PATH CONFLICT: UNDERSTANDING AND MITIGATING HALLUCINATIONS IN MULTIMODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) are often assumed to hallucinate because they under-utilize visual evidence and over-rely on linguistic priors. We show that this explanation is incomplete: hallucinated generations can exhibit *stronger*, rather than weaker, visual-pathway influence. To understand this counter-intuitive behavior, we introduce **Path Conflict**, which characterizes directional inconsistency between Image-to-Text (I2T) and Text-to-Text (T2T) pathways during token prediction. We further define Conflict Ratio and Conflict Strength to quantify the frequency and magnitude of such conflicts at the attention-head level. Our analysis reveals that hallucinated samples exhibit substantially stronger Path Conflict, driven primarily by amplified I2T contributions while T2T contributions remain comparatively stable. This suggests that hallucination can arise from *misaligned visual influence*, rather than simply insufficient visual grounding. Based on this finding, we propose a lightweight, training-free **Conflict-aware Intervention** that selectively regulates I2T and T2T flows within conflict-prone heads. Experiments on POPE, MCQ-POPE, and CHAIR consistently outperform strong inference-time baselines, including AllPath, while pathway-specific ablations reveal the asymmetric and complementary roles of I2T and T2T. Our results highlight inter-path directional conflict as an important mechanism-level perspective for understanding and mitigating MLLM hallucination.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023; Chen et al., 2023; Dai et al., 2023; Zhu et al., 2024) are commonly thought to hallucinate because they fail to sufficiently rely on visual evidence and instead fall back on linguistic priors (Ji et al., 2023; Li et al., 2023b; Gunjal et al., 2024; Lovenia et al., 2024; Qian et al., 2026; Bai et al., 2025; Huang et al., 2024; Leng et al., 2023; Zhang et al., 2024; Arif et al., 2025; Yang et al., 2025). This explanation has shaped a large body of work on hallucination mitigation: strengthen visual grounding, increase visual attention, or suppress excessive dependence on previously generated text. Yet our analysis reveals a strikingly different behavior:

hallucinated generations can receive stronger, rather than weaker, visual influence.

This observation challenges a simple "insufficient vision" account of multimodal hallucination. If the visual pathway is already strongly involved, why does the model still produce content that contradicts the image?

We argue that the missing piece is not merely *how much* information each pathway contributes, but *whether different pathways agree on the prediction*. During autoregressive generation, visual evidence influences the current token through the Image-to-Text (I2T) pathway, while previously generated linguistic context contributes through the Text-to-Text (T2T) pathway. Existing mechanistic studies have largely analyzed these information sources separately, measuring the contribution of individual modules, attention heads, or pathways to hallucinated predictions (Qian et al., 2026; Bai et al., 2025; Liu et al., 2024a; Huang et al., 2024; Leng et al., 2023; Zhang et al., 2024; Arif et al., 2025; Yang et al., 2025; Tang et al., 2025). However, two pathways can both exert substantial influ-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

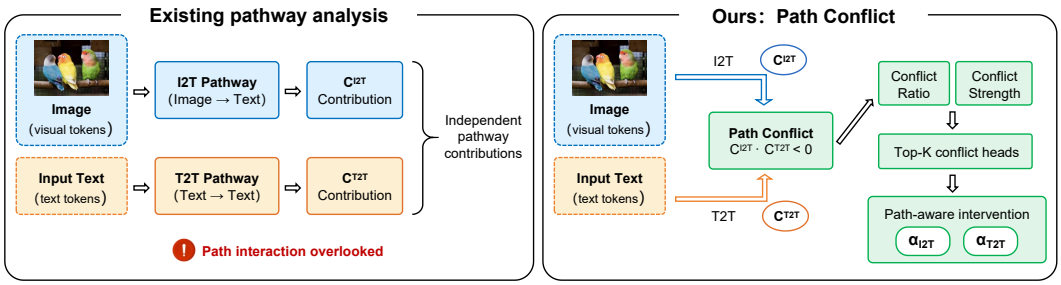

Figure 1: Beyond pathway strength: Path Conflict captures whether I2T and T2T pathways push the same prediction in compatible directions, enabling targeted intervention.

ence while pushing the same token prediction in opposite directions. Such directional disagreement is invisible to pathway-wise analysis.

To capture this overlooked interaction, we introduce **Path Conflict**: the phenomenon that I2T and T2T pathways exert opposite directional effects on the same token prediction. Rather than asking whether visual or textual information is strong enough, Path Conflict asks a different question: *do visual evidence and linguistic context push the model toward the same decision?* We quantify this behavior at the attention-head level using two complementary measures. **Conflict Ratio** captures how frequently the two pathways disagree in direction, while **Conflict Strength** measures the magnitude of their competing contributions.

Analyzing Path Conflict reveals a distinct signature of hallucinated generation. Conflicts are highly concentrated in a small subset of attention heads, and hallucinated samples exhibit substantially stronger conflicts than non-hallucinated ones. More importantly, these stronger conflicts are driven primarily by a pronounced increase in I2T contribution, whereas T2T contribution changes comparatively little. In other words, the visual pathway is not simply absent or suppressed during hallucination. It can become *more influential but directionally misaligned* with the linguistic context. This suggests a different failure mode: hallucination can emerge not from insufficient visual grounding, but from **misaligned visual influence**.

This finding naturally raises a second question: if Path Conflict reflects a hallucination-associated failure mode, can resolving it alter generation behavior? To answer this, we identify attention heads with high Conflict Strength as **Conflict Heads** and introduce a lightweight, training-free **Conflict-aware Intervention**. Instead of uniformly modifying an attention head, we independently regulate its I2T and T2T information flows during inference, allowing the two pathways to be adjusted according to their distinct roles in the conflict.

Experiments on LLaVA-v1.5-7B across POPE, MCQ-POPE, and CHAIR consistently improve over strong inference-time baselines, including AllPath. More revealingly, pathway-specific interventions expose a clear asymmetry between the two information flows. Attenuating I2T alone already reduces hallucination, while attenuating T2T alone degrades performance; jointly regulating both pathways yields the strongest overall results. These findings indicate that I2T is the primary intervention target, whereas T2T becomes beneficial only when coordinated with I2T. Together, our results shift the view of multimodal hallucination from *insufficient pathway utilization* to *conflicting pathway interaction*: what matters is not only how strongly visual and textual information influence generation, but whether they push the model in compatible directions.

2 Related Work

2.1 Hallucination in Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) have achieved strong performance across a wide range of vision-language tasks, but they remain susceptible to hallucination, producing responses that are fluent yet inconsistent with the visual input (Bai et al., 2025; Chen et al., 2026). Compared with hallucination in unimodal LLMs, MLLM hallucination is particularly characterized by failures in cross-modal grounding, where generated content deviates from or contradicts visual evidence.

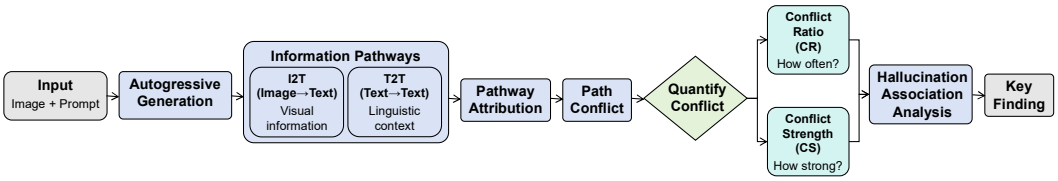

Figure 2: Path Conflict analysis pipeline: pathway attribution identifies I2T and T2T contributions, Conflict Ratio and Conflict Strength quantify their directional disagreement, and the resulting conflict patterns are analyzed for their association with hallucination.

Existing studies have attributed such failures to multiple factors, including training-data bias, visual information loss, insufficient cross-modal alignment, and diminishing utilization of visual tokens during autoregressive generation (Bai et al., 2025; Yin et al., 2024). A common interpretation is that hallucination emerges when visual evidence is under-utilized and generation becomes overly dominated by linguistic priors.

This view has motivated a broad range of mitigation strategies. At the training stage, instruction tuning and preference optimization are used to improve visual alignment and response faithfulness (Liu et al., 2023; Huang et al., 2024; Leng et al., 2023; Yu et al., 2024). At inference time, contrastive decoding and attention-based intervention attempt to reduce hallucination without additional parameter updates (Huang et al., 2024; Leng et al., 2023). In parallel, benchmarks such as CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023b) have been developed to systematically evaluate object-level hallucination. While these approaches substantially improve hallucination robustness, most focus on correcting model outputs or strengthening visual grounding, leaving open the question of how different internal information sources jointly shape hallucinated predictions.

2.2 Mechanistic Analysis of Multimodal Information Flow

Recent work has increasingly examined MLLM hallucination from an internal-mechanism perspective, moving from coarse model components toward finer-grained attention heads and information pathways. Yang et al. (Yang et al., 2025) use causal mediation analysis to compare the roles of different Transformer components and show that multi-head attention contributes more strongly to hallucinated token generation than MLP layers. They further identify hallucination-related attention heads and mitigate hallucination by reducing these heads' reliance on textual tokens. Tang et al. (Tang et al., 2025) analyze anomalous attention behaviors at both token and head levels, associating hallucination with insufficient visual attention, excessive dependence on previously generated text, and over-attention to system prompts. Their results further demonstrate that selectively intervening on internal attention patterns can alter hallucination behavior.

Going beyond individual components, Qian et al. (Qian et al., 2026) study hallucination from the perspective of information propagation and propose Intervene-All-Paths (AllPath). By formulating the Transformer as a causal structure, AllPath decomposes multimodal generation into several information pathways, including image-to-text and text-to-text propagation, and quantifies their individual contributions to model predictions. Their analysis shows that hallucination cannot always be attributed to a single dominant pathway, but instead reflects the joint effects of multiple information sources.

These studies establish an important progression from *where* hallucination-related signals occur to *which pathways* contribute to generation. However, they largely analyze modules, heads, or pathways independently. A remaining question is whether different information pathways provide mutually compatible signals when predicting the same token. In particular, pathway-wise contribution analysis does not reveal whether visual and linguistic pathways reinforce or oppose each other. Our work addresses this missing interaction by introducing **Path Conflict**, which explicitly characterizes directional inconsistency between Image-to-Text (I2T) and Text-to-Text (T2T) pathways. Rather than measuring pathway strength alone, we study whether heterogeneous information flows agree on the prediction, and further use this interaction signal to identify conflict-related heads and guide targeted inference-time intervention.

3 PATH CONFLICT: ANALYSIS AND EVIDENCE

Our introduction highlights a counter-intuitive observation: hallucinated generations can exhibit stronger visual-pathway influence rather than weaker visual grounding. This suggests that pathway strength alone is insufficient to explain hallucination. We therefore turn to the interaction between two information flows involved in autoregressive generation: Image-to-Text (I2T), which carries visual information to generated tokens, and Text-to-Text (T2T), which propagates previously generated linguistic context. Instead of asking only *how much* each pathway contributes, we ask whether the two pathways *agree on the direction of the prediction*.

3.1 FROM PATHWAY CONTRIBUTION TO PATH CONFLICT

Following AllPath (Qian et al., 2026), we quantify pathway contributions at the attention-head level. For layer l, head h, and generation step t, let $C^{\mathrm{I2T}}_{l,h,t}$ and $C^{\mathrm{T2T}}_{l,h,t}$ denote the contributions of the I2T and T2T pathways to the current token prediction, respectively. The magnitude reflects contribution strength, while the sign indicates whether the pathway promotes or suppresses the prediction.

We define **Path Conflict** as directional disagreement between the two pathways:

$$C^{\mathrm{I2T}}_{l,h,t} \cdot C^{\mathrm{T2T}}_{l,h,t} < 0. \tag{1}$$

A negative product indicates that visual and textual information push the same prediction in opposite directions. Unlike pathway-wise analysis, which considers each information flow independently, Path Conflict explicitly captures their interaction.

To quantify this phenomenon, we introduce two complementary measures. **Conflict Ratio (CR)** measures how frequently the two pathways disagree:

$$CR_{l,h} = \frac{N^{\mathrm{conflict}}_{l,h}}{N^{\mathrm{valid}}_{l,h}}, \tag{2}$$

where $N^{\mathrm{conflict}}_{l,h}$ denotes the number of generation positions satisfying Eq. 1, and $N^{\mathrm{valid}}_{l,h}$ is the number of valid positions.

Conflict frequency alone, however, does not capture how strongly the two pathways compete. We therefore define **Conflict Strength (CS)** as

$$CS_{l,h} = \frac{1}{N^{\mathrm{conflict}}_{l,h}} \sum_{t \in T^{\mathrm{conflict}}_{l,h}} \left|C^{\mathrm{I2T}}_{l,h,t}\right| \left|C^{\mathrm{T2T}}_{l,h,t}\right|, \tag{3}$$

where $T^{\mathrm{conflict}}_{l,h}$ denotes the set of conflicting generation positions. CR therefore characterizes *how often* a head exhibits directional disagreement, while CS measures *how strongly* the two pathways compete when such disagreement occurs.

3.2 PATH CONFLICT AND HALLUCINATION

We next examine whether Path Conflict is systematically associated with hallucinated generation. As shown in Fig. 3, conflicts are highly non-uniform across the model and concentrate in a small subset of attention heads. This suggests that directional disagreement between visual and textual information is not a diffuse property of the network, but is localized to specific internal components.

More importantly, Tab. 1 shows that hallucinated samples exhibit stronger Path Conflict than non-hallucinated ones. The difference is particularly pronounced in Conflict Strength, indicating that hallucination is more strongly associated with the magnitude of competing pathway influences than with conflict frequency alone.

To understand what drives this stronger conflict, we further decompose the I2T and T2T contributions within high-conflict heads. For a sample set $\mathcal{D}$, we compute the mean absolute contribution of pathway $P \in \{\mathrm{I2T}, \mathrm{T2T}\}$ as

$$\mathrm{Mean}|P|_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{s \in \mathcal{D}} \mathrm{mean}|P|_s, \tag{4}$$

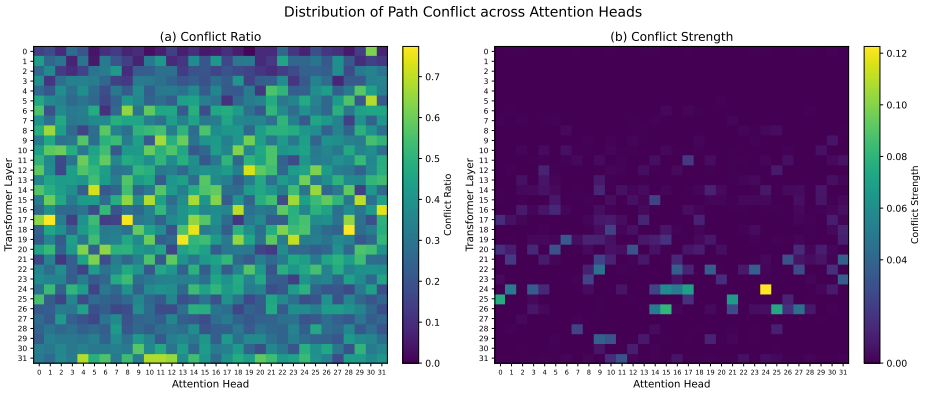

Figure 3: Path Conflict is highly non-uniform across attention heads, with strong conflicts concentrated in a small subset of heads that motivate targeted intervention.

Table 1: Path Conflict statistics for safe and hallucinated samples. Hallucinated samples consistently exhibit higher Conflict Ratio and substantially stronger Conflict Strength.

Sample Type	Conflict Ratio	Conflict Strength
Safe (POPE)	0.4076	0.2389
Hall (POPE)	0.4235	0.7330
Safe (CHAIR)	0.3479	0.6624
Hall (CHAIR)	0.3571	2.1305

Table 2: Hallucinated samples exhibit markedly amplified I2T contributions, while T2T contributions remain comparatively stable.

Sample	**POPE**		**CHAIR**	
	$\lvert\text{I2T}\rvert$	$\lvert\text{T2T}\rvert$	$\lvert\text{I2T}\rvert$	$\lvert\text{T2T}\rvert$
Safe	0.0476	0.0193	0.0408	0.0188
Hall	0.1327	0.0192	0.1534	0.0134

where $\mathcal{D} \in \{\text{Safe}, \text{Hallu}\}$.

As shown in Tab. 2, hallucinated samples exhibit a pronounced increase in I2T contribution, whereas the T2T contribution changes comparatively little. This result is important because it rules out a simple interpretation in which hallucination is caused by stronger linguistic dominance suppressing visual information. Instead, the visual pathway itself can become *more influential yet directionally misaligned* with the linguistic context.

This observation reveals a failure mode beyond insufficient visual grounding: hallucination can emerge from **misaligned visual influence**. In other words, the key issue is not merely whether the model uses visual information, but whether visual and textual pathways push the prediction in compatible directions.

These findings motivate the intervention in the next section. Rather than uniformly suppressing a single information source, we explicitly target high-conflict attention heads and independently regulate their I2T and T2T pathways to resolve conflicting information flow.

4 Conflict-aware Intervention

Section 3 shows that hallucination-associated conflicts are concentrated in a small subset of attention heads and are primarily accompanied by amplified I2T contributions. This motivates a targeted intervention: rather than uniformly modifying attention, we identify high-conflict heads and independently regulate their visual and textual information flows.

4.1 Conflict Head Selection

We rank all attention heads by the Conflict Strength defined in Eq. 3 and select the top-K heads as **Conflict Heads**:

$$\mathcal{H}_{\text{conflict}} = \text{TopK}_{(l,h)\in\mathcal{H}}\, CS_{l,h}, \tag{5}$$

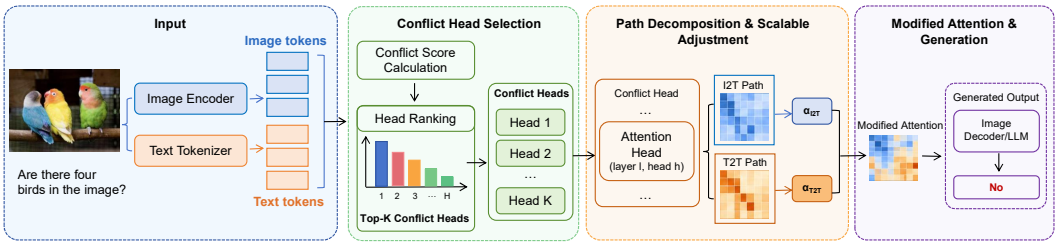

Figure 4: Conflict-aware Intervention resolves Path Conflict through head-level localization and pathway-specific regulation of I2T and T2T information flows.

where $\mathcal{H}$ denotes the set of all attention heads and K controls the intervention budget. We use Top-K selection rather than a fixed threshold because the absolute scale of conflict statistics may vary across datasets and generation settings, while Top-K provides a consistent number of intervention targets.

4.2 Path-aware Attention Intervention

For each Conflict Head, we separately regulate the I2T and T2T attention flows. Let $A_{l,h,q,k}$ denote the original attention weight from query token q to key token k, and let $\mathcal{V}$ and $\mathcal{T}$ denote the visual and text token sets, respectively. We introduce two scaling factors, α_{I2T} and α_{T2T}, and define

$$\hat{A}_{l,h,q,k} = \begin{cases} \alpha_{\text{I2T}} A_{l,h,q,k}, & q \in \mathcal{T},\ k \in \mathcal{V}, \\ \alpha_{\text{T2T}} A_{l,h,q,k}, & q, k \in \mathcal{T}, \\ A_{l,h,q,k}, & \text{otherwise}, \end{cases} \qquad (l,h) \in \mathcal{H}_{\text{conflict}}. \tag{6}$$

The first case regulates the I2T pathway, while the second adjusts the T2T pathway. Since Section 3 shows that hallucinated samples exhibit substantially amplified I2T contributions, I2T serves as the primary intervention target. T2T is retained as a complementary pathway whose role is evaluated independently rather than being uniformly suppressed together with I2T.

After path-specific scaling, we renormalize the attention distribution:

$$\widetilde{A}_{l,h,q,k} = \frac{\hat{A}_{l,h,q,k}}{\sum_{k'} \hat{A}_{l,h,q,k'}}. \tag{7}$$

When $\alpha_{\text{I2T}} < 1$ or $\alpha_{\text{T2T}} < 1$, the corresponding pathway is attenuated, while a coefficient of 1 leaves it unchanged. Attention heads outside $\mathcal{H}_{\text{conflict}}$ remain untouched.

This design intervenes at both the *head* and *pathway* levels: Conflict Head Selection localizes where directional disagreement is concentrated, while path-specific scaling determines which information flow to regulate. Unlike uniformly suppressing an entire head, our intervention preserves non-targeted information flow within the same component. It requires no additional training or parameter updates and can be directly applied during autoregressive inference.

5 Experiments

5.1 Experimental Setup

Model and benchmarks. We evaluate Conflict-aware Intervention on LLaVA-v1.5-7B using four benchmarks: POPE (Li et al., 2023b), MCQ-POPE (Qian et al., 2026), CHAIR (Rohrbach et al., 2018), and MME (Fu et al., 2025). POPE evaluates binary object-presence verification under Random, Popular, and Adversarial settings; MCQ-POPE extends this setting to four-choice questions; CHAIR evaluates object hallucination in open-ended image captioning. MME provides a broader evaluation of multimodal perception and cognition abilities, covering multiple perception categories and cognitive tasks beyond hallucination-specific evaluation. We report Accuracy and F1 for POPE

and MCQ-POPE, CHAIRs/CHAIRi for CHAIR, and the overall perception and cognition scores for MME, where lower CHAIR scores indicate less hallucination.

Implementation. We build on the pathway attribution framework of AllPath (Qian et al., 2026). Attention heads are ranked by Conflict Strength, and the top-K heads are selected as Conflict Heads. During inference, their I2T and T2T attention flows are independently scaled by α_{I2T} and α_{T2T}, respectively, following Eq. 6. No additional training or parameter updates are required.

Table 3: Conflict-aware Intervention consistently outperforms pathway-wise baselines, including AllPath, across three hallucination benchmarks.

Method	POPE						MCQ-POPE						CHAIR	
	Random		Popular		Adversarial		Random		Popular		Adversarial			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	$C_s \downarrow$	$C_i \downarrow$
Vanilla	85.1	83.7	83.7	82.4	80.9	80.0	72.8	72.9	68.0	68.0	64.4	64.5	52.2	14.6
VCD (Leng et al., 2023)	86.3	85.2	84.5	83.5	81.4	80.9	78.2	78.2	72.2	72.2	67.9	67.9	58.2	16.1
ICD (Wang et al., 2024)	84.4	82.2	83.8	81.6	81.5	79.5	69.1	69.1	67.2	67.2	64.7	64.6	51.4	14.7
PAI (Liu et al., 2024b)	86.4	85.0	84.9	83.6	82.5	81.4	78.0	78.0	71.2	71.2	68.0	67.9	28.8	7.9
AD-HH (Yang et al., 2025)	85.0	83.6	83.7	82.4	80.9	79.9	78.5	78.6	71.1	71.1	68.1	68.1	33.2	7.5
AllPath (Qian et al., 2026)	87.2	86.0	86.0	84.9	82.8	82.1	80.5	80.5	74.0	74.0	69.7	69.7	26.6	7.2
Ours	**87.8**	**87.1**	**86.6**	**86.0**	**83.2**	**83.1**	**82.0**	**82.0**	**76.8**	**76.8**	**70.6**	**70.6**	**25.4**	**5.9**
Δ	+0.6	+1.1	+0.6	+1.1	+0.4	+1.0	+1.5	+1.5	+2.8	+2.8	+0.9	+0.9	-1.2	-1.3

5.2 Main Results

Table 3 reports the overall results. Conflict-aware Intervention consistently improves over strong inference-time baselines, including AllPath, across all three benchmarks. On POPE, it improves F1 by 1.0–1.1 points across the three settings. On MCQ-POPE, the gains are larger, reaching 2.8 points in both Accuracy and F1 on the Popular split. On CHAIR, it further reduces CHAIRs and CHAIRi by 1.2 and 1.3 points, respectively.

These gains are consistent across binary verification, multiple-choice reasoning, and open-ended captioning, suggesting that the benefit is not tied to a specific evaluation format. More importantly, the improvement over AllPath is particularly informative: while AllPath intervenes based on the contribution of individual pathways, our method additionally exploits their *directional interaction*. This indicates that whether I2T and T2T agree on a prediction provides complementary information beyond pathway strength alone.

5.3 Generalization on MME

To further examine whether Conflict-aware Intervention introduces unintended degradation on general multimodal capabilities, we evaluate our method on the MME benchmark. As shown in Table 4, our method achieves comparable performance with existing hallucination mitigation approaches across different perception categories.

Table 4: Generalization comparison with existing methods on the MME benchmark.

Method	Object-Level		Attribute-Level		Total↑
	Existence	Count	Position	Color	
Vanilla	$180.0_{(\pm 8.66)}$	$113.9_{(\pm 10.05)}$	$116.7_{(\pm 15.90)}$	$129.4_{(\pm 18.28)}$	$540.0_{(\pm 39.69)}$
VCD Leng et al. (2023)	$177.8_{(\pm 2.55)}$	$122.8_{(\pm 20.84)}$	$122.2_{(\pm 1.92)}$	$141.7_{(\pm 2.89)}$	$564.4_{(\pm 20.84)}$
ICD Wang et al. (2024)	$185.0_{(\pm 5.00)}$	$121.7_{(\pm 2.89)}$	$118.3_{(\pm 6.01)}$	$151.7_{(\pm 7.64)}$	$576.7_{(\pm 11.67)}$
PAI Liu et al. (2024b)	$185.0_{(\pm 5.00)}$	$122.8_{(\pm 14.94)}$	$114.4_{(\pm 4.19)}$	$144.4_{(\pm 5.05)}$	$566.7_{(\pm 15.90)}$
AD-HH Yang et al. (2025)	$179.4_{(\pm 5.09)}$	$117.8_{(\pm 12.51)}$	$116.1_{(\pm 13.37)}$	$132.2_{(\pm 8.39)}$	$545.5_{(\pm 29.80)}$
AllPath Qian et al. (2026)	$188.3_{(\pm 2.89)}$	$126.1_{(\pm 0.96)}$	$132.2_{(\pm 8.39)}$	$146.8_{(\pm 1.09)}$	$593.4_{(\pm 12.33)}$
Ours	$\mathbf{190.0}_{(\pm 5.00)}$	$\mathbf{128.3}_{(\pm 16.63)}$	$\mathbf{119.4}_{(\pm 15.57)}$	$\mathbf{151.1}_{(\pm 6.10)}$	$\mathbf{588.9}_{(\pm 28.87)}$

Despite being specifically designed for hallucination reduction, Conflict-aware Intervention preserves the general perception ability of the underlying MLLM. The results indicate that selectively regulating conflicting information pathways does not impair general multimodal understanding.

5.4 Ablation Studies

5.4.1 Which Path Should Be Intervened?

We next isolate the roles of I2T and T2T by keeping the Conflict Heads fixed and varying which pathway is regulated. As shown in Table 5, the three intervention settings exhibit markedly different behaviors.

Table 5: Pathway-specific intervention reveals asymmetric roles of I2T and T2T: I2T is the primary target, while joint regulation performs best.

I2T Scale	T2T Scale	POPE						MCQ-POPE						CHAIR	
		Random		Popular		Adversarial		Random		Popular		Adversarial			
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	$C_s \downarrow$	$C_i \downarrow$
0.3	0.2	87.8	87.1	86.6	86.0	83.2	83.1	82.00	82.0	76.8	76.8	70.6	70.6	25.4	5.9
0.3	1.0	87.6	86.9	86.4	85.8	82.8	82.7	79.5	79.5	74.8	74.8	69.0	69.0	26	6.8
1.0	0.2	86.7	85.3	85.5	84.2	82.6	81.6	78.6	78.7	72.0	72.1	67.7	67.7	27	7.5
1.0	1.0	87.2	86.0	85.9	84.9	82.8	82.1	78.1	78.1	73.8	73.9	68.5	68.5	26.9	7.5

I2T is the primary intervention target. Attenuating I2T alone consistently improves over the no-intervention baseline, confirming the observation in Section 3 that hallucination-associated conflicts are primarily accompanied by amplified visual-pathway influence.

T2T is not an effective standalone target. In contrast, suppressing T2T alone degrades performance across benchmarks. This suggests that the textual pathway still carries useful contextual information and that hallucination cannot be reduced by simply weakening linguistic influence.

Joint regulation works best. Jointly regulating I2T and T2T achieves the strongest overall performance. Interestingly, T2T intervention becomes beneficial only when coupled with I2T attenuation. This asymmetric behavior supports our central hypothesis: the key failure is not excessive reliance on a single pathway, but the way visual and linguistic pathways interact when they provide conflicting signals.

Overall, these results reinforce the mechanism identified in Section 3. Hallucination-related conflict is primarily driven by misaligned I2T influence, while T2T acts as a complementary pathway whose intervention is effective only in coordination with I2T.

5.4.2 How to Select Intervention Strength?

We further investigate the influence of different intervention strengths for I2T and T2T pathways. As shown in Table 6, moderate attenuation of both pathways achieves the best performance, with $\alpha_{\text{I2T}} = 0.3$ and $\alpha_{\text{T2T}} = 0.2$.

The results further support our finding that I2T pathway conflict is the primary factor contributing to hallucination. Moderate I2T attenuation achieves the best performance, while insufficient suppression leaves excessive visual conflict signals and overly strong suppression removes useful visual evidence. In contrast, T2T scaling has a smaller effect, indicating that linguistic context should be preserved rather than aggressively suppressed. These results validate our asymmetric pathway regulation strategy, which selectively mitigates conflicting visual signals while retaining beneficial textual information.

5.4.3 Effect of Conflict Head Selection

Our method relies on identifying attention heads where I2T and T2T pathways exhibit strong directional conflicts. To verify the effectiveness of this selection strategy, we compare three head selection schemes: ranking by Conflict Strength (CS), ranking by Conflict Ratio (CR), and random selection.

Table 6: Effect of different I2T and T2T intervention strengths on hallucination mitigation.

I2T Scale	T2T Scale	POPE						MCQ-POPE					
		Random		Popular		Adversarial		Random		Popular		Adversarial	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
0.1	0.2	87.0	86.6	85.9	85.6	82.2	82.5	82.4	82.4	75.4	75.5	69.8	69.8
0.3	0.2	87.8	87.1	86.6	86.0	83.2	83.1	82.0	82.0	76.8	76.8	70.6	70.6
0.5	0.2	87.6	86.7	86.3	85.5	83.0	82.6	81.5	81.5	75.7	75.7	70.3	70.3
0.7	0.2	87.4	86.3	86.1	85.1	82.6	82.1	81.0	81.0	75.3	75.4	69.8	69.7
0.3	**0.4**	87.6	86.9	86.4	85.7	83.0	82.8	82.7	82.7	76.4	76.4	70.0	70.0
0.3	**0.6**	87.7	87.0	86.4	85.8	83.0	82.8	81.3	81.3	76.1	76.1	70.0	70.0

As shown in Table 7, selecting heads according to Conflict Strength consistently achieves the best performance across POPE, MCQ-POPE, and CHAIR. In contrast, selecting heads based only on Conflict Ratio leads to inferior results, suggesting that the frequency of conflicts alone is insufficient to identify influential hallucination-related heads. Random selection performs worse, further demonstrating that the observed improvements originate from targeted intervention on conflict-aware heads rather than generic attention modification.

Table 7: Ablation study of different Conflict Head selection strategies.

Head Selection	POPE						MCQ-POPE						CHAIR	
	Random		Popular		Adversarial		Random		Popular		Adversarial			
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	$C_s \downarrow$	$C_i \downarrow$
Conflict Strength	87.8	87.1	86.6	86.0	83.2	83.1	82.00	82.0	76.8	76.8	70.6	70.6	25.4	5.9
Conflict Ratio	87.1	86.8	86.1	85.5	82.5	82.2	81.5	81.6	75.8	75.8	70.2	70.2	26.2	6.2
Random	86.1	84.9	85.9	85.3	82.9	82.4	81.0	81.0	75.1	75.1	69.7	69.7	25.9	7.4

These results support our analysis in Section 3 that hallucination is associated with specific attention heads exhibiting strong pathway disagreement, and validating Conflict Strength as an effective criterion for locating such heads.

6 LIMITATIONS

Our study has several limitations. First, the current experiments are conducted on LLaVA-v1.5-7B and primarily focus on object-level hallucination benchmarks. Although the observed Path Conflict pattern is consistent across POPE, MCQ-POPE, and CHAIR, its generality across different MLLM architectures and broader hallucination types remains to be further investigated. Second, we focus on the interaction between I2T and T2T pathways, while multimodal generation may involve other information flows whose conflicts are not explicitly modeled. Finally, our intervention uses fixed pathway-wise scaling coefficients within selected Conflict Heads. More adaptive strategies that condition intervention strength on the current token, sample, or conflict state may further improve the effectiveness and generality of conflict-aware mitigation.

7 CONCLUSION

We identify **Path Conflict** as a distinct failure mode of MLLM hallucination, where visual and textual pathways push token predictions in opposite directions. Hallucinated generations exhibit stronger conflict, primarily driven by amplified I2T contributions, suggesting that hallucination can arise from *misaligned visual influence* rather than insufficient visual grounding. Based on this insight, we propose a lightweight, training-free **Conflict-aware Intervention** that selectively regulates I2T and T2T flows within high-conflict heads, consistently improving hallucination robustness across multiple benchmarks. Our results shift the focus from pathway strength to pathway agreement in understanding multimodal hallucination.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation and improving the readability of the manuscript. We also used generative AI tools for literature search and identifying relevant references during the research process. All AI-assisted outputs were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including all text, claims, experimental results, and analyses.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Kazi Hasan Ibn Arif, Sajib Acharjee Dip, Khizar Hussain, Lang Zhang, and Chris Thomas. Paint: Paying attention to informed tokens to mitigate hallucination in large vision-language model, 2025. URL https://arxiv.org/abs/2501.12206.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey, 2025. URL https://arxiv.org/abs/2404.18930.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023. URL https://arxiv.org/abs/2209.06794.

Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection, 2026. URL https://arxiv.org/abs/2507.19024.

Wenliang Dai, Junnan Li, DONGXU LI, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 49250–49267. Curran Associates, Inc., 2023. doi: 10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/9a6a435e75419a836fe47ab6793623e6-Paper-Conference.pdf.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models, 2025. URL https://arxiv.org/abs/2306.13394.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models, 2024. URL https://arxiv.org/abs/2308.06394.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation, 2024. URL https://arxiv.org/abs/2311.17911.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, March 2023. ISSN 1557-7341. doi: 10.1145/3571730. URL http://dx.doi.org/10.1145/3571730.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding, 2023. URL https://arxiv.org/abs/2311.16922.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023a. URL https://arxiv.org/abs/2301.12597.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models, 2023b. URL https://arxiv.org/abs/2305.10355.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024a. URL https://arxiv.org/abs/2402.00253.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms, 2024b. URL https://arxiv.org/abs/2407.21771.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models, 2024. URL https://arxiv.org/abs/2310.05338.

Jiaye Qian, Ge Zheng, Yuchen Zhu, and Sibei Yang. Intervene-all-paths: Unified mitigation of lvlm hallucinations across alignment formats, 2026. URL https://arxiv.org/abs/2511.17254.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation, 2025. URL https://arxiv.org/abs/2506.12609.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding, 2024. URL https://arxiv.org/abs/2403.18715.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/8001c3568152d134d821cd46d4d84768-Paper-Conference.pdf.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL http://dx.doi.org/10.1093/nsr/nwae403.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback, 2024. URL https://arxiv.org/abs/2312.00849.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms, 2024. URL https://arxiv.org/abs/2411.09968.

Deyao Zhu, jun chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/50623630a2372839c078474efa6c0cb8-Paper-Conference.pdf.