

Object Presence Is Not Relation Evidence: Relation-Grounded Faithful Reasoning for Multimodal Large Language Models

Xiyao Dong* Yilong Chen* Xiaojin Zhang Kun He†

Huazhong University of Science and Technology

{dongxiyao, ylchen, xiaojinzhang, brooklet60}@hust.edu.cn

Abstract

Recent Multimodal Large Language Models (MLLMs) excel in precise entity recognition and powerful reasoning, significantly boosting performance and efficiency in complex vision-language tasks. Despite these gains, this advanced reasoning process introduces a subtle yet critical vulnerability. We identify an under-explored relational reasoning failure mode in which models over-prioritize semantic coherence, thereby inferring ungrounded interactions from mere object entities. We term this phenomenon Relation Evidence Amnesia (REA), wherein MLLMs correctly localize individual entities but fail to enforce true visual dependencies as the generative process becomes hijacked by semantic rationalization. To mitigate REA, we propose Relation-Grounded Faithful Reasoning (RGFR), an inference-time decoding intervention that operates without modifying target model parameters. Adopting a decoupled verification philosophy, RGFR employs a semantic-driven soft-gating mechanism to simultaneously incorporate explicit entity constraints and extract counterfactual causal evidence, followed by a Unified Schema-Aware Rectifier that rigorously mediates these visual-causal constraints during the decoding process. Extensive experiments across relation benchmarks demonstrate that RGFR improves relation recognition accuracy by over 8.1% compared to baselines, while largely preserving task utility. These results highlight the critical role of decoupled, evidence-grounded rectification in achieving faithful relational reasoning for multimodal models.

1 Introduction

Built upon the success of Large Language Models (LLMs), Multimodal Large Language Models (MLLMs) integrate visual perception with natural

language reasoning, significantly enhancing performance on complex scene understanding tasks (Liu et al., 2023; OpenAI et al., 2024; Yang et al., 2025a; Ma et al., 2026). Despite these advancements, recent research reveals that their generative capabilities are severely bottlenecked by hallucinations (Guan et al., 2024; Huang et al., 2024a; Wang et al., 2025b). Specifically, MLLMs exhibit an inherent tendency to maintain linguistic inertia and semantic coherence, which induces the unfaithful rationalization of physical interactions, generating text that is not faithfully grounded in the visual input (Zhou et al., 2024; Wang et al., 2024a; Yang et al., 2025b; Samyoun et al., 2026).

Existing mitigation strategies primarily rely on training-time alignment or inference-time interventions. Training-time approaches struggle to eradicate deep-seated semantic biases without massive data scale. Meanwhile, inference-time defenses (Yin et al., 2024; Leng et al., 2024; Wu et al., 2024a; Lee and Song, 2025) attempt to intercept ungrounded generation via contrastive decoding or output calibration. However, they rely on indiscriminate global perturbations. While global noise helps verify object existence, it encounters fundamental limitations in relation verification, as uniform noise inevitably obscures the fine-grained spatial features required to identify physical interactions. Motivated by these limitations, we pose two critical research questions: (1) how internal reasoning trajectories evolve to rationalize ungrounded relations despite accurate object perception, and (2) whether this rationalization can be dynamically intercepted during decoding to restore faithful visual grounding.

In this work, we take a first step toward answering these questions by systematically analyzing the internal decoding dynamics of MLLMs. Through this analysis, we identify a fundamental cognitive pathology we term Relation Evidence Amnesia (REA). We observe that MLLMs fail not from a

*The first two authors contribute equally.

†Corresponding author.

lack of basic visual perception, but because semantic rationalization overrides factual observation. Once a model commits to a likely entity co-occurrence, it mechanically diverts attention away from genuine visual evidence, cognitively suppressing fine-grained spatial boundaries and physical contact signals. As detailed in Section 3, REA manifests in three orthogonal failure modes: Coordinate Blindness, Co-presence Bias, and Textual Prior Overconfidence.

To address this, we propose Relation-Grounded Faithful Reasoning (RGFR), an inference-time framework that dynamically intercepts this rationalization process. RGFR uses a semantic-driven soft-gating mechanism to activate two parallel modules: a Geometric Verifier for explicit entity constraints and a Causal Probing Branch that isolates genuine visual evidence from spurious co-presence. Finally, a Unified Schema-Aware Rectifier turns relation verification into a strict procedural constraint. By using visual-causal evidence to correct relation predictions biased by language priors, RGFR prevents linguistic inertia from overriding visual reality.

Our main contributions are summarized as follows:

- We identify and mechanistically analyze Relation Evidence Amnesia (REA), a vulnerability where linguistic inertia drives MLLMs to suppress critical visual evidence, manifesting as coordinate blindness, co-presence bias, and textual prior overconfidence.
- We introduce RGFR, an inference-time framework that fundamentally mitigates this pathology through dynamic interventions at the decoding stage, employing an explicitly decoupled entity-causal verification and a semantic-driven rectification mechanism.
- Extensive experiments on relation benchmarks demonstrate that RGFR improves relation recognition accuracy by over 8.1% compared to state-of-the-art baselines, while preserving the model’s general utility.

2 Related Work

2.1 Training-free Multimodal Hallucination Mitigation.

Existing training-free methods primarily mitigate hallucinations during inference, either via advanced decoding strategies (Leng et al., 2024;

Wang et al., 2024b; Park et al., 2024; Huo et al., 2025) or internal representation diagnostics (Liu et al., 2024a; Chen et al., 2024; Gong et al., 2024; Sarkar et al., 2025). Although effective, these traditional paradigms are predominantly tailored for object-centric hallucinations. To bridge this gap, targeted mitigation techniques focusing on relation hallucination have rapidly emerged. For instance, ChainMPQ (Wu et al., 2026) tackles this failure mode by incorporating multi-perspective sub-questions and an interleaved text-image reasoning chain; Reefknot (Zheng et al., 2025) counters relationship misjudgments through specialized structural alignment and benchmarking; and CAP (Wu et al., 2025) mitigates relational errors via constraint-aware prompting with explicit spatial-logical rules. This transition highlights a growing consensus that mitigating multimodal hallucination requires moving beyond entity-level presence to mastering structural and relational consistency (Zhang et al., 2026b).

2.2 Uncertainty-aware reasoning correction.

Quantifying and calibrating model confidence has emerged as a rigorous paradigm to diagnose and rectify structural hallucinations (Zhang et al., 2025; Li et al., 2025; Zhang et al., 2026a). Prior works rely on various expressions of model uncertainty to detect low-confidence states, including token-level entropy (Kadavath et al., 2022), semantic consistency over generated distributions (Kuhn et al., 2023), and verbalized probabilities through self-introspection (Huang et al., 2025). In our work, we operate on binary relation verification, utilizing entropy solely to rescale the penalty of a counterfactual branch. The closest to our approach is the contemporaneous work of (Xu et al., 2026), which uses entropy-aware switching between latent and discrete reasoning modes and injects visual anchors when reasoning becomes uncertain.

2.3 Relation-aware representation beyond hallucination mitigation.

The importance of relations has been recognized in other multimodal tasks (Hu et al., 2023; Huang et al., 2024b). For instance, relation-aware approaches are deployed to erase unsafe knowledge in safety unlearning (Anilkumar et al., 2026), encode joint entity-relation embeddings in multi-modal knowledge graph completion (Niu and Li, 2026), and capture structured spatial-semantic contexts in fine-grained 3D visual grounding (He et al., 2021).

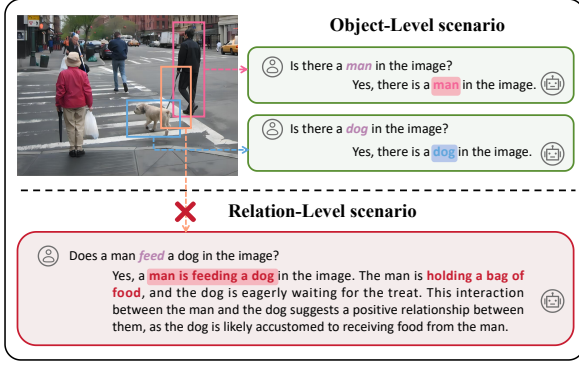


Figure 1: Comparison of model behaviors under Object-Level scenario and Relation-Level scenario.

These methods reinforce a broader point that relations are first-class semantic objects in multimodal systems and should not be reduced to isolated concepts. Our work brings the same relation-aware viewpoint to hallucination mitigation, while remaining entirely training-free and inference-time.

3 Motivation

3.1 Reasoning Vulnerability in Relation Grounding

Prior work largely frames MLLM hallucinations as perceptual failures, primarily focusing on the inability to ground absent objects (Yin et al., 2024). In the relational domain, however, we identify a fundamentally different mechanism: relation hallucinations are driven not by perceptual blindness, but by *overconfident reasoning*. Driven by user instructions and strong language priors, the model actively rationalizes unsupported relations between visually grounded entities.

Specifically, the reasoning process treats its accurate object-level perception as a misleading premise, misinterpreting the mere co-occurrence of these entities as sufficient interaction evidence to maintain narrative fluency. Consequently, even increasingly effective object-level hallucination mitigation remains insufficient to suppress relation hallucinations.

As illustrated in Figure 1, a clear discrepancy emerges between entity perception and relation faithfulness. In Object-Level scenarios, the model reliably anchors to relevant entities, such as the man and the dog, confirming that errors do not stem from perceptual failure. In Relation-Level scenarios, however, the model misinterprets mere co-presence as definitive interaction evidence, incorrectly asserting that the man is feeding the dog.

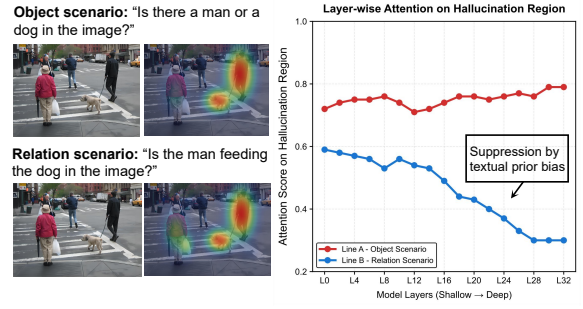


Figure 2: (Left) Attention shift: Entity co-presence triggers linguistic inertia, redirecting attention away from visual interaction evidence. (Right) Layer-wise scores: Deep layers suppress relation-specific visual cues, unlike shallow layers that retain them, indicating prior-driven override.

It further fabricates a fluent rational, such as claiming the man is holding food, to render the unsupported interaction narratively plausible. This contrast points to a cognitive suppression effect rather than perceptual blindness: once entities are grounded, prior-driven reasoning overrides the finer visual constraints necessary for relational verification. This motivates our mechanistic analysis of Relation Evidence Amnesia in Section 3.2.

3.2 Mechanistic Validation of Relation Evidence Amnesia

Building on Section 3.1, we investigate the mechanistic roots of this suppression and term it *Relation Evidence Amnesia* (REA). REA is not a perceptual failure but a cognitive override: when processing relational queries, the model systematically discards valid visual interaction cues in favor of language priors. More broadly, REA manifests in three potentially overlapping modes: *Coordinate Blindness*, the failure to enforce spatial constraints despite correct entity localization; *Co-presence Bias*, the tendency to treat mere entity co-occurrence as interaction evidence, as in Figure 1, where the model incorrectly concludes that the man is feeding the dog; and *Textual Prior Overconfidence*, the preference for linguistically plausible relations despite weak or contradictory visual evidence, reflected in the fabricated claim that the man is holding food.

To validate this, we perform a layer-wise cross-modal attention analysis on relation verification benchmarks (Wu et al., 2024b). As illustrated in Figure 2, our analysis reveals a critical Perception-Cognition Gap. In shallow perception layers, the visual encoder objectively attends to interaction boundaries; however, as reasoning progresses to

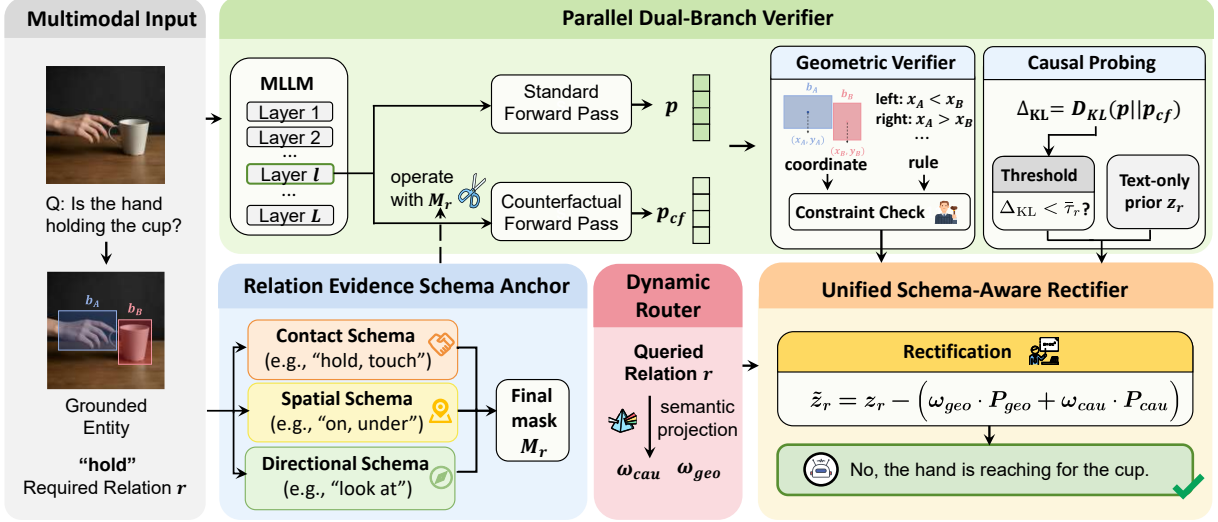


Figure 3: Illustration of RGFR. The framework employs three key components: (1) Semantic-Driven Dynamic Router, (2) Parallel Dual-Branch Verifier, and (3) Unified Schema-Aware Rectifier to effectively mitigate relation hallucinations in MLLMs.

deep causal layers, severe Attention Disorientation occurs: attention on actual relation evidence collapses, shifting predominantly to the salient entity centers. Driven by the imperative to satisfy textual priors, the model actively suppresses critical visual interaction constraints.

This phenomenon provides mechanistic verification of REA. Driven by the user’s directive, the model’s language prior actively suppresses visual interaction constraints to maintain narrative consistency. This exposes a critical insight: the alignment between entity perception and relation cognition is not static, but can be dynamically skewed by the model’s own reasoning process, making final-layer confidence or coarse global perturbations insufficient for relation-faithful verification. A faithful verifier should therefore track relation-specific evidence across the internal reasoning process and test whether the final relation decision remains grounded in the visual cues that actually support the queried predicate. To quantitatively assess this systematic pattern, we conduct an outcome-conditioned diagnostic on the DALL-E subset of MMRel; details are provided in Appendix A.3.

4 Method

4.1 Overview

We propose Relation-Grounded Faithful Reasoning (RGFR), a novel training-free decoding-time intervention framework designed to mitigate Relation Hallucinations by interrupting the prior-driven rationalization of interactions. As illustrated

in Figure 3, RGFR implements a Parallel Dual-System Rectification strategy via three key components: (1) A Semantic-Driven Dynamic Router that acts as a continuous soft-gating mechanism, zero-shot leveraging the LLM’s intrinsic semantic space to allocate adaptive weights for the simultaneously active intervention branches. (2) A Parallel Dual-Branch Verifier that simultaneously enforces symbolic spatial constraints and executes counterfactual causal probing, utilizing a Relation Evidence Schema (RES) to surgically blind and extract non-compensatory visual signals. (3) A Unified Schema-Aware Rectifier that fuses geometric constraints with prior-calibrated causal uncertainty to dynamically penalize unfaithful language priors and surface grounded relations. Through this parallel reasoning-phase intervention, RGFR transforms the MLLM into a rigorous visual-causal reasoner.

4.2 Semantic-Driven Dynamic Router

While traditional routing often implies a rigid binary decision, our Semantic-Driven Dynamic Router acts as a continuous soft-gating mechanism. It evaluates the linguistic properties of the queried relation r to allocate adaptive rectification weights for the simultaneously active parallel intervention branches. Operating prior to visual intervention, it serves as a zero-shot cognitive gate that maps r into a continuous probability simplex using the LLM’s pre-trained semantic space.

Guided by the semantics of spatial and physical interactions, the Router categorizes evidence

reliance into geometric and causal axes. Let $\mathcal{E}(x)$ denote the static unembedding function projecting token x into a d -dimensional vector. To establish this continuous projection, we define two compact anchor sets: a geometric set $\mathcal{A}_{geo}$ consisting of spatial prepositions, and a causal set $\mathcal{A}_{cau}$ comprising action verbs. Their semantic centroids, $\mathbf{e}_k$, are computed offline:

$$\mathbf{e}_k = \frac{1}{|\mathcal{A}_k|} \sum_{a \in \mathcal{A}_k} \mathcal{E}(a), \quad k \in \{geo, cau\} \quad (1)$$

This offline extraction incurs zero latency during inference.

To avoid manual thresholds and hard splits, we define the semantic similarity score as $s_k = \cos(\mathcal{E}(r), \mathbf{e}_k)$. The dynamic geometric weight ω_{geo} is derived via a temperature-scaled softmax projection:

$$\omega_{geo} = \frac{\exp(s_{geo}/\tau)}{\exp(s_{geo}/\tau) + \exp(s_{cau}/\tau)} \quad (2)$$

where τ controls the distribution sharpness, and the causal probing weight is its complement $\omega_{cau} = 1 - \omega_{geo}$.

Consequently, purely spatial relations are heavily weighted toward coordinate validation ($\omega_{geo} \rightarrow 1$), physical actions lean predominantly toward causal probing ($\omega_{cau} \rightarrow 1$), and composite relations organically receive balanced dual-branch weights. This continuous rectification ensures that both verifiers jointly contribute to the final inference without any rigid cutoffs.

4.3 Parallel Dual-Branch Verifier

Following the semantic routing, RGFR activates two complementary branches to defend against coordinate blindness and co-presence bias.

Geometric Verifier Branch. To bypass the coordinate blindness of attention mechanisms, this branch enforces symbolic spatial constraints directly on the localized bounding boxes b_A and b_B . If the physical configuration contradicts the spatial predicate, it outputs an infinite penalty $P_{geo} = \infty$; otherwise, $P_{geo} = 0$. This deterministic verifier operates entirely outside the neural network, incurring zero computational overhead.

Causal Probing Branch. For complex actions, this branch executes a bifurcated counterfactual intervention. First, operating strictly after entity localization, a Relation Evidence Schema (RES)

translates relations into deterministic spatial sub-masks $\mathbf{M}^{(s)} \in \{0, 1\}^{H \times W}$. Based on physical mechanics, RES defines three isolated canonical schemas: Contact (proximity gaps), Spatial (structural alignment), and Directional (trajectory blinding).

To observe the model’s reliance on genuine evidence, we perform a counterfactual forward pass via a bifurcated strategy with KV-cache sharing, isolating the masking to deep causal layers. Let $\mathbf{S}^{(l)}$ be the pre-softmax attention matrix at layer l . The counterfactual attention for schema s is surgically blinded:

$$\mathbf{S}_{cf}^{(l,s)} = \mathbf{S}^{(l)} \odot (\mathbf{J} - \mathbf{M}^{(s)}) - \infty \cdot \mathbf{M}^{(s)} \quad (3)$$

where $\mathbf{J}$ is the all-ones matrix. Crucially, interactions are governed by the Weakest Link Principle (e.g., strong spatial alignment cannot compensate for missing physical contact). Thus, we aggregate the schema-specific evidence sensitivities $e_s^{(l)}$ using a non-compensatory geometric mean:

$$E_r^{(l)} = \prod_{s \in \mathcal{S}_r} \left(e_s^{(l)} \right)^{w_{r,s}} \quad (4)$$

where $w_{r,s}$ is the prior weight of schema s . This unbounded penalty ensures that the absence of any single critical evidence directly dismantles the co-presence hallucination.

4.4 Unified Schema-Aware Rectifier

While the parallel verifier exposes missing evidence, MLLMs may still hallucinate relations due to strong linguistic inertia. To completely dismantle this, we introduce a prior-calibrated rectification mechanism.

We first quantify the model’s epistemic uncertainty regarding visual evidence via the Kullback-Leibler (KL) divergence between the factual ($\mathbf{p}$) and counterfactual ($\mathbf{p}_{cf}$) predictive distributions: $\Delta_{KL} = D_{KL}(\mathbf{p} \parallel \mathbf{p}_{cf})$. A low Δ_{KL} indicates that blinding the evidence minimally shifted predictions, exposing a severe prior-driven hallucination.

To aggressively penalize relations susceptible to textual overconfidence, we calibrate the penalty using the unconditional text-only prior $\mathbf{p}_{prior,r}$. The causal penalty P_{cau} is conditionally activated when Δ_{KL} falls below a threshold τ :

$$P_{cau} = \bar{\alpha} \cdot (1 + \gamma \cdot \mathbf{p}_{prior,r}) \cdot \mathbb{I}(\Delta_{KL} < \tau) \cdot z_r \quad (5)$$

where $\bar{\alpha}$ is the penalty base, γ scales the prior, z_r is the original logit, and $\mathbb{I}(\cdot)$ is the indicator.

Table 1: Accuracy (%), Precision (%), and F1 (%) comparison of the baseline without intervention (Vanilla) and state-of-the-art baselines: VCD, CAP, and Calibrate, with RGFR (Ours) on the MMRel and R-Bench benchmarks. The best results appear in **bold**, and the second-best in underlined.

Model	Method	MMRel			R-Bench		
		Acc	Prec	F1	Acc	Prec	F1
LLaVA-1.5	Vanilla	57.50	55.15	69.53	69.72	63.26	76.10
	VCD (Leng et al., 2024)	<u>60.05</u>	<u>57.16</u>	<u>70.79</u>	71.23	64.59	76.97
	CAP (Wu et al., 2025)	58.08	55.52	69.97	70.14	63.59	76.38
	Calibrate (Zheng et al., 2025)	58.41	56.03	69.67	<u>74.76</u>	<u>74.66</u>	<u>83.85</u>
	RGFR (Ours)	81.97	80.69	83.16	77.96	77.83	84.41
InstructBLIP	Vanilla	58.54	56.10	70.18	76.72	80.91	82.65
	VCD (Leng et al., 2024)	58.62	55.84	70.17	80.48	83.59	<u>86.98</u>
	CAP (Wu et al., 2025)	<u>62.44</u>	<u>59.75</u>	<u>71.47</u>	80.18	81.28	85.79
	Calibrate (Zheng et al., 2025)	60.88	57.80	71.41	<u>80.75</u>	84.34	85.61
	RGFR (Ours)	84.14	82.99	84.99	83.67	<u>83.89</u>	87.99
Qwen2.5-VL	Vanilla	77.85	74.60	80.55	76.64	<u>91.58</u>	79.59
	VCD (Leng et al., 2024)	79.78	76.71	81.96	81.53	89.66	<u>85.20</u>
	CAP (Wu et al., 2025)	<u>82.67</u>	<u>79.49</u>	<u>84.12</u>	80.83	87.42	85.06
	Calibrate (Zheng et al., 2025)	79.72	77.01	81.86	<u>81.64</u>	89.55	85.17
	RGFR (Ours)	87.44	89.73	87.22	82.20	91.59	85.31
InternVL3.5	Vanilla	77.01	71.91	80.84	83.87	85.33	<u>88.21</u>
	VCD (Leng et al., 2024)	76.45	72.74	79.74	82.39	86.75	86.65
	CAP (Wu et al., 2025)	<u>80.06</u>	<u>75.43</u>	<u>82.72</u>	<u>84.44</u>	87.32	88.03
	Calibrate (Zheng et al., 2025)	77.22	71.52	81.24	84.08	<u>86.80</u>	88.16
	RGFR (Ours)	87.85	86.98	88.31	86.35	86.41	89.87

Ultimately, the Rectifier fuses the geometric constraints and the neural causal penalty via the semantic routing weights (Eq. 2). The final rectified logit $\tilde{z}_r$ is:

$$\tilde{z}_r = z_r - \left(\omega_{geo} \cdot P_{geo} + \omega_{cau} \cdot P_{cau} \right) \quad (6)$$

This explicit rectification suppresses ungrounded rationalization, forcing genuinely visual-grounded relations to organically surface during decoding.

5 Experiments

5.1 Experimental Setup

Benchmarks. We employ R-Bench (Wu et al., 2024b) and MMRel (Nie et al., 2025) for comprehensive relation hallucination evaluation. From MMRel, we utilize the VG and DALL-E categories, totaling 11871 samples. From R-Bench, we focus on the image-level subset, totaling 7787 samples. These selections ensure broad coverage of fine-grained annotations, systematically evaluating both pure spatial configurations and complex physical

interactions. Finally, to verify that our intervention avoids over-penalization, we employ MM-Vet (Yu et al., 2024) to evaluate general utility and the inference efficiency (latency). See detailed descriptions of these datasets in Appendix A.1.

Evaluation Models and Baselines. We evaluate the effectiveness of RGFR on four diverse open-source MLLMs: LLaVA-1.5-7B (Liu et al., 2023), InstructBLIP-7B (Dai et al., 2023), Qwen2.5-VL-7B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025a). To demonstrate the superiority of our framework, we compare RGFR against recent state-of-the-art mitigation strategies. These include Visual Contrastive Decoding (VCD) (Leng et al., 2024), Constraint-Aware Prompting (Wu et al., 2025) and Calibrate (Zheng et al., 2025). All baseline methods are evaluated on a unified test dataset using consistent metrics. Baseline details are provided in Appendix A.2.

Implementation Details and Metrics. RGFR operates without training or parameter updates,

Table 2: Fine-grained accuracy (%) on the MMRel dataset across Spatial and Action subsets. The results explicitly validate the effectiveness of RGFR’s dual-branch verifiers against coordinate blindness and co-presence bias.

Model	Method	Accuracy	
		Spatial	Action
LLaVA-1.5	Vanilla	50.06	64.95
	CAP (Wu et al., 2025)	50.26	65.90
	Calibrate (Zheng et al., 2025)	50.21	66.61
	RGFR (Ours)	90.19	73.71
InstructBLIP	Vanilla	50.46	66.63
	Calibrate (Zheng et al., 2025)	50.87	70.88
	CAP (Wu et al., 2025)	52.42	72.47
	RGFR (Ours)	89.18	78.10
Qwen2.5-VL	Vanilla	76.62	81.22
	Calibrate (Zheng et al., 2025)	76.36	83.09
	CAP (Wu et al., 2025)	80.97	84.37
	RGFR (Ours)	90.20	84.61
InternVL3.5	Vanilla	71.77	82.25
	Calibrate (Zheng et al., 2025)	73.41	81.03
	CAP (Wu et al., 2025)	76.56	83.56
	RGFR (Ours)	89.05	85.22

using the released checkpoints of each evaluated MLLM under the same decoding and evaluation protocol. Our implementation is built on a PyTorch-based inference pipeline and Hugging Face model interfaces, utilizing Grounding DINO (Liu et al., 2024b) for object bounding box extraction. All experiments are conducted on a server equipped with 8 NVIDIA RTX 3090 GPUs. For MMRel and R-Bench, we report Accuracy (%), Precision (%), and F1-score (%) to evaluate relation hallucination mitigation. For the fine-grained Spatial and Action subsets, we report Accuracy (%). For MM-Vet, we follow official evaluation protocol and report the utility score, while inference efficiency is measured by average latency per query.

5.2 Main Results

Our proposed RGFR consistently surpasses existing training-free mitigation baselines across all four evaluated MLLMs on both the MMRel and R-Bench benchmarks, demonstrating a robust capability for visually grounded relation reasoning against co-presence bias (Table 1 and Table 2).

Performance on different Benchmarks. As shown in Table 1, RGFR achieves absolute accuracy gains of up to 25.6 percentage points over the baselines across the evaluated MLLMs. This validates the dual-system architecture’s efficiency in orchestrating grounded reasoning from complex

Table 3: Failure analysis on the DALL-E subset of MMRel. Distribution of remaining errors across failure types.

Failure type	Share (%)
Missing entity localization	22.5
Incorrect or imprecise box	14.3
Overlapping or ambiguous entities	18.2
Ambiguous relation evidence	27.6
Out-of-schema predicate	17.4

visual contexts. On MMRel, RGFR increases the overall accuracy of LLaVA-1.5 from 57.50% to 81.97%. The result indicates that our intervention outperforms existing training-free methods in handling relation-heavy samples. Similar improvements on R-Bench demonstrate the versatility of RGFR. Furthermore, RGFR transfers effectively to recent larger-scale models, improving the accuracy of InternVL3.5 from 77.01% to 87.85% and mitigating ungrounded rationalization in Qwen2.5-VL.

Robustness against Cognitive Pathologies. To assess resilience against specific relational vulnerabilities, we evaluate the models on the Spatial and Action subsets of MMRel (Table 2). For LLaVA-1.5, RGFR substantially improves accuracy from 50.06% to 90.19% in the Spatial category through strict geometric constraints, and from 64.95% to 73.71% in the Action category by isolating physical interaction evidence. RGFR also delivers consistent gains for stronger models, including Qwen2.5-VL and InternVL3.5. These results indicate that RGFR effectively defends against sophisticated, prior-driven illusions, specifically neutralizing both coordinate blindness and co-presence bias.

Failure Analysis. To analyze the remaining failures, we categorize the erroneous predictions on the DALL-E subset of MMRel under the default setting. As shown in Table 3, ambiguous relation evidence is the largest category (27.6%), while missing localization and imprecise bounding boxes together account for 36.8% of the remaining errors. Future work could further improve visual grounding and relation disambiguation.

5.3 Hyperparameter and Schema Sensitivity

To assess hyperparameter and schema sensitivity, we evaluate Qwen2.5-VL on the same DALL-E subset of MMRel, varying one factor at a time (Table 4). Performance remains stable under moderate variations but declines with an overly strict ground-

Table 4: Hyperparameter and schema sensitivity on the DALL-E subset of MMRel. Bold indicates defaults.

Factor	Tested Setting	Performance (Acc.)
Routing temperature	[0.5, 1.0 , 2.0]	[84.2%, 87.8% , 81.8%]
KL threshold	[0.05, 0.10 , 0.20]	[84.4%, 87.8% , 86.3%]
Prior penalty scale (α)	[0.5, 1.0 , 2.0]	[84.8%, 87.8% , 85.1%]
Grounding DINO threshold	[0.30 , 0.40, 0.50]	[87.8% , 86.4%, 83.3%]
Relation schema variant	[reduced, full , expanded]	[84.2%, 87.8% , 84.5%]

ing threshold or a poorly scaled prior penalty. The full relation schema performs best, while the expanded schema yields comparable accuracy and the reduced schema performs worse. Overall, RGFR maintains a practical operating range but remains sensitive to extreme settings.

5.4 Intervention Layer Study

To assess the effect of intervention depth, we divide the LLaVA-1.5 language backbone into three contiguous intervals representing early, middle, and late stages, and include an all-layer setting as a control. As shown in Table 5, the late-stage interval $L22-L32$ performs best in accuracy, precision, and F1, while the early-stage interval performs worst. The all-layer setting is slightly less effective, suggesting that targeted late-stage intervention is preferable to broader layer coverage.

Table 5: Intervention layer analysis on MMRel with LLaVA-1.5.

Layers	Accuracy	Precision	F1
$L0-L10$	59.85	57.10	66.82
$L11-L21$	62.40	60.55	68.90
$L22-L32$	81.97	80.69	83.16
$L0-L32$	74.65	72.78	75.25

5.5 Ablation Study

To analyze the contributions of RGFR’s key components, we conduct ablation studies focusing on its core mechanisms: the dual-branch verifiers, the interaction mask, and the weighted penalty. Experiments are carried out on the DALL-E subset of the MMRel benchmark using three representative MLLMs: LLaVA-1.5, Qwen2.5-VL, and InternVL3.5.

On the Dual-Branch Verifiers. We evaluate whether handling distinct spatial and physical constraints requires decoupled strategies by comparing four settings on the DALL-E subset: (1) vanilla inference, (2) Geometric Verifier only, (3) Causal Probing Verifier only, and (4) full RGFR. As shown in Table 6, the two verifiers exhibit com-

Table 6: Ablation study of the dual-branch verifiers on the DALL-E subset. We compare Accuracy (%) across the Spatial and Action subsets.

Model	Method	Spatial	Action	Overall
LLaVA-1.5	Vanilla	52.34	48.65	50.49
	Geometric Only	76.51	49.12	62.81
	Causal Only	51.08	71.43	61.25
	RGFR	77.15	72.38	74.76
Qwen2.5-VL	Vanilla	75.20	71.30	73.25
	Geometric Only	88.45	70.55	79.50
	Causal Only	74.82	85.60	80.21
	RGFR	89.12	86.45	87.78
InternVL3.5	Vanilla	72.85	69.45	71.15
	Geometric Only	85.10	68.20	76.65
	Causal Only	73.05	82.55	77.80
	RGFR	85.88	83.20	84.54

plementary effects across all three MLLMs. The Geometric-only setting substantially improves Spatial accuracy but performs less effectively on Action, whereas the Causal-only setting shows the opposite pattern. By dynamically routing queries to their appropriate verifiers, the full RGFR framework successfully achieves the best of both worlds, yielding the highest accuracy across all dimensions and validating the structural necessity of our decoupled architecture.

On Targeted Interaction Masking. To rigorously assess the mechanistic advantage of our Relation Evidence Schema (RES), we compare three configurations: (1) a no-masking baseline, (2) indiscriminate global noise, and (3) our targeted masking approach. Results are reported in Figure 4, evaluating Accuracy (%) across three MLLMs.

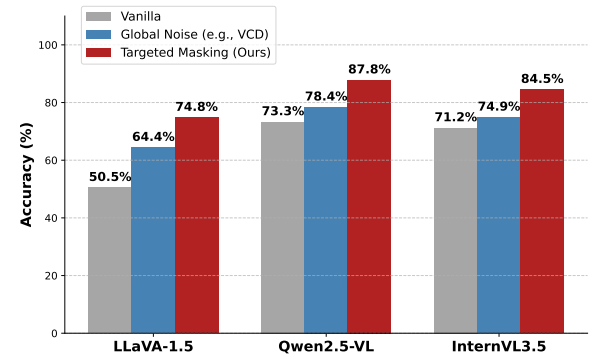


Figure 4: Comparison of masking strategies on the DALL-E subset. Targeted masking consistently outperforms no-masking baseline and global noise across three MLLMs.

As illustrated in Figure 4, the no-masking baseline (gray bars) suffers from semantic co-presence

bias. While global noise (blue bars) mitigates object-level hallucinations, its indiscriminate perturbations tend to blur spatial boundaries essential for physical constraints, thereby compromising relational accuracy. In contrast, RGFR’s targeted masking (red bars) achieves the highest accuracy across all models, proving that surgically isolating interaction pathways is crucial for robust relational reasoning.

Table 7: Cost Analysis on Utility and Efficiency. Comparison of Utility and Inference Latency (seconds per query) on Qwen2.5-VL. RGFR achieves a superior utility-efficiency balance compared to baselines.

Method	Utility $\uparrow$	Latency $\downarrow$
<i>Baseline</i>		
Qwen2.5-VL (Vanilla)	69.0	3.096s
<i>Decoding Interventions</i>		
+ VCD	54.4	7.121s
+ CAP	57.0	6.159s
<i>Our Method</i>		
+ RGFR	62.4	5.637s

5.6 Evaluation of Inference Utility and Efficiency

We assess the trade-off between general multi-modal utility on MM-Vet and inference efficiency via Qwen2.5-VL per-query latency. As detailed in Table 7, test-time interventions inevitably introduce computational overhead. However, RGFR achieves a far superior utility-efficiency balance compared to existing inference-stage interventions. Specifically, RGFR increases latency by 2.541s per query. In contrast, VCD and CAP incur much higher delays of 4.025s and 3.063s while suffering sharper utility drops to 54.4 and 57.0, respectively. These results prove RGFR mitigates relation hallucinations with manageable latency overhead. Expanded evaluations are provided in Appendix A.6.

6 Conclusion

We explored relation hallucinations in MLLMs. Our analysis reveals that Relation Evidence Amnesia drives models to rationalize non-existent interactions. To address this, we propose RGFR, a training-free intervention that decouples geometric grounding from physical constraints. RGFR interrupts hallucinatory priors by dynamically rectifying between explicit entity anchors and targeted masking, transforming verification from generative guesswork into rigid, boundary-preserving calibra-

tion. Extensive experiments across diverse MLLMs establish a new state-of-the-art on relational benchmarks, improving recognition accuracy by over 8.1% while preserving utility with minimal latency. Ultimately, our work establishes a new paradigm explicitly guiding inference-time reasoning based on actual visual evidence, paving the way for trustworthy multi-modal reasoning.

Limitations

Our study focuses on relation hallucinations in single-image, text-query settings, where the key challenge is to verify whether a queried relation is visually supported. While RGFR consistently improves relation faithfulness across MMRel and R-Bench, its current evaluation does not fully cover more complex multimodal scenarios, such as multi-turn dialogue, video-based relation reasoning, long-context agentic workflows, or open-ended dense relational description. In addition, RGFR is a decoding-time intervention and therefore introduces extra inference overhead compared with vanilla generation. Although our experiments show a favorable utility–efficiency trade-off, further optimization may be needed for latency-sensitive deployment. Its calibration may also require adjustment when transferred to substantially different model architectures. Extending RGFR to broader modalities, richer relation taxonomies, and other MLLM architectures remains future work.

Ethical Considerations

This work mitigates MLLM relation hallucinations by enforcing visual grounding over language priors. We use existing relation hallucination benchmarks solely for controlled evaluation. By introducing explicit inference-time verification, RGFR may benefit applications in which incorrect spatial or action relations could mislead users. However, RGFR may still fail on ambiguous, low-quality, or out-of-distribution inputs and exhibit model-dependent behavior. It should thus serve as a complementary reliability mechanism rather than a substitute for robust data curation, model alignment, or human oversight in high-stakes applications.

Acknowledgments

This work is supported by National Natural Science Foundation of China (U22B2017), and International Cooperation Foundation of Hubei Province, China (2024EHA032).

References

- Vishnu Narayanan Anilkumar, Abhijith Sreesylesh Babu, Trieu Hai Vo, Mohankrishna Kolla, and Alexander Cuneo. 2026. [Relationship-aware safety unlearning for multimodal llms](#). *Preprint*, arXiv:2603.14185.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report.
- Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024. Halc: Object hallucination reduction via adaptive focal-contrast decoding.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.
- Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. 2024. [DAMRO: Dive into the attention mechanism of LVLm to reduce object hallucination](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7696–7712, Miami, Florida, USA. Association for Computational Linguistics.
- Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models.
- Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. 2021. [Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding](#). In *Proceedings of the 29th ACM International Conference on Multimedia*, MM '21, page 2344–2352. ACM.
- Xuming Hu, Zhijiang Guo, Zhiyang Teng, Irwin King, and Philip S. Yu. 2023. [Multimodal relation extraction with cross-modal retrieval and synthesis](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 303–311, Toronto, Canada. Association for Computational Linguistics.
- Wen Huang, Hongbin Liu, Minxin Guo, and Neil Zhenqiang Gong. 2024a. Visual hallucinations of multimodal large language models.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. [Look before you leap: An exploratory study of uncertainty analysis for large language models](#). *IEEE Transactions on Software Engineering*, 51(2):413–429.
- Zhipeng Huang, Zhizheng Zhang, Zheng-Jun Zha, Yan Lu, and Baining Guo. 2024b. Relationvlm: Making large vision-language models understand visual relations.
- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislaw Fort, and 17 others. 2022. Language models (mostly) know what they know.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation.
- Jihoon Lee and Min Song. 2025. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.
- Shenshen Li, Xing Xu, Wenxin Meng, Jingkuan Song, Chong Peng, and Heng Tao Shen. 2025. [Mitigating hallucinations in large vision-language models via reasoning uncertainty-guided refinement](#). *IEEE Transactions on Multimedia*, 27:7380–7391.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. [Visual instruction tuning](#). *Preprint*, arXiv:2304.08485.
- Shi Liu, Kecheng Zheng, and Wei Chen. 2024a. Paying more attention to image: A training-free method for alleviating hallucination in llms.
- Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2024b. [Grounding dino: Marrying dino with grounded pre-training for open-set object detection](#). *Preprint*, arXiv:2303.05499.
- Yiwei Ma, Weihuang Lin, Zhibin Wang, Jiayi Ji, Xiaoshuai Sun, Chia-Wen Lin, and Rongrong Ji. 2026. [Boosting Multi-Modal Large Language Model With Enhanced Visual Features](#). *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 48(04):4524–4538.
- Jiahao Nie, Gongjie Zhang, Wenbin An, Yun Xing, Yappeng Tan, Alex C. Kot, and Shijian Lu. 2025. Mmrel: Benchmarking relation understanding in multi-modal large language models.

- Guanglin Niu and Bo Li. 2026. [Radd: Retrieval-augmented discrete diffusion for multi-modal knowledge graph completion](#). *Preprint*, arXiv:2604.25693.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2024. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models.
- Sirat Samyoun, Yingtai Xiao, and Jian Du. 2026. [Attribution-guided multi-object hallucination and bias detection in vision-language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4529–4548, Rabat, Morocco. Association for Computational Linguistics.
- Sreetama Sarkar, Yue Che, Alex Gavin, Peter A. Bearel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation.
- Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025a. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency.
- Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024b. [Mitigating hallucinations in large vision-language models with instruction contrastive decoding](#). *Preprint*, arXiv:2403.18715.
- Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. 2025b. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink.
- Jiarui Wu, Zhuo Liu, and Hangfeng He. 2025. Mitigating hallucinations in multimodal spatial relations through constraint-aware prompting.
- Kai Wu, Boyuan Jiang, Zhengkai Jiang, Qingdong He, Donghao Luo, Shengzhi Wang, Qingwen Liu, and Chengjie Wang. 2024a. Noiseboost: Alleviating hallucination with noise perturbation for multimodal large language models.
- Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024b. Evaluating and analyzing relationship hallucinations in large vision-language models.
- Yike Wu, Yiwei Wang, and Yujun Cai. 2026. [Chainmpq: Interleaved text-image reasoning chains for mitigating relation hallucinations](#). *Preprint*, arXiv:2510.06292.
- Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyang Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, Yifan Peng, Mingquan Lin, and Zongyuan Ge. 2026. [Thinking in uncertainty: Mitigating hallucinations in mllms with latent entropy-aware decoding](#). *Preprint*, arXiv:2603.13366.
- Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025a. Visionthink: Smart and efficient vision language model via reinforcement learning.
- Tiancheng Yang, Lin Zhang, Jiaye Lin, Guimin Hu, Di Wang, and Lijie Hu. 2025b. Tracing and mitigating hallucinations in multimodal llms via dynamic attention localization.
- 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).
- Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities.
- Xinglang Zhang, Yunyao Zhang, ZeLiang Chen, Junqing Yu, Wei Yang, and Zikai Song. 2026a. [Logical phase transitions: Understanding collapse in LLM logical reasoning](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18836–18860, San Diego, California, United States. Association for Computational Linguistics.
- Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Chen Xu, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2025. Siren’s song in the ai ocean: A survey on hallucination in large language models.
- Yunyao Zhang, Xinglang Zhang, Junxi Sheng, Wenbing Li, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Zikai Song. 2026b. [Semantic-aware logical reasoning via a semiotic framework](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18349–18374, San Diego, California, United States. Association for Computational Linguistics.
- Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2025. Reefknot: A comprehensive

benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. [Analyzing and mitigating object hallucination in large vision-language models](#). *Preprint*, arXiv:2310.00754.

A Appendix

A.1 Benchmark Details

MMRel. MMRel (Nie et al., 2025) is a relation-understanding benchmark designed to evaluate whether multimodal large language models can faithfully judge relations between visually grounded entities. Instead of focusing only on object existence, MMRel constructs relation-centric yes/no questions that require the model to verify whether a queried subject–predicate–object relation is supported by the image. The benchmark covers three complementary relation categories: spatial relations, action relations, and comparative relations. Its images are drawn from both real-world and generated sources, including Visual Genome images and DALL-E generated images, which makes it suitable for testing relation grounding under both natural and controlled visual compositions. In our experiments, we use the Visual Genome and DALL-E portions of MMRel, totaling 11,871 samples. We further report fine-grained results on the Spatial and Action subsets, which directly correspond to the two failure modes studied in the main paper: coordinate blindness and co-presence bias.

R-Bench. R-Bench (Wu et al., 2024b) is a benchmark for evaluating relationship hallucinations in large vision–language models. It focuses on whether a model can distinguish visually supported relations from relations that are linguistically plausible but unsupported by the image. The benchmark contains both image-level and instance-level evaluation settings; the image-level setting asks models to verify relation statements at the whole-image level, while the instance-level setting further localizes relation judgments to specific object instances. This design is well aligned with our goal of testing whether models over-rely on object co-occurrence as relation evidence. We use the image-level subset in our experiments, totaling 7,787 evaluated samples under our filtering and evaluation setup. These samples provide broad coverage of relation hallucination cases in which entity presence alone is insufficient for a faithful answer.

MM-Vet. MM-Vet (Yu et al., 2024) is a general-purpose benchmark for evaluating integrated multimodal capabilities. It contains 218 open-ended questions paired with 200 images and measures six core abilities: recognition, optical character recognition, knowledge, language generation, spatial awareness, and mathematical reasoning. Unlike

relation-specific binary benchmarks, MM-Vet evaluates free-form answers and therefore provides a useful stress test for whether a mitigation method harms general visual reasoning and response quality. We use MM-Vet as a utility benchmark rather than as a relation hallucination benchmark. This allows us to assess whether RGFR improves relation faithfulness while preserving the broader multimodal competence of the base model.

A.2 Baseline Details

Visual Contrastive Decoding. Visual Contrastive Decoding (VCD) (Leng et al., 2024) is a training-free decoding method originally proposed to mitigate object hallucinations in large vision–language models. It contrasts the model distribution conditioned on the original image with the distribution conditioned on a distorted visual input, typically created by adding visual noise, and uses the contrast to reduce visually unsupported generations. In our setting, VCD serves as a representative decoding-time intervention baseline. Its global perturbation strategy is useful for testing whether generic visual contrast is sufficient for relation hallucination mitigation. However, relation verification often depends on fine-grained boundaries and localized interaction cues, so indiscriminate image-level perturbation can remove precisely the visual evidence needed for faithful relation judgments.

Constraint-Aware Prompting. Constraint-Aware Prompting (CAP) (Wu et al., 2025) is a prompting-based method for mitigating hallucinations in multimodal spatial relations. It augments the model input with explicit spatial and logical constraints so that the model is encouraged to reason about relation validity rather than relying only on language priors. We include CAP as a strong text-side baseline because it targets relation errors without modifying model parameters. In our unified evaluation, CAP uses the same images and binary relation questions as the other methods, allowing us to compare instruction-level constraint guidance with RGFR’s visual-evidence-grounded intervention.

Calibrate. Calibrate refers to the relation-hallucination mitigation strategy introduced with Reefknot (Zheng et al., 2025). Reefknot studies relationship hallucination in multimodal large language models and proposes calibration-oriented mitigation to reduce unsupported relation judgments. In our experiments, Calibrate represents

the output-calibration family of baselines: rather than explicitly masking relation evidence or enforcing geometric constraints, it adjusts the model’s relation decision by calibrating confidence or internal signals associated with the generated answer. This makes it complementary to VCD and CAP, and provides a useful comparison against RGFR’s explicit relation-grounded arbitration.

A.3 Relation Evidence Amnesia Diagnostic

We conduct an outcome-conditioned diagnostic on the DALL-E subset of MMRel using Qwen2.5-VL. We group the resulting predictions as correct, hallucinated, and RGFR-corrected under the same evaluation protocol. We measure the REA Index as the difference between entity-region attention and relation-evidence attention; larger values indicate stronger relative dependence on entity regions. We also report masking impact as the percentage of relation predictions that change after masking either predicate-relevant relation-evidence regions or localized entity regions. Table 8 summarizes both measures across the three outcome groups.

Table 8: Outcome-conditioned REA analysis on the DALL-E subset of MMRel. Mask impact is the prediction-change rate (%).

Group	REA	Relation mask	Entity mask
Correct	0.14	34.2	18.5
Hallucinated	0.52	9.3	41.7
RGFR-corrected	0.11	36.8	16.4

The results show a double dissociation: hallucinated predictions have a higher REA Index and are more sensitive to entity masking, whereas correct and RGFR-corrected predictions are more sensitive to relation-evidence masking. These results support REA as a systematic failure pattern in the evaluated setting and motivate explicit verification of predicate-specific evidence.

A.4 Predicate-to-Schema Composition

Contact, Spatial, and Directional denote reusable evidence primitives, not an exhaustive partition of relation predicates. Table 9 gives representative compositions. The router assigns continuous geometric and causal weights, and the causal branch can combine multiple primitives. New predicates can therefore reuse an existing composition; predicates that cannot be mapped reliably fall back to the base model and are counted as out-of-schema failures.

Table 9: Predicate-to-schema composition examples. A checkmark indicates a relevant evidence primitive.

Predicate	Contact	Spatial	Directional
riding	✓	✓	optional
holding	✓	✓	–
wearing	✓	✓	–
carrying	✓	✓	optional
looking at	–	✓	✓

A.5 Additional Evaluation Protocols

White-box relation verification. For MMRel and R-Bench, we retain each benchmark’s original question wording and apply the released chat template for each backbone. The scored answer alternatives are the model’s Yes and No continuations. RGFR reuses the same prompt for factual and counterfactual passes; only the relation-evidence mask differs. The mask is constructed after Grounding DINO localizes the queried subject and object. Contact masks cover overlap or nearest-boundary regions, Spatial masks cover relevant horizontal or vertical gaps, support, and alignment regions, and Directional masks cover the subject-to-object trajectory. Decoding and evaluation settings are held fixed between Vanilla and RGFR for each backbone. Hardware and model checkpoints are reported in Section 5.1.

Table 10: Commercial-API evaluation on the DALL-E subset of MMRel. RGFR-API is an API-compatible variant of white-box RGFR; latency is seconds per sample.

Model	Method	Acc.	F1	Calls	Latency
GPT-5.5	Vanilla API	74.8	74.1	1	2.65s
GPT-5.5	RGFR-API	85.2	84.8	3	8.62s
Gemini 3.5 Flash	Vanilla API	73.5	72.8	1	1.85s
Gemini 3.5 Flash	RGFR-API	81.6	81.2	3	5.85s

RGFR-API protocol. Commercial APIs do not expose the internal states needed by white-box RGFR. RGFR-API therefore makes three calls per sample. The first receives the original image and relation query. The second receives the relation-evidence-masked image and the same query. The third receives the original query and the two responses and is instructed to return the final binary decision after comparing their evidence. Vanilla API uses the original-image call only. We use the same DALL-E subset of MMRel, prompts, and evaluation criteria for both conditions. Table 10 reports accuracy, F1, calls, and latency under this pro-

Table 11: Negation robustness on the DALL-E subset of MMRel.

Method	Original	Negated	Consistency
Vanilla	73.3	56.5	48.0
RGFR	87.8	78.0	71.5

tocol. RGFR-API is an approximation of white-box RGFR and consequently has substantially higher latency.

Negation protocol. For each sample in the DALL-E subset of MMRel, we pair the original query with a manually constructed negated form and reverse the binary ground-truth label, producing one original–negated logical pair per sample. Original and negated accuracy are computed separately. Paired logical consistency is the percentage of pairs for which the two predictions exhibit the expected reversal. The resulting rates are reported in Table 11.

A.6 Expanded Utility and Efficiency

Table 12 extends the utility–efficiency analysis to MME and InternVL3.5-8B under the same hardware and inference configuration. Across all four backbone–benchmark settings, RGFR achieves a higher score and lower latency than VCD and CAP while remaining slower than vanilla decoding. These results support a favorable relative trade-off among the evaluated interventions, not negligible overhead over vanilla decoding.

Table 12: Utility scores and inference latency across backbones and benchmarks. Latency is measured in seconds per query. Bold marks the best intervention method.

Backbone	Benchmark	Metric	Vanilla	VCD	CAP	RGFR
Qwen2.5-VL	MM-Vet	Score	69.0	54.4	57.0	62.4
		Latency	3.096	7.121	6.159	5.637
	MME	Score	2316.7	2117.9	2183.7	2291.5
		Latency	1.957	5.842	4.914	3.271
InternVL3.5	MM-Vet	Score	49.9	45.3	45.9	48.6
		Latency	2.229	4.539	3.852	2.945
	MME	Score	2258.3	2089.8	2126.4	2215.6
		Latency	1.452	4.125	3.512	2.415

A.7 Relative Error Reduction on R-Bench

Because R-Bench has stronger vanilla baselines than MMRel, Table 13 reports relative error reduction in addition to accuracy. We compute relative error reduction (RER) as follows:

$$\text{RER} = \frac{\text{Acc}_{\text{RGFR}} - \text{Acc}_{\text{Vanilla}}}{100 - \text{Acc}_{\text{Vanilla}}}. \quad (7)$$

RGFR reduces error by 15.4%–29.9% across the four backbones.

Table 13: Accuracy and relative error reduction (RER) on R-Bench across four backbones.

Model	Vanilla	RGFR	RER
LLaVA-1.5	69.72	77.96	27.2%
InstructBLIP	76.72	83.67	29.9%
Qwen2.5-VL	76.64	82.20	23.8%
InternVL3.5	83.87	86.35	15.4%

A.8 Qualitative Case Study

Figure 5 presents representative MMRel cases in a boxed image–text format. The cases show two complementary behaviors: RGFR suppresses an unsupported spatial relation caused by entity co-presence, and it also recovers a visually supported action relation that the base model fails to recognize.

Figure 6 further presents representative cases from the R-Bench image-level subset. These examples illustrate that RGFR can both recover visually supported relations that are missed by the base model and reject plausible but unsupported relations induced by object co-occurrence.

NEGATIVE CASE: MMRel Spatial



Prompt: Is there a treehouse left to a mirror in the image?

Ground Truth: No

Base Model: Yes. Although both entities are correctly identified, the model fails to enforce their spatial configuration and accepts the reversed left-to relation.

RGFR: No. RGFR verifies the grounded coordinates and identifies that the treehouse is to the right of the mirror.

The corrected decision follows the visual relation rather than relying on entity presence alone.

POSITIVE CASE: MMRel Action



Prompt: Does a forklift lift a bulldozer in the image?

Ground Truth: Yes

Base Model: No. The model detects the two machines but misses the lifting interaction between them, rejecting a relation that is grounded by the contact and support configuration.

RGFR: Yes. RGFR focuses on the action-bearing visual evidence and verifies that the forklift is supporting and lifting the bulldozer, recovering the true relation.

The corrected decision follows the actual action relation rather than treating the objects as independent visual evidence.

Figure 5: Qualitative case study of RGFR on MMRel spatial and action relation examples.

POSITIVE CASE: R-Bench Image-Level



Prompt: Is the person in the image riding the exercise bike while others are watching?

Ground Truth: Yes

Base Model: No. The model overlooks the riding interaction and treats the visible exercise-bike scene as insufficient evidence for the queried relation.

RGFR: Yes. RGFR preserves the image-grounded action evidence: the person is riding the exercise bike while nearby people are watching.

The corrected answer verifies the full relation expressed in the image instead of discarding the action context.

NEGATIVE CASE: R-Bench Image-Level



Prompt: Is the white sports car parked next to a blue car?

Ground Truth: No

Base Model: Yes. The model grounds the white sports car but misattributes a nearby vehicle as the queried blue car, accepting an unsupported next-to relation.

RGFR: No. RGFR re-checks the local visual evidence around the white car and identifies that the adjacent vehicle is not blue, so the queried relation is absent.

The revised answer follows the actual relation and attribute evidence rather than relying on generic car co-occurrence.

Figure 6: Qualitative case study of RGFR on R-Bench image-level relation hallucination examples.