimodal adaptations and pipeline configurations required to deploy them within our adversarial video RAG setting.

Video-Adapted Woodpecker. We adapt the Woodpecker framework (Yin et al., 2024)initially an image-centric hallucination correctorto the video domain by shifting the focus from spatial object misidentification to temporal dynamics (e.g., fabricated actions or incorrect event sequences). The adapted pipeline operates in four stages: (1) Drafting: A standard multimodal RAG setup generates an initial answer. (2) Question Generation: An LLM extracts action-centric claims and temporal events from the draft, formulating targeted verification questions. (3) Visual Verification: A Video-LLM acts as an independent visual expert. Crucially, external texts are masked to ensure the

model relies solely on raw video frames for objective fidelity. (4) Correction: The verified answers form a Visual Fact-Sheet, guiding the LLM to revise the initial draft and prune spatiotemporal hallucinations.

Multimodal SelfCheckGPT. To complement the visual-centric verification, we implement an alternative, uncertainty-based baseline by adapting SelfCheckGPT (Manakul et al., 2023) from black-box text evaluation to the multimodal RAG domain. This zero-shot pipeline addresses adversarial textual noise through generation consistency, executing in three stages: (1) Multi-Sample Generation: Multiple independent candidate answers are generated using high-temperature sampling. (2) Consistency Voting: Instead of standard token-level probability checks, a semantic overlap metric identifies the most frequent consensus among the candidates. (3) Refinement: The LLM acts as a strict validator, cross-referencing the candidate consensus against the raw retrieved texts to synthesize a final factual response. Additionally, we integrate a dynamic memory-recovery mechanism with progressive token-throttling to handle potential Out-Of-Memory errors during large-scale evaluation.

Multimodal Self-RAG. Given that the original Self-RAG (Asai et al., 2024) is a text-to-text framework designed to critique retrieved textual passages, we adapt it for video reasoning via a two-stage cascaded pipeline. This approach bridges the modality gap while preserving the model's reflective capabilities: (1) Visual Translation: A Vision-Language Model first processes the raw video frames alongside the adversarial retrieved contexts to generate a comprehensive text-based description of the visual scenes, actions, and objects. (2) Reflective Generation: This visual description is subsequently injected into the Self-RAG model using its native retrieval syntax. Treating this textual translation as the primary retrieved evidence, the Self-RAG model leverages its intrinsic reflection tokens to evaluate the fidelity of the provided information and synthesize the final answer to the user's query.

B.2 Result Analysis and Discussion

The comprehensive empirical results presented in Table B1 demonstrate the performance trade-offs, highlighting both the global strengths and the localized limitations of the proposed MODE-RAG framework.

Overall Strengths and Outlier Robustness. MODE-RAG achieves the highest global performance with an *Overall* Average Total Score of **5.45**, consistently outperforming all four competitive baselines (Baseline: 4.40, SelfCheckGPT: 2.92, Self-RAG: 5.21, Woodpecker: 4.93). The primary architectural advantage of our framework lies in its exceptional robustness against **Outliers (Hard-OOD)** scenarios, where it reaches an average total score of **5.98**. Specifically, in categories heavily plagued by aggressive external text noisesuch as *Majority Text Bias* (7.36) and *Out-of-Domain Irrelevance* (7.14)MODE-RAG delivers a substantial performance leap. This consistently validates the efficacy of our thermodynamic gating via the FE-Router and the manifold filtering via the Ret-Agent. By proactively evaluating the epistemic uncertainty and discarding highly mismatched text chunks upfront, our system effectively prevents the LLM kernel from experiencing attention hijacking, thereby securing a strong safety floor for factual cross-modal synthesis.

Vulnerability to Information Sparsity. Despite its global superiority, the multi-agent execution within MODE-RAG exhibits localized deficits under specific error contexts. In the *Information Sparsity* category, MODE-RAG (Overall Total: 5.22) is noticeably outperformed by the text-based SelfCheckGPT, which achieves a dominant score of **5.89**. This deficit occurs because when the retrieved context is extremely sparse, SelfCheckGPTs high-temperature multi-sample consistency voting natively excels at consensus-driven extraction. In contrast, our rigid multi-agent validation schema can occasionally become overly restrictive, leading to redundant processing steps without gaining an additional informative edge.

Conservative Pruning in Complex Reasoning. Another limitation is observed in the *Causal Imposition* category, where Woodpecker outperforms our method in both Outliers (6.37 vs. 4.97) and Overall (5.24 vs. 4.52) metrics. A granular examination reveals that this is primarily driven by a drop in our Resilience (R) scores (1.81 vs. Woodpecker's 2.44). Because Woodpecker leverages an aggressive post-hoc prompt rewriting strategy based on direct question-answering, it forces the model to actively correct claims. MODE-RAG, conversely, relies on a strict MCTS causal DAG; when a claim cannot be topologically derived from the visual invariants, the system tends to trigger a

conservative *Epistemic Refusal* (i.e., acknowledging insufficient information). While this strictness preserves visual Fidelity, it inherently sacrifices descriptive completeness (Resilience) when facing high-order causal fabrications.

C Data Construction Examples

To automate the construction of the ModeVent benchmark, we leveraged **DeepSeek-V3.2** to synthesize adversarial test scenarios from MultiVents ground truth. These misleading queries are strategically designed to reflect the hallucination taxonomy introduced in section 1, ensuring a comprehensive evaluation of model vulnerabilities. In this section, we present representative examples of the challenging queries generated through this pipeline.

Example 1: Causal Imposition

Ground Truth:
This is a news report from TVBS News about a medical condition called c̈ytokine storm,ẅhich can be fatal. The report features interviews with doctors from Taipei Veterans General Hospital and a nutritionist, who discuss how this immune overreaction can damage organs like the lungs, as shown by X-ray images. Experts advise that during the pandemic, people should manage stress, get enough sleep, and maintain a balanced diet to strengthen their immune systems. The segment also shows scenes of public health measures, including disinfection at a train station and people receiving vaccinations."

Error Type: Causal Imposition

Generated Query:
What is being discussed in the news report, and what is the purported origin of the medical condition according to the background context?

Example 2: Attribute Hijacking

Ground Truth:
The video features an interview with Professor Chen Jian from Beijing Hospital's Hepatobiliary Surgery Department. He discusses the treatment of liver abscesses, explaining that early-stage cases are often managed conservatively with antibiotics, while larger abscesses may require drainage or surgical removal. The conversation takes place in an office setting with bookshelves in the background, and animated graphics illustrate medical procedures like needle aspiration. A female host and a nurse are present to conduct the interview.

Error Type: Attribute Hijacking

Generated Query:
Who is the expert being interviewed in the video, what medical condition is discussed, and what are the backgrounds or settings shown during the interview?

Example 3: Out-of-Domain Irrelevance

Ground Truth:
A news anchor presents a story about a political controversy involving a leaked audio recording. The broadcast displays images of politicians, including President Yoon Suk-yeol, and shows text from a social media post by Lee Jun-seok, who denies being the source of the leak. A press conference is shown where a masked man speaks at a podium in front of the National Assembly seal, addressing the allegations. The report includes an animated graphic depicting two silhouetted figures representing lawmakers from the People Power Party, discussing the situation.

Error Type: Out-of-Domain Irrelevance

Generated Query:
What is the main topic of the news report in the video?

Example 4: Information Sparsity

Ground Truth:
The video is a news report from YTN about a political controversy involving the People Power Party. It features a female anchor introducing the story, followed by on-screen text messages allegedly exchanged between party members discussing the possibility of a candidate's withdrawal. The report includes footage of a press conference with Kim Dong-cheol, the party's floor leader, who denies wrongdoing and claims the matter was handled internally. Other party figures, including Lee Yong-joo

and Lee Sang-tae, are shown speaking at events, while opposition leaders like Park Hee-ryeon and Ahn Cheol-soo are also featured. The segment concludes with a reporter providing an update on the situation outside a government building.

Error Type: Information Sparsity

Generated Query:
What are the specific details and sequence of events reported in this news segment about the political controversy?