Treble Counterfactual VLMs: A Causal Approach to Hallucination

Anonymous ACL submission

Abstract

Vision-Language Models (VLMs) excel at tasks such as image captioning and visual question answering but frequently produce hallucinated outputs that deviate from the actual visual input or prompt. While prior work links hallucination to biases in data or representation, their causal origins remain unclear. We propose a causal framework to analyze and mitigate hallucination in VLMs. Our key hypothesis is that hallucinations arise from unintended direct influences of the vision or text modality that bypass the intended multi-modal fusion. To examine this, we construct a causal graph of the VLM and use counterfactual analysis to estimate the Natural Direct Effect (NDE) of each modality and their interaction. By systematically identifying and suppressing these direct effects, we encourage outputs that are more faithfully grounded in true cross-modal reasoning. Our approach consists of three steps: (1) designing structural causal graphs to distinguish correct fusion pathways from spurious modality shortcuts, (2) estimating modality-specific and cross-modal NDE using perturbed image representations, hallucinated text embeddings, and degraded visual inputs, and (3) implementing a test-time intervention module to dynamically adjust the model's dependence on each modality. Experimental results demonstrate that our method significantly reduces hallucination while preserving task performance, providing a robust and interpretable framework for improving VLM reliability.

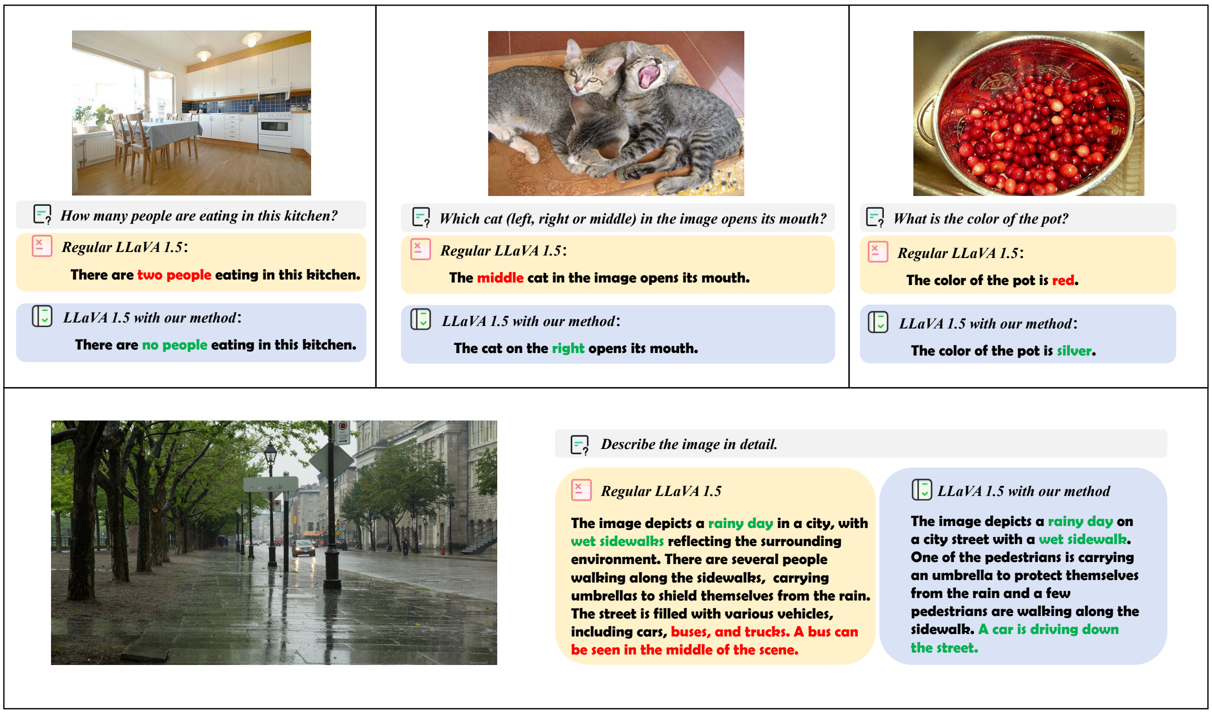
1 Introduction

Vision-Language Models (VLMs) have made significant progress in multi-modal tasks such as image captioning (Mokady et al., 2021), visual question answering, and visual reasoning (Li et al., 2023a; Alayrac et al., 2022; Liu et al., 2023b; Radford et al., 2021). By integrating visual and textual inputs, VLMs generate descriptive outputs that enhance machine understanding of multi-modal contexts (Chowdhery et al., 2023). They typically comprise a vision encoder for extracting image features and a language model for generating outputs conditioned on both modalities. Advances in large-scale pre-training and transformer architectures have further improved their generalization (Zhai et al., 2022), making VLMs key to AI applications.

Hallucination in VLMs. Despite strong performance, VLMs are prone to hallucination (Ji et al., 2023): producing outputs *inconsistent* with the visual input or textual prompt, often introducing incorrect or fabricated information. This reduces reliability in high-stakes domains such as medical imaging (Goddard, 2023), autonomous driving (Chen et al., 2024a), and surveillance (Zhao et al., 2020). While several factors contribute to hallucination, e.g., modality misalignment and learned biases, its root causes remain understudied, which needs systematic investigation and mitigation.

Existing Approaches. (see Appx. A for more details) Prior work has explored various ways to understand and reduce hallucination in VLMs (Ji et al., 2023; Zhou et al., 2023; Rohrbach et al., 2018; Yang et al., 2025), with different explanations and mitigations. Some studies link hallucinations to biases in training data (Zhou et al., 2023), where models latch onto spurious correlations rather than truly learning visual-text relationships. Others point to overreliance on language priors (Yang et al., 2025; Rohrbach et al., 2018), leading to text-focused outputs that overlook visual context. Additional research highlights biased feature learning (Kayhan et al., 2021; Chen et al., 2024b), which can cause certain patterns to dominate the representations and distort multi-modal reasoning. However, most approaches focus on *statistical* or *empirical* analyses and often do not differentiate VLMs from large language models (LLMs), overlooking the distinct challenges inherent in multi-modal architectures in VLMs.

Our Causal Perspective. In this work, we propose



How many people are eating in this kitchen?

Regular LLaVA 1.5:

There are two people eating in this kitchen.

LLaVA 1.5 with our method:

There are no people eating in this kitchen.

Which cat (left, right or middle) in the image opens its mouth?

Regular LLaVA 1.5:

The middle cat in the image opens its mouth.

LLaVA 1.5 with our method:

The cat on the right opens its mouth.

What is the color of the pot?

Regular LLaVA 1.5:

The color of the pot is red.

LLaVA 1.5 with our method:

The color of the pot is silver.

Describe the image in detail.

Regular LLaVA 1.5

The image depicts a rainy day in a city, with wet sidewalks reflecting the surrounding environment. There are several people walking along the sidewalks, carrying umbrellas to shield themselves from the rain. The street is filled with various vehicles, including cars, buses, and trucks. A bus can be seen in the middle of the scene.

LLaVA 1.5 with our method

The image depicts a rainy day on a city street with a wet sidewalk. One of the pedestrians is carrying an umbrella to protect themselves from the rain and a few pedestrians are walking along the sidewalk. A car is driving down the street.

Figure 1: Case study illustrating the impact of our method on VLM hallucination. The figure compares outputs from the original model and our enhanced approach, highlighting reductions in hallucinated content and improved alignment with the visual context. Our method effectively mitigates incorrect descriptions by refining modality interactions, leading to more accurate and reliable multi-modal reasoning.

a causal framework to analyze and mitigate hallucination in VLMs. We construct a *causal graph* (Neuberg, 2003) for VLMs, hypothesizing that hallucination may arise due to unintended direct influences from either the vision or text modality, bypassing the intended multi-modal fusion process (Kiros et al., 2014). Specifically, each modality can have independent direct effects on the output, leading to inconsistencies between generated answers and their intended multi-modal context. Based on this premise, we employ counterfactual analysis (Lewis, 2013) to estimate the Natural Direct Effect (NDE) (Robins and Greenland, 1992) of each modality and systematically remove these extraneous influences. By doing so, we ensure that responses are primarily driven by joint vision-text reasoning, thereby reducing hallucination and improving reliability.

This can be described as a three-step methodology. *First*, we design structural causal graphs (Neuberg, 2003) to capture the relationships between vision, text, and outputs, distinguishing correct fusion pathways from spurious shortcuts. *Second*, we systematically estimate the NDE of vision, text, and their cross-modal interaction. For vision, we generate perturbed images by applying multiple random masks, then measure how these perturbations shift latent representations. For text, we create "hallucinated" captions via a language model (Zhao et al., 2023) and compare their embeddings with those of the original input. *Finally*, we develop a dynamic test-time intervention module that adjusts the model's reliance on each modality, effectively reducing hallucination while preserving overall performance. Our method requires only 50 randomly selected samples to estimate intervention directions. These directions generalize well across different benchmarks and VLM architectures, indicating that the modality-specific biases we correct are stable and transferable. This efficient estimation enables broad applicability without retraining or model-specific tuning.

Our key contributions are as follows:

- **Causal Analysis of Hallucination.** We present a structured causal framework for VLMs, showing the unintended direct effects from both vision and text that bypass proper multi-modal fusion. By conducting rigorous counterfactual analysis, our approach uncovers how each modality's direct influence underlies hallucinations.

- **Test-time Hallucination Reduction.** We develop a lightweight method to mitigates hallucination in VLMs by proper multi-modal fusion and reasoning, without requiring model retraining or additional parameters.

- **Effectiveness.** Our approach consistently outperforms existing methods on two VLMs across two diverse benchmarks. For instance, it improves the F1 score of LLaVA 1.5 by over 10% on the POPE benchmark. Notably, our method remains robust across random, popular, and adversarial scenarios, with broad applicability and resilience.

- **Accessibility and Reproducibility.** Our intervention is model-agnostic, incurs no training or inference cost, and is fully test-time deployable. We release all code and data to support future research: `https://anonymous.4open.science/r/Treble-Counterfactual-VLMs-16B4`.

2 Related Works

For a full discussion, please refer to Appx. A.
Hallucination in Vision-Language Models. VLMs combine visual encoders with LLMs to enable multimodal reasoning (Dai et al., 2023; Liu et al., 2023b), but often suffer from hallucinations—outputs inconsistent with visual input (Bang et al., 2023; Huang et al., 2021). This includes inventing non-existent objects or relying on language priors. Prior work attributes hallucinations to biases such as object co-occurrence or spatial misalignment (Li et al., 2023a; Zhou et al., 2023), and proposes mitigation via retraining or post-hoc correction (Yin et al., 2024; Yue et al., 2024).
Causal Perspectives. Causality offers tools to separate genuine multimodal reasoning from spurious modality dominance (Li et al., 2022; Wang and Vasconcelos, 2020). In VLMs, causal graphs and counterfactual analysis have been used to expose and reduce hallucinations by tracing modality-specific effects (Li et al., 2023b, 2024a). Our work builds on this foundation to provide a lightweight, test-time causal intervention.

3 Preliminaries

Related works are shown in Appx. A. In this section, we propose a series of structural causal graphs (SCGs) (§3.1) for different scenarios to illustrate the superficial correlations between visual inputs, language inputs, and generated answers (§3.2). We then analyze the hallucination problem in VLMs and provide a causal interpretation to explain its underlying causes (§3.3).

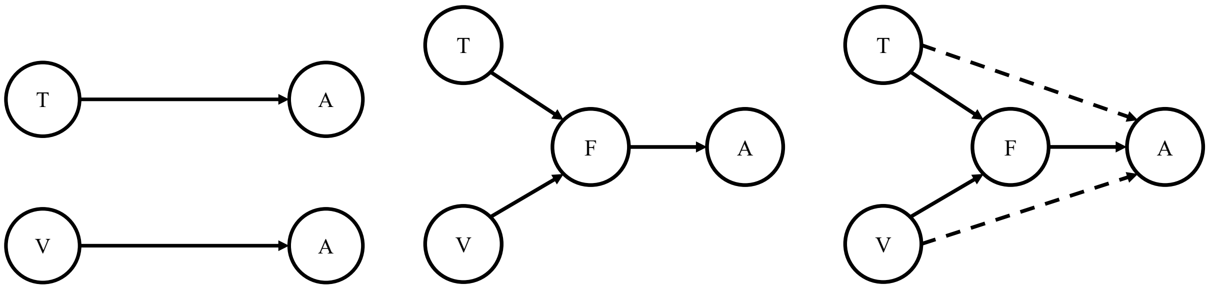
3.1 Structural Causal Graph

The SCGs for different scenarios are illustrated in Fig. 2. The effects of visual input V and textual input T on the output A can be categorized into two types: single-modal impact (Traditional computer vision tasks or Large Language Models) and multi-modal impact (Vision-Language Models). As shown in Fig. 2a, the single-modal impact captures the direct influence of V or T on A through $V \rightarrow A$ or $T \rightarrow A$. In contrast, the multi-modal impact represents the indirect effect of V and T on A via the multi-modal fused knowledge F, formulated as $(V, T) \rightarrow F \rightarrow A$, as shown in Fig. 2b. The underlying rationale behind the SCG is explained as follows:

- $T \rightarrow A$: This represents the data flow in traditional Large Language Models (LLMs), where natural language inputs (typically comprising instructions and data) are processed by the LLM to generate the corresponding output A.
- $V \rightarrow A$: This corresponds to traditional computer vision tasks, such as image captioning, where images are provided as input, and the output A is generated solely based on visual information without language-based context.
- $(V, T) \rightarrow F \rightarrow A$: This illustrates the mechanism of modern Vision-Language Models. The visual input V is first processed by a vision backbone (e.g., a convolutional neural network or a transformer-based vision encoder) to extract high-level visual features. These visual features are then projected into a shared embedding space compatible with the LLM. Simultaneously, the textual input T is encoded by the LLM. The multi-modal fusion module combines the visual and textual representations to form the fused knowledge F. Finally, the LLM leverages this fused knowledge F to generate the answer A, integrating both vision and language modalities for coherent and context-aware outputs.

3.2 Potential Biased Independent Influence

Although the optimal Vision-Language Model is expected to generate answers solely based on the combined vision and text input pairs, in practice, vision and text inputs may still exert direct and independent influences on the output A (Kiros et al., 2014). As illustrated in Fig. 2c, these unintended direct influences are highlighted by dashed arrows, indicating potential shortcut paths that bypass the multi-modal fusion process. Such direct influences can lead to the hallucination problem, where the



(a) Causal graph for traditional single-modal model.

(b) Causal graph for Vision-Language Model.

(c) Causal graph for biased Vision-Language Model.

Figure 2: Causal graphs for single-modal models and Vision-Language Models (VLMs) are shown. An optimal VLM generates answers conditioned on both vision and text input pairs. However, vision and text inputs may individually exert a direct influence on the output. This direct influence can lead to the hallucination problem in VLMs, where the generated answers are inconsistent with the provided visual or textual context. T: Text input. V: Vision input. F: Fusion. A: Answer.

generated answer A does not align with the provided visual context or textual input.

- $T \rightarrow A$: The textual input T may directly influence the output A without considering the visual information. For instance, the model might rely heavily on language priors or contextual cues from the text alone, resulting in answers that ignore relevant visual details. This direct influence can lead to hallucinated responses that appear semantically plausible based on the text but remain inconsistent with the actual visual content.
- $V \rightarrow A$: Similarly, the visual input V may directly affect the output A without proper alignment with the textual input. In this scenario, the model might over-rely on visual patterns or features, producing answers that are disconnected from the given textual instructions or questions. This form of direct influence also contributes to hallucinations, where the output appears visually grounded but fails to reflect the intended textual semantics.

These dashed causal paths emphasize the inherent challenge in VLMs: ensuring that the answer A is truly conditioned on the coherent fusion of both V and T, rather than being dominated by a single modality. Addressing these unintended direct influences is essential for mitigating hallucination problems and improving the overall reliability and consistency of VLMs.

3.3 Causal Perspective on VLM Hallucination

From a causal perspective, the hallucination problem in VLMs arises when the model over-relies on a single modality, leading to outputs that are misaligned with the intended multi-modal context. Specifically, unintended direct influences from either the vision or text modality, or their interaction, can dominate the output generation process, causing hallucinated responses. To systematically examine and mitigate these biases, we focus on the *Natural Direct Effect (NDE)* as a means to quantify the direct contributions of each modality and their interaction.

Definition 1 (Causal Notations)**.** *Causal notations are used to translate causal assumptions from structural causal graphs into formal mathematical expressions, allowing precise quantification of modality influences on model outputs. Formally, given the causal graph illustrated in Fig. 2c, the answer A is influenced by three paths: $T \rightarrow A$, $V \rightarrow A$, and $F \rightarrow A$. The corresponding causal notation is as follows:*

$$A_{T,V} = A(t, v, F(t, v)), \quad (1)$$

where t and v are text and visual inputs, and $F(\cdot)$ denotes the multi-modal fusion process.

Definition 2 (Natural Direct Effects (NDE))**.** *The Natural Direct Effect (NDE) measures the direct impact of a modality on the output A while holding the multi-modal fusion process consistent. We consider three types of NDEs to capture both the individual and interactive effects of the vision and text modalities:*

1) Vision Direct Effect ($\mathbf{NDE_V}$): The direct influence of the vision modality is assessed by altering the vision input while keeping the textual input fixed. Formally:

$$\text{NDE}_V = Y(t, v, F(t, v)) - Y(t, v_*, F(t, v_*)), \quad (2)$$

where v denotes the original vision input and v_* represents the treated vision input. This formulation captures how much the vision modality alone contributes to the output, independent of multimodal fusion consistency.

2) Text Direct Effect ($\mathbf{NDE_T}$): The direct influence of the text modality is measured by modifying the textual input while keeping the visual input constant:

$$\mathrm{NDE}_T = Y(t, v, F(t, v)) - Y(t_*, v, F(t_*, v)), \quad (3)$$

where t is the original text input and t_* represents the treated text input. This equation reflects how text alone influences the output, independent of visual grounding.

3) Cross-Modality Direct Effect ($\mathbf{NDE_{V,T}}$): While the vision modality treatment assesses the direct influence of vision by altering visual inputs, it does not capture how vision complements textual information in multi-modal reasoning. In practice, vision often provides contextual cues that enhance text interpretation. Thus, it is essential to evaluate how vision interacts with text to influence the output.

To this end, we propose the *Cross-Modality Direct Effect ($NDE_{V,T}$)*, which quantifies the complementary role of vision when combined with text. Unlike vision treatment, which isolates vision's standalone contribution, this analysis evaluates scenarios where textual input is paired with a partially informative image versus a non-informative one. The formulation is:

$$\mathrm{NDE}_{V,T} = Y(t, v_*, F(t, v_*)) - Y(t, v_{\text{null}}, F(t, v_{\text{null}})), \quad (4)$$

where v_{null} denotes a non-informative visual input. A high $\mathrm{NDE}(V, T)$ indicates meaningful visual-textual complementarity, while a low or negative value suggests that vision introduces noise, potentially leading to hallucinations.

By focusing on these direct effects, our causal analysis framework provides a clear diagnostic approach to understanding and mitigating hallucination in VLMs. This framework highlights the necessity of balanced multi-modal fusion, where each modality contributes appropriately to the final prediction without dominating the reasoning process.

4 Methodology

Building on prior work in editing vision-language model intermediate representations (Liu et al., 2024; Jiang et al., 2024), we quantify the *Natural Direct Effects (NDEs)* of different modalities by analyzing representation shifts before and after applying modality-specific perturbations. This allows us to analyze separately the contributions of vision and text, along with their interaction, to the final model output.

Measuring $\mathbf{NDE_V}$. To measure the vision modality's direct effect, we introduce perturbations to the visual input and assess impacts on representations.

Given an image input I, we extract its vision representation $V_{i,k}^{I}$ from the i-th layer at the k-th visual token. We then apply m different random masks, C_j for $j \in \{1, \ldots, m\}$, to corrupt the image, producing masked versions $M_j(I)$. The vision encoder processes each perturbed input $M_j(I)$, yielding the corresponding representations $V_{i,k}^{M_j(I)}$. To estimate the perturbed vision representation, we take the avg. of these masked representations $\bar{V}_{i,k}^{I}$.

The direct effect of the vision modality for the image I is then quantified as the difference between the original and perturbed representations:

$$D_{i,k}^{I} = \bar{V}_{i,k}^{I} - V_{i,k}^{I}. \quad (5)$$

To obtain a global-level estimate of NDE_V (as opposed to the instance-level effect $D_{i,k}^{I}$), we sample N images and compute their respective direct effects, systematically stacking them into a structured matrix:

$$[D_{i,k}^{I_1}, D_{i,k}^{I_2}, ..., D_{i,k}^{I_N}]. \quad (6)$$

Following Liu et al. (2024), we perform PCA on this matrix and use the first principal direction as the global-level estimate of NDE_V.

Measuring $\mathbf{NDE_T}$. To measure the direct effect of the text modality, we introduce controlled textual hallucinations and analyze their influence on representations.

We randomly sample N image captions C_N and generate their hallucinated counterparts C_N^h using a GPT model. For each caption, we extract the last-token representation from the i-th layer, denoted as $T_i^{C_N}$ for the original text and $T_i^{C_N^h}$ for the hallucinated version. The direct effect of text modality can be computed as:

$$D_i^T = T_i^{C_N^h} - T_i^{C_N}. \quad (7)$$

To estimate global-level NDE_T, we stack the text direct effect vectors for all sampled captions into a matrix and apply PCA, obtaining the first principal direction as the final measure of NDE_T.

Settings	Method	LLaVA 1.5				InstructBlip			
		Accuracy	Precision	Recall	F1 score	Accuracy	Precision	Recall	F1 score
Random	Regular	83.49	88.83	76.70	82.34	80.42	78.93	83.21	81.01
	VCD	86.84	87.15	86.68	86.91	84.10	84.21	85.36	84.78
	Opera	87.53	**94.52**	79.80	86.53	85.07	**88.39**	80.73	84.39
	Our Method	**89.10**	90.59	**87.27**	**88.89**	**88.83**	88.04	**89.87**	**88.95**
Popular	Regular	79.98	82.47	76.72	79.48	76.10	73.22	82.94	77.78
	VCD	82.65	87.15	80.60	83.74	79.94	77.84	83.33	80.49
	Opera	84.21	**88.00**	79.80	83.70	78.33	73.85	87.73	80.19
	Our Method	**87.53**	87.73	**87.27**	**87.50**	**83.27**	**79.39**	**89.87**	**84.30**
Adversarial	Regular	76.03	76.11	76.80	76.45	72.37	68.78	83.06	75.24
	VCD	77.31	73.43	86.47	79.42	**76.32**	**73.24**	84.08	78.29
	Opera	80.88	**82.16**	79.76	80.94	75.50	70.50	87.73	78.17
	Our Method	**81.70**	78.90	**87.27**	**82.87**	76.23	70.84	**89.87**	**79.22**

Table 1: Performance comparison on POPE (Regular, Popular, and Adversarial) across two state-of-the-art Vision-Language Models (LLaVA 1.5 and InstructBlip). The best performance in each column is indicated in bold, and the second-best is underlined. Our proposed causal intervention method consistently outperforms existing methods (VCD, Opera), demonstrating improved accuracy and reduced hallucination across different evaluation settings.

Measuring $\mathrm{NDE}_{\mathbf{V,T}}$. To quantify the cross-modality direct effect of vision and text, we evaluate how vision complements textual information in multi-modal reasoning. Unlike NDE_V, which isolates vision's standalone impact, $NDE_{V,T}$ comprehensively captures the extent to which vision enhances or distorts textual semantic grounding.

We begin by sampling N images I_N and their corresponding textual descriptions C_N. For each image, we generate two perturbed versions: 1) I_{black} — a fully black image, containing no meaningful visual information. This setting ensures that the vision encoder receives an input with no structured content while preserving input dimensions and format. 2) I_{null} — a no-input condition, where the model receives no visual input at all. This serves as an extreme reference case to assess the model's reliance on textual information alone.

For each case, we obtain the visual representations $V_{i,k}^{I_{\text{black}}}$ and $V_{i,k}^{I_{\text{null}}}$ at the i-th layer and k-th token. The cross-modality direct effect is as:

$$D_{i,k}^{V,T} = V_{i,k}^{I_{\text{black}}} - V_{i,k}^{I_{\text{null}}}. \tag{8}$$

A high $\mathrm{NDE}_{V,T}$ suggests that vision provides complementary information to text, improving multi-modal understanding. Conversely, a low or negative $\mathrm{NDE}_{V,T}$ suggests that vision introduces noise or misalignment, potentially leading to hallucinated responses.

For global-level analysis, we stack the cross-modality direct effect vectors across N samples and apply PCA, using the first principal direction as the final estimate of $\mathrm{NDE}_{V,T}$.

Test-time Intervention. We integrate the computed Natural Direct Effects, NDE_V, NDE_T, and the cross-modal component $NDE_{V,T}$, to adjust the outputs of both the vision and text encoders during inference. We modify the intermediate representations at every layer and token position as follows:

$$V_{i,k}^{I'} = V_{i,k}^{I} + a \cdot NDE_V, \tag{9}$$

$$T_i^{C_N'} = T_i^{C_N} + b \cdot NDE_{V,T} + c \cdot NDE_T. \tag{10}$$

Our intervention method operates entirely at test time, offering a lightweight and architecture-agnostic solution compatible with all mainstream VLMs. The intervention directions are derived once from a random collection of $N = 50$ examples from MSCOCO (Lin et al., 2014), and remain

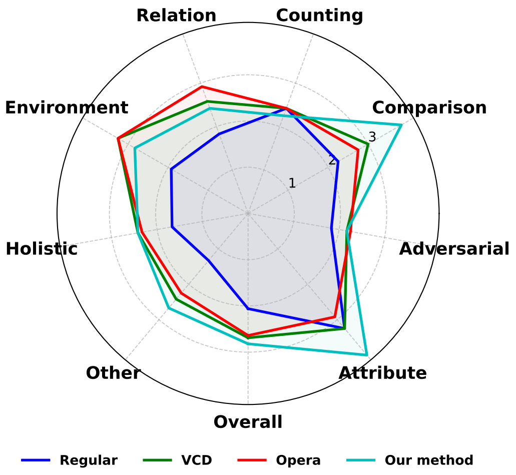


Figure 3: Overall performance and detailed score of different methods on the 8 question categories of MMHal-Bench. Our method achieves the best overall performance and significantly outperforms existing methods (VCD, Opera) in Attribute and Comparison.

unchanged throughout all evaluations. This unified configuration across datasets and tasks highlights the broad generalizability of the approach.

5 Experiments

5.1 Datasets and Evaluation Metrics

Datasets. We evaluate on two hallucination benchmarks: (1) MMHal-Bench (Sun et al., 2023), and (2) POPE (Li et al., 2023c). See details in Appx. B.

Evaluation Metrics for MMHal-Bench. According to the evaluation results in MMHal-Bench, GPT-4 (OpenAI, 2023) can achieve a 94% agreement rate with human judgments. Therefore, we use GPT-4o-mini (OpenAI, 2024) to analyze and score the responses of LMMs. Following the assessment method in MMHal-Bench, we provide GPT-4o-mini with the question and the VLM's response. Additionally, we supply the category name of the image content and a standard human-generated answer to improve the accuracy of response evaluation. Ultimately, GPT-4o-mini returns the VLM's scores across the 8 question categories and its hallucination rate.

Evaluation Metrics for POPE. Since POPE consists entirely of Yes/No questions, the correctness of VLM responses can be directly determined based on the ground-truth answers. This allows for the calculation of accuracy, precision, recall, and F1 score, with F1 score as the primary metric.

5.2 Implementation Details

We evaluate the effectiveness of our method on three widely used 7B VLMs, LLaVA 1.5 (Liu et al., 2023b), InstructBLIP (Dai et al., 2023), and Qwen2.5-VL-7B-Instruct (Bai et al., 2025). Additionally, we evaluate our method against two state-of-the-art baselines for alleviating hallucinations in the decoding stage: VCD (Leng et al., 2024) and Opera (Huang et al., 2024). Our default hyperparameter is sampling size $N = 50$. To ensure a fair comparison, we set $a = b = c = 0.9$ for all models across all experiments. Experiments are conducted using PyTorch with Nvidia RTX A6000 GPUs.

5.3 Experimental Results

Tab.1, Tab.2, and Fig. 3 demonstrate the effectiveness of our method compared to the SOTA approaches in three VLMs and two benchmarks. Our method consistently achieves best or near-best results in all metrics. More retults and analysis in Appx. C.

Results from Tab. 1 highlight key trends across Random, Popular, and Adversarial settings for LLaVA 1.5 and InstructBlip. In the Random setting, our method significantly improves accuracy (e.g., 83.49 to 89.10 in LLaVA 1.5) and recall (76.70 to 87.27), demonstrating the effectiveness of removing unintended direct modality influences. In the Popular setting, our method mitigates reliance on language priors, leading to higher accuracy (e.g., 79.98 to 87.53 in LLaVA 1.5) and F1 scores. Under the challenging Adversarial setting, our approach remains robust, significantly improving recall (76.80 to 87.27 in LLaVA 1.5) and F1 scores. These results validate that our causal intervention mechanism systematically reduces hallucination while enhancing resilience in diverse conditions.

Tab. 2 further demonstrates our method's superiority across MMHal-Bench categories, achieving the highest average performance (2.82). It excels in Attribute (4.00), Comparison (3.83), and Other (2.67) categories, indicating enhanced multi-modal reasoning. Strong performance in Holistic (2.42) and Environment (2.83) categories confirms that reducing unintended modality influences improves vision-text alignment.

Overall, our causal intervention framework effectively reduces hallucination, leading to more accurate and reliable multi-modal reasoning across diverse tasks. These results underscore the importance of addressing unintended modality biases in

Method	**Average**	Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
Regular	2.06	3.25	1.83	2.25	2.40	1.83	1.92	1.67	1.33
VCD	2.69	3.25	2.18	3.00	**2.42**	2.58	3.25	**2.42**	2.42
Opera	2.64	2.92	**2.25**	2.75	2.41	**2.92**	**3.26**	2.33	2.25
Our Method	**2.82**	**4.00**	2.17	**3.83**	2.25	2.42	2.83	**2.42**	**2.67**

Table 2: Performance comparison on MMHal-Bench with LLaVA 1.5. The best performance in each column is indicated in bold, and the second-best is underlined. Our proposed causal intervention method consistently outperforms existing methods (VCD, Opera), demonstrating improved accuracy and reduced hallucination across different evaluation settings.

PCA dim	**Average**	Attribute	Adversarial	Comparison	Counting	Relation	Environment	Holistic	Other
Regular	2.06	3.25	1.83	2.25	2.40	1.83	1.92	1.67	1.33
1	2.82	4.00	2.17	3.83	2.25	2.42	2.83	2.42	2.67
3	2.51	3.58	1.67	3.58	1.92	2.5	3.08	1.67	2.08
5	2.42	3.58	1.67	3.08	1.75	2.08	3.08	1.58	2.5

Table 3: Performance of LLaVA 1.5 on MMHal-Bench with different PCA dimensions. ‘Regular’ denotes the baseline method without any enhancement.

Number of samples	**Average↑**	**Hallucination rate↓**
Regular	2.06	64.58
25	2.45	51.04
50	2.82	45.83
75	2.62	45.83
100	2.58	50.00

Table 4: Performance of LLaVA 1.5 on MMHal-Bench with different numbers of samples. ‘Regular’ denotes the baseline method without any enhancement.

VLMs to improve robustness.

5.4 In-Depth Analysis

Measuring NDE with Different PCA Dimensions. Tab. 3 shows that using a single principal component (PCA dim = 1) yields the highest overall performance (2.82), outperforming PCA dim = 3 (2.51) and PCA dim = 5 (2.42). This suggests that restricting modality influence to a single direction effectively mitigates hallucinations while preserving multi-modal reasoning. Performance declines in Adversarial (from 2.17 to 1.67) and Holistic (2.42 → 1.58) categories with higher PCA dimensions indicate that excessive components may reintroduce noise, weakening robustness and interpretability. These results highlight that a minimal but targeted reduction in the influence of the modality enhances the accuracy of reasoning.

Effect of Sample Size. As shown in Tab. 4, using 50 samples achieves the best performance (2.82), outperforming both smaller (25 samples, 2.45) and larger settings (75 and 100 samples). Gains are most evident in Attribute (4.00) and Comparison (3.83), indicating improved hallucination mitigation. Performance drops at 75 and 100 samples suggest redundancy or overfitting, particularly in Adversarial and Holistic categories. These findings indicate that an optimal sample size (50) ensures robust estimation of modality influences while avoiding excessive noise, leading to better reasoning and reduced hallucinations.

Qualitative Analysis. To further demonstrate the effectiveness of our approach, we provide extensive visualizations comparing outputs before and after applying our method. These qualitative examples highlight reductions in hallucination and improved alignment with visual context. Detailed case studies can be found in the appx. D.

6 Conclusion

In this work, we introduced a causal framework to analyze and mitigate hallucination in VLMs. By constructing structural causal graphs and estimating the Natural Direct Effect of each modality, we identified unintended direct modality influences as a key contributor to hallucination. Our proposed test-time intervention mechanism effectively reduces modality bias, ensuring that generated outputs are more accurately grounded in fused multi-modal information. Empirical results across multiple benchmarks demonstrate that our method improves the reliability of VLMs while maintaining task performance.

7 Limitation & Ethical Consideration

Limitation: The causal framework may not capture all hallucination sources, especially in open-ended tasks. Also, the intervention introduces inference overhead, impacting real-time use. Future work can refine the causal model, develop task-specific adaptive interventions, and integrate contrastive learning for better multi-modal alignment.
Ethics Statement: Our method improves the reliability of the VLM by reducing hallucinations and improving trust in AI applications such as healthcare and autonomous systems. However, it does not eliminate biases in training data, and strict hallucination control may limit creative applications. Future work should balance factual consistency with flexibility across different use cases. This research improves the factual grounding of VLM without altering training data. Although our approach reduces hallucination, it does not guarantee complete accuracy, requiring users to apply additional validation in sensitive applications. Responsible deployment is key to effectively prevent misuse or excessive overreliance on AI-generated outputs.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo 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. *Preprint*, arXiv:2502.13923.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. 2024a. End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. 2024b. Multi-object hallucination in vision language models. *Advances in Neural Information Processing Systems*, 37:44393–44418.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. 2025. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. *Preprint*, arXiv:2503.19622.

Jerome Goddard. 2023. Hallucinations in chatgpt: a cautionary tale for biomedical researchers. *The American journal of medicine*, 136(11):1059–1060.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Yichong Huang, Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2021. The factual inconsistency problem in abstractive text summarization: A survey. *arXiv preprint arXiv:2104.14839*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

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

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2024. Interpreting and editing vision-language representations to mitigate hallucinations. *arXiv preprint arXiv:2410.02762*.

Osman Semih Kayhan, Bart Vredebregt, and Jan C Van Gemert. 2021. Hallucination in object detection—a study in visual part verification. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2234–2238. IEEE.

Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 595–603, Bejing, China. PMLR.

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.

David Lewis. 2013. *Counterfactuals*. John Wiley & Sons.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Li Li, Wei Ji, Yiming Wu, Mengze Li, You Qin, Lina Wei, and Roger Zimmermann. 2024a. Panoptic scene graph generation with semantics-prototype learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(4):3145–3153.

Li Li, You Qin, Wei Ji, Yuxiao Zhou, and Roger Zimmermann. 2024b. Domain-wise invariant learning for panoptic scene graph generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3165–3169.

Li Li, Chenwei Wang, You Qin, Wei Ji, and Renjie Liang. 2023b. Biased-predicate annotation identification via unbiased visual predicate representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, page 4410–4420.

Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2928–2937.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Aligning large multi-modal model with robust instruction tuning. *CoRR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. 2024. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Leland Gerson Neuberg. 2003. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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

James M Robins and Sander Greenland. 1992. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, and 1 others. 2023. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*.

Pei Wang and Nuno Vasconcelos. 2020. Scout: Self-aware discriminant counterfactual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8981–8990.

Yiming Yang, Yangyang Guo, Hui Lu, and Yan Wang. 2025. Vidlbeval: Benchmarking and mitigating language bias in video-involved lvlms. *arXiv preprint arXiv:2502.16602*.

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

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an eos decision perspective. *arXiv preprint arXiv:2402.14545*.

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. 2022. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113.

Ao Zhang, Hao Fei, Yuan Yao, Wei Ji, Li Li, Zhiyuan Liu, and Tat-Seng Chua. 2023. Vpgtrans: Transfer visual prompt generator across llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 20299–20319.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, and 1 others. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Zheng Zhao, Shay B Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. *arXiv preprint arXiv:2009.13312*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Related Works

Hallucination in Vision-Language Models. Recent work has developed VLMs by integrating visual encoders with pre-trained LLMs (Dai et al., 2023; Liu et al., 2023b; Zhu et al., 2023). This allows LLMs to interpret vision tokens from a pretrained backbone, achieving strong multimodal understanding (Zhang et al., 2023). However, these models also inherit the LLMs' tendency to generate ungrounded content, commonly termed "hallucination" (Bang et al., 2023; Huang et al., 2021; Favero et al., 2024). A major issue in VLM hallucinations is the incorrect inclusion of objects absent from the visual input (Bang et al., 2023; Huang et al., 2021; Li et al., 2023c; Wang et al., 2023). Studies suggest this often involves common or co-occurring objects in training data (Li et al., 2023a). Moreover, VLMs struggle with instructions requiring the recognition of absent objects, prompting research on improving model robustness (Liu et al., 2023a). Some studies attribute hallucinations to object co-occurrence, model uncertainty, and spatial positioning in text, proposing post-hoc correction methods (Zhou et al., 2023). Hallucination, originally studied in NLP, has become a concern in multimodal models due to its impact on performance (Ji et al., 2023). Common mitigation strategies rely on additional training to improve alignment with ground truth (Yue et al., 2024; Gao et al., 2025), but these methods demand significant data and computation. Training-free alternatives, such as self-feedback correction, auxiliary knowledge models, and enhanced decoding, offer practical solutions but often primarily focus on text rather than addressing vision-induced hallucinations (Yin et al., 2024).

Causality-Inspired Vision-Language Models. Causal inference provides a powerful framework for understanding and controlling the underlying mechanisms in machine learning models. By estimating causal effects, it enables the removal of spurious correlations, disentanglement of meaningful model behaviors, and identification of invariant features that enhance generalization across diverse scenarios (Li et al., 2022). Recently, causal methods have been increasingly applied to computer vision, benefiting tasks such as visual explanation (Wang and Vasconcelos, 2020), image and video recognition (Li et al., 2023b), scene graph generation (Li et al., 2024b), and representation learning (Li et al., 2024a). In the context of VLMs, causal analysis is particularly valuable for addressing hallucination, as it allows us to separate genuine multi-modal reasoning from biased modality dominance. By leveraging causal graphs and counterfactual reasoning, we can systematically diagnose and mitigate modality-specific artifacts, ensuring that model predictions are grounded in meaningful cross-modal interactions rather than unintended shortcuts.

B Additional Experimental Settings

As briefly discussed in §5.1, we evaluate our method on two benchmarks.

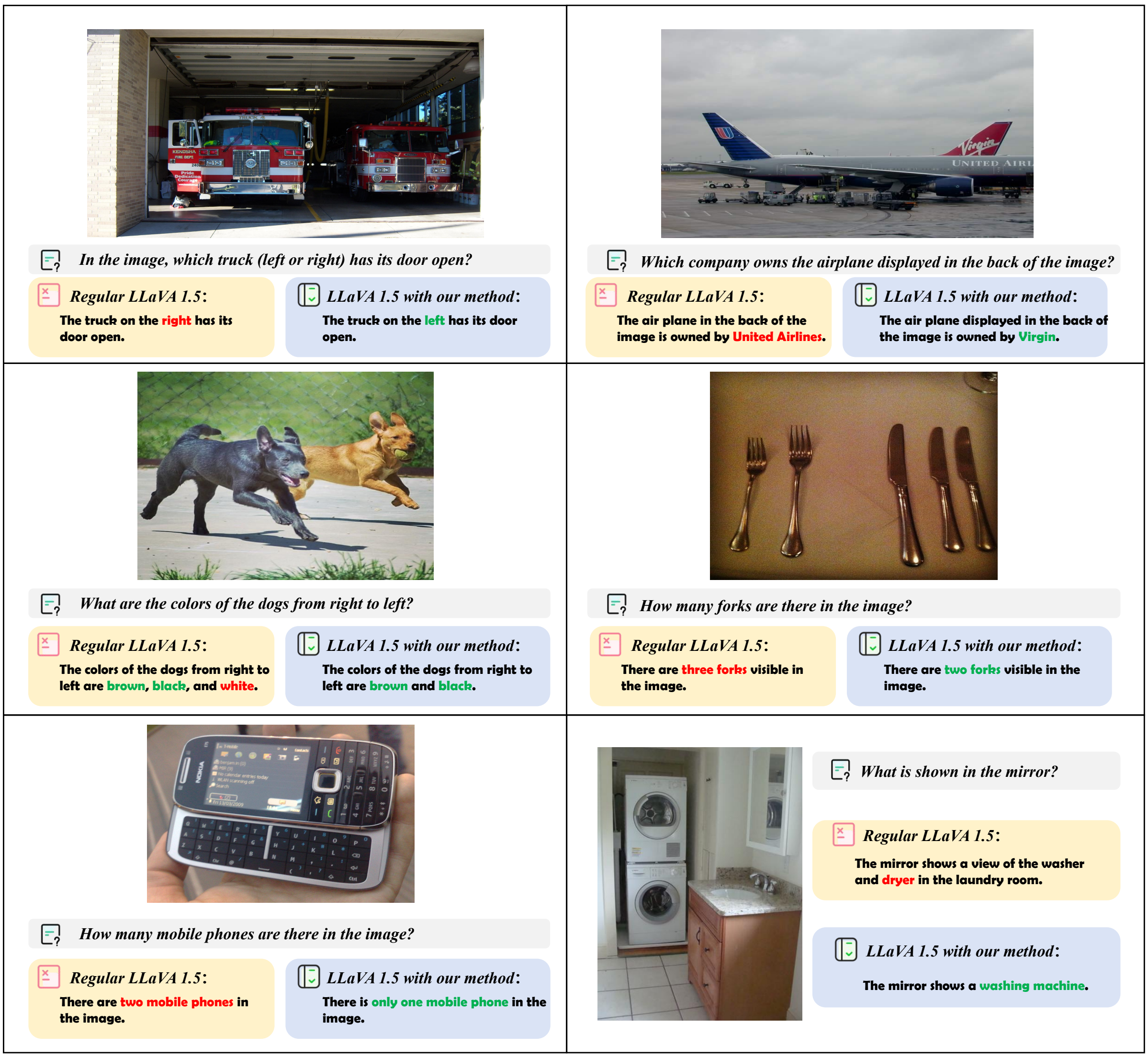
(1) **MMHal-Bench** (Sun et al., 2023) is designed to evaluate hallucinations in VLMs' responses. It includes 96 image-question pairs across 8 question categories and 12 object topics from MSCOCO (Lin et al., 2014). It specifically targets types of questions where VLMs are prone to making false claims about image content, including object attributes, adversarial objects, comparison, counting, spatial relations, environment, holistic description, and other cases, such as misreading text or icons. Evaluation is conducted using GPT-4o-mini, which compares model responses against human-generated answers to determine hallucination presence, and additional context is provided to enhance its judgment.

(2) **POPE** (Li et al., 2023c) (Polling-based Object Probing Evaluation) is a polling-based evaluation benchmark for assessing object hallucination in VLMs. It formulates the evaluation of object hallucination as a binary classification task by prompting VLMs with questions that require "Yes" or "No" responses. POPE maintains a balanced distribution, ensuring an equal split between queries for existing and non-existing objects, and utilizes three sampling strategies: random, popular, and adversarial. It collects 500 images from each of the MSCOCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019), and then samples objects that VLMs are prone to hallucinate, generating a total of 27,000 challenging Yes/No questions to assess the model's ability to correctly identify objects in images. POPE adopts Accuracy, Precision, Recall, and F1-score as evaluation metrics.

C Additional Experimental Analysis

As briefly discussed in §5.3, we evaluate our method on two benchmarks.

The results summarized in Tab. 1 reveal several notable trends when comparing our proposed



In the image, which truck (left or right) has its door open?

Regular LLaVA 1.5: The truck on the right has its door open.

LLaVA 1.5 with our method: The truck on the left has its door open.

Which company owns the airplane displayed in the back of the image?

Regular LLaVA 1.5: The air plane in the back of the image is owned by United Airlines.

LLaVA 1.5 with our method: The air plane displayed in the back of the image is owned by Virgin.

What are the colors of the dogs from right to left?

Regular LLaVA 1.5: The colors of the dogs from right to left are brown, black, and white.

LLaVA 1.5 with our method: The colors of the dogs from right to left are brown and black.

How many forks are there in the image?

Regular LLaVA 1.5: There are three forks visible in the image.

LLaVA 1.5 with our method: There are two forks visible in the image.

How many mobile phones are there in the image?

Regular LLaVA 1.5: There are two mobile phones in the image.

LLaVA 1.5 with our method: There is only one mobile phone in the image.

What is shown in the mirror?

Regular LLaVA 1.5: The mirror shows a view of the washer and dryer in the laundry room.

LLaVA 1.5 with our method: The mirror shows a washing machine.

Figure A: Case study illustrating the impact of our method on VLM hallucination. The figure compares outputs from the original model and our enhanced approach, highlighting reductions in hallucinated content and improved alignment with the visual context. Our method effectively mitigates incorrect textual descriptions by refining modality interactions, leading to more accurate and reliable multi-modal reasoning.

method to existing approaches across Random, Popular, and Adversarial settings for both LLaVA 1.5 and InstructBlip. Under the Random setting, our method achieves a clear advantage. For instance, with LLaVA 1.5, accuracy increases from 83.49 in the Regular baseline to 89.10, while recall improves from 76.70 to 87.27. In InstructBlip, similar gains are observed: accuracy rises from 80.42 to 88.83, and recall from 83.21 to 89.87. These improvements indicate that our test-time intervention module, which systematically estimates and removes the unintended direct influences from each modality, effectively reduces hallucinations and leads to better alignment between the generated outputs and the intended multi-modal context.

In the Popular setting, our approach again outperforms the alternatives. For LLaVA 1.5, our method boosts accuracy from 79.98 (Regular) to 87.53 and enhances the F1 score from 79.48 to 87.50. InstructBlip also benefits, with accuracy improving from 76.10 to 83.27 and F1 score rising from 77.78 to 84.30. These results suggest that by mitigating the model's over-reliance on language priors and counteracting spurious correlations present in the training data, our method promotes a more balanced integration of visual and textual cues. The most challenging conditions are observed under the Adversarial setting. Here, the LLaVA 1.5 model's recall jumps significantly from 76.80 to 87.27, and the F1 score improves from 76.45 to 82.87. Although the

Settings	Method	Qwen2.5-VL-7B-Instruct			
		Acc	Prec	Rec	F1
Random	Regular	84.43	**99.71**	69.07	81.61
	VCD	**86.44**	98.90	70.23	82.14
	Opera	85.80	98.40	69.90	81.90
	Ours	85.50	98.17	**71.60**	**83.16**
Popular	Regular	83.87	**98.02**	69.13	81.08
	VCD	**85.63**	96.91	70.47	81.60
	Opera	85.10	96.40	70.60	81.50
	Ours	84.37	96.15	**71.60**	**82.08**
Advers.	Regular	83.40	**96.91**	69.00	80.61
	VCD	**84.53**	94.30	70.43	80.64
	Opera	84.00	94.90	71.10	81.00
	Ours	83.77	95.13	**71.60**	**81.52**

Table A: Performance of Qwen2.5-VL-7B-Instruct across three POPE evaluation settings (Regular, Popular, Adversarial). Best values are in **bold** and second-best are underlined.

improvements in InstructBlip are more modest in terms of accuracy (from 72.37 to 76.23), both recall and F1 scores show meaningful enhancements. This pattern indicates that our approach is robust even when the input signals are intentionally degraded or perturbed, highlighting its potential for real-world applications where input quality may vary. Overall, the experimental data suggest that our causal intervention mechanism—grounded in counterfactual analysis and Natural Direct Effect estimation—is effective in systematically reducing hallucination in VLMs. By eliminating unintended direct modality influences, our method not only improves the accuracy of vision-text fusion but also enhances the model's resilience across diverse and challenging scenarios.

The experimental results presented in Table 2 demonstrate the effectiveness of our proposed causal intervention approach in mitigating hallucination and improving the accuracy of vision-language models (VLMs) across multiple reasoning categories in the MMHal-Bench benchmark. Compared to existing methods, our approach consistently achieves the highest average performance score (2.82), outperforming both VCD (2.69) and Opera (2.64), as well as the regular baseline (2.06). A closer examination of the category-wise results reveals that our method exhibits notable improvements in specific reasoning types. In particular, it achieves the highest performance in Attribute (4.00), Comparison (3.83), and Other (2.67) categories. The superior performance in Attribute reasoning suggests that our method enhances the model's ability to accurately associate visual details with textual descriptions, a critical factor in reducing hallucinated object properties. Similarly, the strong performance in Comparison tasks indicates improved cross-instance reasoning, likely due to our causal intervention strategy, which ensures that both visual and textual modalities contribute meaningfully to the generated response rather than relying on language priors. In contrast, while our method does not achieve the highest score in Adversarial, Counting, and Relation categories, it remains competitive, showing marginal differences from the top-performing methods. For instance, in the Adversarial category, our score (2.17) is comparable to Opera (2.25), suggesting that while causal intervention reduces hallucination, certain adversarial perturbations may still challenge the model's robustness. Additionally, in Counting (2.25), our approach is slightly lower than VCD (2.42), possibly indicating that direct modality influence alone may not fully address numerical inconsistencies, which often require improved object permanence reasoning. Importantly, our approach demonstrates a balanced improvement across multiple reasoning types, particularly excelling in categories where multi-modal fusion plays a crucial role, such as Holistic (2.42) and Environment (2.83). These results support our hypothesis that hallucination arises due to unintended direct influences from individual modalities, and by systematically mitigating these effects, our method enhances the model's ability to generate more reliable and contextually grounded outputs. Overall, these findings validate the effectiveness of our causal intervention framework in reducing hallucination and improving reasoning accuracy across diverse evaluation settings. The performance gains across multiple reasoning categories highlight the necessity of explicitly addressing unintended modality biases in VLMs, reinforcing the potential of causal analysis as a key tool in advancing the robustness of multi-modal models.

D Qualitative Result

As briefly discussed in §5.4, we provide more qualitative results to showcase the effectiveness of our method, as shown in Fig. A.