Causal Tracing of Object Representations in Large Vision Language Models: Mechanistic Interpretability and Hallucination Mitigation

Qiming Li[1]*, Zekai Ye[1]*, Xiaocheng Feng[1,2]†, Weihong Zhong[1], Weitao Ma[1], Xiachong Feng[3]†

[1]Harbin Institute of Technology
[2]Peng Cheng Laboratory
[3]The University of Hong Kong
{qmli,zkye}@ir.hit.edu.cn

Abstract

Despite the remarkable advancements of Large Vision-Language Models (LVLMs), the mechanistic interpretability remains underexplored. Existing analyses are insufficiently comprehensive and lack examination covering visual and textual tokens, model components, and the full range of layers. This limitation restricts actionable insights to improve the faithfulness of model output and the development of downstream tasks, such as hallucination mitigation. To address this limitation, we introduce **Fine-grained Cross-modal Causal Tracing (FCCT)** framework, which systematically quantifies the causal effects on visual object perception. FCCT conducts fine-grained analysis covering the full range of visual and textual tokens, three core model components including multi-head self-attention (MHSA), feed-forward networks (FFNs), and hidden states, across all decoder layers. Our analysis is the first to demonstrate that MHSAs of the last token in middle layers play a critical role in aggregating cross-modal information, while FFNs exhibit a three-stage hierarchical progression for the storage and transfer of visual object representations. Building on these insights, we propose **Intermediate Representation Injection (IRI)**, a training-free inference-time technique that reinforces visual object information flow by precisely intervening on cross-modal representations at specific components and layers, thereby enhancing perception and mitigating hallucination. Consistent improvements across five widely used benchmarks and LVLMs demonstrate **IRI** achieves state-of-the-art performance, while preserving inference speed and other foundational performance.

Introduction

Large Vision-Language Models (LVLMs) have rapidly evolved, demonstrating impressive capabilities across diverse tasks. However, existing interpretability studies fall short in capturing the full complexity of visual information flow, thereby limiting progress in critical downstream applications such as hallucination mitigation. Fundamental questions require further investigation, particularly regarding how LVLMs process visual object features and align them with textual semantics in cross-modal representations. Furthermore, how visual and textual tokens elicit distinct functional behaviors from three core model components—such as multi-head self-attention (MHSA), feed-forward networks (FFNs), and hidden states—across layers, remains underexplored. Previous studies have partially addressed these questions but still leave notable limitations. Attention knockout experiments (Neo et al. 2024) demonstrate that LVLMs extract object information from visual object tokens in the middle to late layers. However, it lacks an analysis of the cross-modal interactions between visual and textual tokens, as well as the functional roles of the MLP and hidden states. NOTICE (Golovanevsky et al. 2024) introduces semantic image pairs for image corruption and symmetric token replacement for text corruption to analyze how MHSA and MLP contribute to information aggregation in textual tokens, but overlooks the effect of visual tokens.

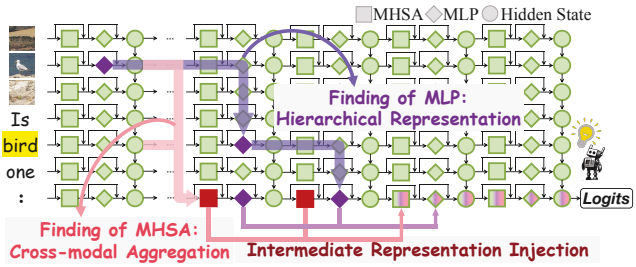

Figure 1: An overview of our proposed **Fine-grained Cross-modal Causal Tracing (FCCT)** findings and **Intermediate Representation Injection (IRI)** method.

To address these limitations, we propose a **Fine-grained Cross-modal Causal Tracing (FCCT)** framework, which systematically analyzes cross-modal causal effects on visual perception by examining visual and textual tokens categorized with their position and semantic role in the input sequence, covering three core model components across layers. By introducing controlled Gaussian perturbations to input images, we induce measurable drops in output probabilities for existing objects. Then we restore specific activations using the clean activations from the original image input. By quantifying the recovery in the LVLM's output probabil-

*These authors contributed equally.
†Corresponding authors

ities, we precisely estimate the causal effect on visual object perception for each core components across token types and layers. **FCCT** is the first to demonstrate that the MHSAs of the last token in middle layers play a critical role in aggregating crucial object visual and textual information, as well as the FFNs exhibit a three-stage hierarchical progression for the storage and transfer of visual object representations.

FCCT not only reveals cross-modal information flow of visual objects, but also provides valuable guidance for hallucination mitigation. Prior studies (Tang et al. 2025) suggest that deep unidirectional information flow can lead to the progressive degradation of fine-grained semantic cues encoded in earlier layers, which may be a potential cause of object hallucination. To prevent mid-layer degradation of critical information during forward and reinforce components with strong causal effects identified by **FCCT**, we further propose **Intermediate Representation Injection (IRI)**, a training-free inference-time technique that injects crucial mid-layer representations into subsequent layers, thereby enhancing visual perception capability and mitigating hallucination. **FCCT** offers fine-grained and quantitative guidance for selecting model components and layers, serving as a theoretical foundation for the design and implementation of the **IRI** method. Consistent improvement across five widely used benchmarks and five advanced LVLMs demonstrates that **IRI** achieves state-of-the-art (SOTA) performance.

In summary, our main contributions are three-fold:

- We propose **FCCT**, a fine-grained causal analysis that covers all types of visual and textual tokens, three core components across the full layer range, providing a comprehensive mechanistic interpretability study of LVLMs.
- We propose **IRI**, a training-free inference-time method that effectively mitigates hallucination while preserving inference speed and other foundational capabilities.
- Consistent improvements across five widely used benchmarks and LVLMs not only demonstrate **IRI**'s SOTA performance, but also validate the findings of **FCCT**.

Related Work

Mechanistic Interpretability of LVLMs While LVLMs have demonstrated remarkable capabilities across various downstream tasks, their mechanistic interpretability remains underexplored. Existing interpretability methods, such as probing (Salin et al. 2022), activation patching (Basu et al. 2024; Palit et al. 2023; Golovanevsky et al. 2024), logit lens (Neo et al. 2024; Huo et al. 2024), in-context learning (Li et al. 2025d,b,c) provide only a coarse-grained analysis of model components or do not fully disentangle the complex interactions between visual and textual token representations. In contrast, our work employs causal tracing with Gaussian noise to precisely quantify the functional roles of MHSA, FFN, and hidden states for both visual and textual tokens across layers, enabling a fine-grained analysis of how LVLMs perceive and process visual object information.

Mitigating Hallucination in LVLMs LVLMs frequently produce content that deviates from visual information, leading to object hallucination. Existing hallucination mitigating strategies can be broadly categorized into three types: **(1) Training-based approaches** enhance model factuality by pre-training or finetuning with carefully curated datasets (Yu et al. 2024; You et al. 2023; Zhang et al. 2025) and novel training objectives (Lyu et al. 2024). These methods can be effective but require substantial data and computational resources. **(2) Contrastive decoding** (Leng et al. 2024; Huang et al. 2024; Zhong et al. 2024) leverages differences between deliberately perturbed decoding paths to promote generations that are more consistent in visual information. However, such methods introduce significant latency at inference time. **(3) Inference-time interventions** modify internal activations such as attention heads outputs (Liu, Zheng, and Chen 2024; Li et al. 2025a; Ye et al. 2025) or hidden states (Liu, Ye, and Zou 2024) to steer the model toward more faithful outputs. However, these methods generally lack interpretability of the selection of layers and components.

Preliminary

We restrict our scope to LVLMs that are based on auto-regressive Transformer architecture (Vaswani et al. 2017), as it is adopted by most SOTA LVLMs. The model receives as input a visual input sequence $\boldsymbol{V} = \{v_1, v_2, \ldots, v_m\}$ and a textual input sequence $\boldsymbol{T} = \{t_1, t_2, \ldots, t_n\}$, where m and n denote the sequence lengths of the visual and textual input. The textual and visual input sequences are concatenated together and processed through L transformer layers of the language decoder, each consisting of multi-head self-attention (MHSA), feed-forward network (FFN) that is usually a multilayer perception (MLP), and a residual stream is applied between each components. The l-th layer hidden state $\boldsymbol{h}^{(l)}$ can be computed from the previous layer:

$$\boldsymbol{h}^{(l)} = \boldsymbol{h}^{(l-1)} + \boldsymbol{a}^{(l)} + \boldsymbol{m}^{(l)}, \tag{1}$$

where $\boldsymbol{a}^{(l)}$ and $\boldsymbol{m}^{(l)}$ are the output of the MHSA component and the FFN component at layer l. Finally, the model predicts the next token in an auto-regressive manner based on the last layer output.

In this paper, we aim to identify which types of visual and textual tokens, model components (i.e., $\boldsymbol{a}^{(l)}$, $\boldsymbol{m}^{(l)}$, and $\boldsymbol{h}^{(l)}$), and layer ranges play a critical role in the perception and comprehension of visual object information in LVLMs. By uncovering the underlying information flow, we seek to provide practical guidance for mitigating object hallucination and related downstream issues.

Fine-Grained Cross-Modal Causal Tracing

Causal tracing (also known as activation patching or causal mediation analysis) is a widely used interpretability technique that selectively replaces internal activations to probe the causal contribution of specific model components (Meng et al. 2022). In the context of large language models (LLMs), causal tracing is frequently employed to examine the storage and retrieval mechanisms of factual associations (Meng et al. 2022), and document-level relevance (Liu, Mao, and Wen 2025). In the context of LVLMs, we are the first to propose using controlled Gaussian noise perturbations on input images for causal tracing. By adding controlled Gaussian noise

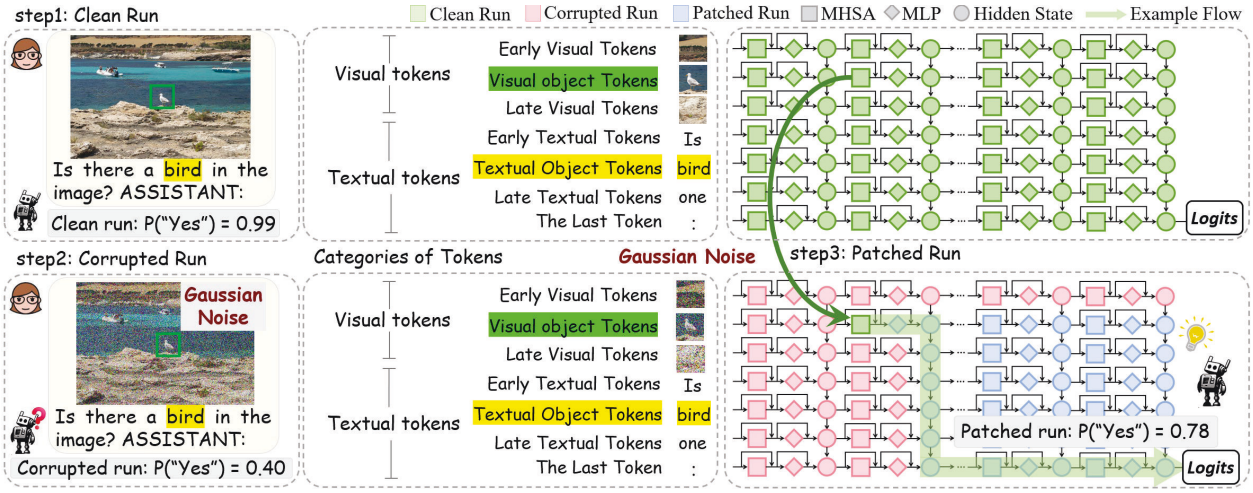

Figure 2: Overview of our proposed **Fine-grained Cross-modal Causal Tracing** method. Activation patching computes the causal effect of a specific component by running the LVLMs three times: a **clean run (step1)** with original image, a **corrupted run (step2)** with image added Gaussian noise, and a **patched run (step3)** with corrupted input but restoring specific component using the value in the clean run. We use **Recovery Rate** to quantify the causal effect of each restored component.

to the entire image and then selectively restoring the activations of specific components, we conduct a fine-grained analysis of the internal mechanisms responsible for visual object perception and comprehension in LVLMs.

Specifically, we select 500 images from the COCO dataset and design object-related questions for the objects present in each image. Following the analysis methodology of ROME (Meng et al. 2022), we define three types of inference runs:

- **Clean Run:** The model is given the original image, and we record the probability $\boldsymbol{P}_{\text{clean}}$ assigned to the token `yes` in response to binary questions of the form "Is there a XXX in the image? Please answer this question with one word (Yes or No)."
- **Corrupted Run:** Gaussian noise is added to the entire image to degrade visual quality, and we record the resulting prediction probability $\boldsymbol{P}_{\text{corrupted}}$.
- **Patched Run:** Starting from the corrupted run, we selectively restore specific internal activations (e.g., MHSA output, MLP output, or hidden states) at certain layers and token categories using activations from the clean run. The prediction probability is denoted as $\boldsymbol{P}_{\text{patched}}$.

To quantify the causal effect of each restored component, we define the following **Recovery Rate** ($\boldsymbol{RR}$) metric:

$$\boldsymbol{RR} = \frac{\boldsymbol{P}_{\text{patched}} - \boldsymbol{P}_{\text{corrupted}}}{\boldsymbol{P}_{\text{clean}} - \boldsymbol{P}_{\text{corrupted}}} \tag{2}$$

This normalized measure reflects the proportion of clean performance regained through targeted restoration; more profoundly, it serves as a quantitative estimate of the component's causal effect on visual object perception. A value close to 1 indicates a strong causal effect, whereas a value near 0 suggests minimal influence.

To systematically conduct causal tracing across visual and textual information flow, we define seven categories based on token's position and semantic role in the input sequence:

① **Early Visual Tokens** occur before any queried object region, which serve as a control group for comparison.

② **Object Visual Tokens** directly encode visual features corresponding to the queried object, which are central for analyzing the internal mechanisms of visual object perception and comprehension.

③ **Late Visual Tokens** occur after any queried object region, which may capture residual visual information.

④ **Early Textual Tokens** occur near the visual&textual sequences boundary, which help analyze how information transitions from visual to language components.

⑤ **Textual Object Tokens** encode textual features corresponding to the queried object, which help reveal how visual object information interacts with textual reference.

⑥ **Late Textual Tokens** occur in the late part of the textual prompt, which help analyze how visual object information propagates across textual stream not directly related.

⑦ **The Last Token** occurs at the end of input sequence, helping analyze cross-modal information aggregation.

By restoring only one component of one layer for a single token category at one time, we systematically derive fine-grained insights into how LVLMs perceive and comprehend visual object information. This enables us to trace the causal pathways through which visual object information is represented, propagated, and aggregated across different layers and model components.

Causal Tracing Results and Key Findings

In this section, we present and analyze the experimental results and key findings of FCCT conducted on two widely used LVLMs: LLaVA-1.5-7B and Qwen-VL-Chat. As illustrated in Figure 3, we present the $\boldsymbol{RR}$**s** of 3 model compo-

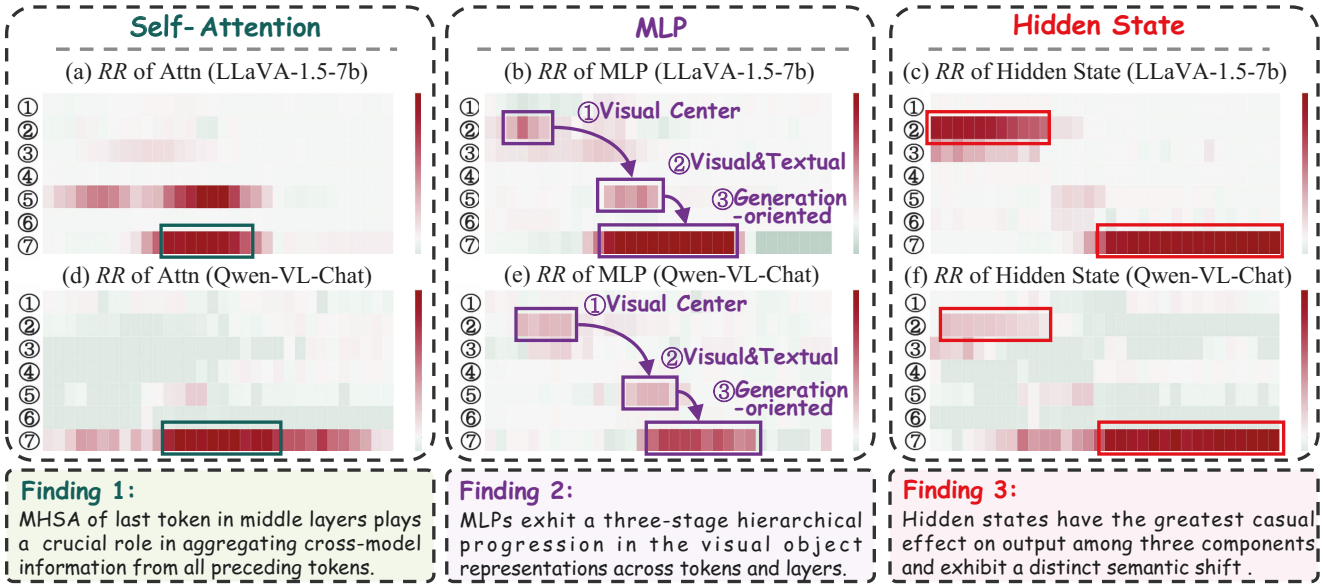

Figure 3: Results and key findings of **FCCT** framework on LLaVA-1.5-7b and Qwen-VL-Chat. The symbols from ① to ⑦ represent the seven token categories defined above: **① Early Visual Tokens, ② Object Visual Tokens, ③ Late Visual Tokens, ④ Early Textual Tokens, ⑤ Textual Object Tokens, ⑥ Late Textual Tokens, and ⑦ The Last Token**.

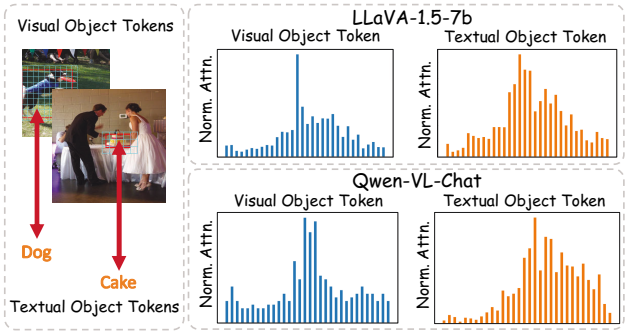

Figure 4: Visualization of normalized attention weights to visual object tokens and corresponding textual object tokens across layers. We report the average result on 3,000 VQAs.

nents across 32 layers under 7 token categories, denoted as ① to ⑦, corresponding to the previous definition.

Cross-modal Aggregation via the Last Token's MHSAs As shown in Figure 2 (a) and (d), the last token's MHSAs in intermediate layers exhibit a strong causal effect, which plays a particularly crucial role in aggregating information from all preceding tokens.

To further investigate the cross-modal aggregation effect of MHSA, we visualized the last token's layer-wise normalized attention weights for the queried visual object tokens and textual object tokens. As shown in Figure 3, we observe a sharp increase in attention weight around Layer 15. This suggests that around these layers, LVLMs begin to align instruction-guided attention with the most relevant visual and textual cues. We hypothesize that these layers mark a transition point where deep cross-modal information aggregation occurs, enabling the model to bind unimodal representations to high-level cross-modal representations.

Three-stage Hierarchical Representations via MLPs As shown in Figure 2 (b) and (e), MLPs exhibits a three-stage progression in the formation of visual object representations: In early layers, visual object tokens are encoded into localized, modality-specific embeddings with limited semantic abstraction; In intermediate layers, textual object tokens interact with visual representations, forming increasingly rich cross-modal semantics; In the deeper layers, under the cross-modal aggregation effect of MHSA, the last token's MLPs progressively accumulate a cross-modal and task-relevant representation.

Overall, this hierarchical progression illustrates how MHSA-driven cross-modal interactions and MLPs together transform unimodal and localized visual representations into cross-modal and globally aggregated representations that are essential for visual object perception in LVLMs.

Hierarchical Semantic Shift of the Hidden States As shown in Figure 2 (c) and (f), hidden states exhibit a distinct semantic shift: in shallow layers (Layers 0–10), visual tokens' hidden states are primarily visual-centric, encoding low-level perceptual patterns. In deep layers (Layers 18–31), the last tokens' are cross-modal and highly task-related to the final prediction. Notably, the phase in which the causal effect of the hidden states gradually strengthens aligns with the intermediate layers around Layer 14, precisely where MHSAs and MLPs jointly contribute most significantly to refining cross-model object representations.

Intermediate Representation Injection

Motivated by the findings from the proposed FCCT analysis, we observe that the last token's MHSA and MLP outputs at intermediate layers are crucial for capturing and aggregating task-related object information from both visual and textual modalities. To take advantage of this observation, we propose **Intermediate Representation Injection (IRI)**, a training-free inference-time technique, which aims to reinforce crucial cross-modal representations to improve visual object information perception and mitigate hallucination.

Method

To reinforce crucial mid-layer cross-modal representations, we selectively inject them into later layers, which are adaptively scaled by their causal effects (Recovery Rates), thereby amplifying the contributions of the most influential components across layers.

Let $RR^{\text{attn}} = \{RR_k^{\text{attn}}\}_{k=1}^{L}$ and $RR^{\text{mlp}} = \{RR_k^{\text{mlp}}\}_{k=1}^{L}$ denote the Recovery Rates of the MHSA and MLP outputs across the L layers, respectively. To ensure that the most critical cross-modal representations are injected into layers with sufficient causal effect, we rank components across layers by their Recovery Rates and independently select:

- A set of top-k_1 MHSA source layers $\mathcal{L}_{\text{src}}^{\text{attn}}$ with the highest RR^{attn} values and a set of top-k_2 target layers $\mathcal{L}_{\text{tgt}}^{\text{attn}}$.
- A set of top-k_1 MLP source layers $\mathcal{L}_{\text{src}}^{\text{mlp}}$ with the highest RR^{mlp} values and a set of top-k_2 target layers $\mathcal{L}_{\text{tgt}}^{\text{mlp}}$.

For each attention source layer $k \in \mathcal{L}_{\text{src}}^{\text{attn}}$, we record the MHSA output of the last token: $\boldsymbol{a}^{(k)} \in \mathbb{R}^d$. For each target layer $l \in \mathcal{L}_{\text{tgt}}^{\text{attn}}$ such that $l > k$, we inject the stored activations into the target MHSA output, and scaling them by RR to modulate their contribution to reflect causal effect:

$$\tilde{\boldsymbol{a}}^{(l)} = \boldsymbol{a}^{(l)} + \lambda_a \cdot \sum_{k \in \mathcal{L}_{\text{src}}^{\text{attn}}} g(k,l) \cdot RR_k^{\text{attn}} \cdot \boldsymbol{a}^{(k)}, \quad (3)$$

Similarly, for each $k \in \mathcal{L}_{\text{src}}^{\text{mlp}}$ and $l \in \mathcal{L}_{\text{tgt}}^{\text{mlp}}$ with $l > k$, we inject the recorded MLP outputs as:

$$\tilde{\boldsymbol{m}}^{(l)} = \boldsymbol{m}^{(l)} + \lambda_m \cdot \sum_{k \in \mathcal{L}_{\text{src}}^{\text{mlp}}} g(k,l) \cdot RR_k^{\text{mlp}} \cdot \boldsymbol{m}^{(k)}, \quad (4)$$

where λ_a and λ_m are scaling coefficients, and $g(k,l)$ ensures that the injected information respects causal ordering:

$$g(k,l) = \begin{cases} 1, & \text{if } l > k \\ 0, & \text{otherwise} \end{cases}$$

To ensure that the injected activation maintains the same norm as the original, we apply the following normalization:

$$\tilde{\boldsymbol{a}}^{(l)} = \tilde{\boldsymbol{a}}^{(l)} \cdot \frac{\|\boldsymbol{a}^{(l)}\|_2}{\|\tilde{\boldsymbol{a}}^{(l)}\|_2}, \tilde{\boldsymbol{m}}^{(l)} = \tilde{\boldsymbol{m}}^{(l)} \cdot \frac{\|\boldsymbol{m}^{(l)}\|_2}{\|\tilde{\boldsymbol{m}}^{(l)}\|_2}, \quad (5)$$

where $||\cdot||$ represents the ℓ_2 norms (Euclidean norms) of the activation vectors.

This intervention ensures that only critical cross-modal intermediate activations are injected into subsequent layers with sufficient causal effect throughout the information flow, enhancing the LVLM's trustworthiness to visual object information and thereby mitigating hallucination.

Experimental Setup

Models. We adopt the widely used LLaVA-1.5-7b (Liu et al. 2024a), Qwen-VL-Chat (Bai et al. 2023), LLaVA-NeXT (Liu et al. 2024b), Qwen2-VL-7B(Wang et al. 2024) and InternVL2-8B (Chen et al. 2024) as baseline LVLMs.

Evaluation. We comprehensively evaluate the methods for both discriminative and generative tasks to measure the effectiveness and robustness of hallucination mitigation.

- **POPE** (Li et al. 2023) employs a binary question-answering format, inquiring LVLMs to answer if a special object exists in the given image. Following previous works, we adopt Accuracy and F1 score as the metrics.
- **MME** (Fu et al. 2023) serves as a comprehensive tool for assessing the capabilities of LVLMs across both 10 perception tasks and 4 cognition tasks. Consequently, task scores are reported as the evaluation metric.
- **CHAIR** (Rohrbach et al. 2018) is a widely used metric for assessing object hallucination in responses of LVLMs. The CHAIR metric comprises two important indicators, denoted as C_S and C_I, with the following calculation formulas:

$$C_S = \frac{|\{\text{Hallucinated objects}\}|}{|\{\text{All mentioned objects}\}|}$$

$$C_I = \frac{|\{\text{Sentences w/ hallucinated objects}\}|}{|\{\text{All sentences}\}|}$$

- **MMHal-Bench** (Sun et al. 2023) comprises 96 meticulously designed questions, which evaluates response-level hallucination rate (VH.%) and informativeness (Score). It asks **GPT-4** to compare model outputs with human responses and object labels for evaluation.
- **MHumanEval** (Yu et al. 2024) is designed to evaluate hallucination performance by **human annotators**. The benchmark contains 146 samples collected from Object HalBench and MMHal-Bench. Given model responses, we ask three human annotators to label the hallucinated segments and compute the mean response-level hallucination rate (Hu.%) as the evaluation metric.

Baselines. We compared our proposed **IRI** method with the following SOTA training-free methods: **VCD** (Leng et al. 2024) contrasts model logits derived from original and distorted visual input to reduce the over-reliance on statistical bias and unimodal priors. **OPERA** (Huang et al. 2024) introduces a penalty term on the model logits during the beam-search decoding to mitigate the over-trust issue. **PAI** (Liu, Zheng, and Chen 2024) intervenes on attention heads by leveraging their original direction and optimizes the output distribution during decoding to mitigate language bias. **VTI** (Liu, Ye, and Zou 2024) mitigates hallucination by steering layer-wise hidden states during inference to enhance visual feature stability.

Method	LLaVA-1.5-7b					Qwen-VL-Chat					LLaVA-NeXT				
	Exist.	Count	Pos.	Color	Total	Exist.	Count	Pos.	Color	Total	Exist.	Count	Pos.	Color	Total
Regular	175.7	124.7	114.0	151.0	565.4	170.0	135.0	123.3	170.0	598.3	180.0	105.0	150.0	151.7	586.7
VCD	180.3	131.7	125.0	155.0	592.0	180.0	133.3	131.7	175.0	620.0	185.0	125.0	133.3	168.3	611.6
OPERA	165.0	116.0	133.3	149.0	563.3	180.0	140.0	**138.3**	175.0	633.3	183.8	121.3	**155.0**	162.1	622.2
PAI	190.0	**148.3**	126.7	160.0	625.0	175.0	141.6	132.5	177.5	626.6	185.0	128.3	148.3	170.8	632.4
VTI	185.0	140.0	135.0	165.7	625.7	180.0	142.5	133.0	178.0	633.5	186.7	126.7	150.0	172.5	635.9
IRI(ours)	**195.0**	140.0	**140.0**	**168.3**	**648.3**	**185.0**	**145.0**	135.0	**180.0**	**645.0**	**190.0**	**135.0**	**155.0**	**177.5**	**657.5**

Table 1: Results on MME hallucination subset. The best performances are bolded and the second-best are underlined.

Setting	Method	LLaVA-1.5-7b		Qwen-VL-Chat		LLaVA-NeXT	
		Acc	F1	Acc	F1	Acc	F1
Ran.	Regular	83.29	81.33	84.63	82.61	84.78	86.43
	VCD	87.73	87.16	86.93	85.46	88.76	89.57
	OPERA	89.20	88.81	85.71	84.64	90.27	89.71
	PAI	86.33	84.56	85.38	85.54	88.40	87.16
	VTI	89.50	88.89	86.73	85.59	89.23	88.68
	IRI(ours)	**89.76**	**89.32**	**87.38**	**87.42**	**90.68**	**90.21**
Pop.	Regular	81.88	80.06	83.63	81.53	83.23	84.77
	VCD	85.38	85.06	85.17	83.68	87.01	87.70
	OPERA	86.64	86.62	84.82	83.99	87.16	87.68
	PAI	85.33	83.62	84.20	83.10	86.65	86.99
	VTI	87.36	86.69	85.67	84.48	87.33	87.16
	IRI(ours)	**87.67**	**87.07**	**86.24**	**86.89**	**88.25**	**88.04**
Adv.	Regular	78.96	77.57	81.03	79.30	81.19	82.50
	VCD	80.88	81.33	83.10	82.04	84.80	85.23
	OPERA	81.24	81.38	82.67	79.89	85.20	85.54
	PAI	83.17	81.67	82.19	82.06	84.32	83.68
	VTI	82.57	82.11	83.13	82.16	85.35	84.52
	IRI(ours)	**85.17**	**84.18**	**84.83**	**84.52**	**85.67**	**86.26**

Table 2: Results on POPE tasks. We evaluate the accuracy and F1 Score of various widely used LVLMs.

Method	LLaVA-1.5-7b				Qwen-VL-Chat			
	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Len	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Len
Regular	52.8	15.9	77.3	93.4	2.8	3.0	31.0	5.3
VCD	51.0	14.9	77.2	101.9	1.4	1.2	30.8	4.0
OPERA	45.6	13.1	**78.5**	95.3	1.7	1.3	31.9	4.4
PAI	38.3	12.4	76.9	94.4	1.3	1.2	32.2	4.2
VTI	36.9	12.1	76.8	93.8	1.1	1.1	31.4	4.2
IRI(ours)	**34.6**	**11.5**	78.2	95.8	**1.0**	**0.9**	**32.6**	4.4

Table 3: Results on CHAIR. $Max\ new\ tokens$ is 512. Lower C_S and C_i along with higher recall and length indicate better hallucination mitigating performance.

Implementation Details. In our experiments, we uniformly set $k_1 = 3$ and $k_2 = 10$. For LLaVA-1.5-7B, we use $\lambda_a = 0.26$ and $\lambda_m = 0.16$; for Qwen-VL-Chat, $\lambda_a = 0.20$ and $\lambda_m = 0.10$; and for LLaVA-NeXT, $\lambda_a = 0.15$ and $\lambda_m = 0.08$. Both causal tracing analysis and evaluation experiments of our proposed IRI method are performed on 8 $\times$ NVIDIA A100 SXM 80GB GPUs.

Main Results

Based on the experimental results presented in Tables 1-5, we can draw the following key conclusions:

Method	LLaVA-1.5-7b			Qwen-VL-Chat		
	Score↑	VH.%↓	Hu.%↓	Score↑	VH.%↓	Hu.%↓
Regular	1.86	63.5	67.1	2.93	41.1	61.0
VCD	2.12	54.2	66.7	2.77	39.2	61.5
OPERA	2.15	54.2	63.0	2.94	38.4	58.2
PAI	2.27	53.2	62.5	2.87	39.5	56.7
VTI	2.43	52.2	63.4	2.99	38.4	57.4
IRI(ours)	**2.53**	**50.2**	**62.0**	**3.13**	**37.5**	**56.2**

Table 4: Results on MMHal-Bench and MHumanEval. We use **GPT-4** and **human annotators** as evaluation references.

Model	POPE		MME		CHAIR	
	ACC ↑	F1 ↑	Cog.↑	Hall.%↑	$C_S \downarrow$	$C_I \downarrow$
Qwen2-VL-7B	88.49	87.85	556.4	630.0	24.8	7.2
+ IRI	**89.04**	**88.44**	**563.4**	**663.3**	**14.2**	**6.5**
InternVL2-8B	86.67	85.72	566.4	663.0	37.2	9.4
+ IRI	**87.69**	**86.90**	**569.3**	**688.7**	**30.7**	**8.6**

Table 5: Results on more advanced models. Cog. and Hall. denote the cognitive and hallucination subset of MME.

(1) Robust and SOTA Performance: Our proposed IRI method demonstrates robust, SOTA hallucination mitigation performance across both discriminative and generative tasks. Specifically, on the POPE benchmark, IRI achieves an average improvement of +4.89% in Accuracy and +5.20% in F1 Score. For the MME hallucination subset, IRI brings an average absolute gain of +65.7 points in the Total score. On the CHAIR benchmark, IRI reduces the average hallucination metrics (C_S and C_I) by 6.43 points. Finally, on MMHal-Bench, IRI improves the average Score by +0.44 while lowering the average VH Rate by 8.45%. These results demonstrate the effectiveness and robustness of our approach in mitigating hallucinations.

(2) Model-agnostic and generalizable: IRI is not dependent on specific model architectures and can be readily deployed across various LVLMs. We successfully implemented IRI on more advanced models, such as Qwen2-VL-7B and InternVL2-8B, where it continued to provide steady and significant performance enhancements.

(3) Preserving foundational capabilities: IRI effectively mitigates hallucination without sacrificing LVLM's other foundational capabilities. Specifically, it leads to improved scores on MME cognitive tasks and more informative responses, as indicated by higher scores on MMHal-Bench.

Setting	Hyperparameters				POPE	
	λ_a	λ_m	k_1	k_2	ACC ↑	F1 ↑
LLaVA-1.5-7b	-	-	-	-	78.96	77.57
Ablation of Component						
+ IRI w/o MLP	0.22	-	**3**	**10**	84.93	84.02
+ IRI w/o MHSA	-	0.04	**3**	**10**	85.07	84.03
+ IRI w/ Hidden States	0.18	0.04	**3**	**10**	84.50	83.67
Ablation of Layer Range						
+ IRI w/ Fisrt 10 Layers	0.22	0.14	**3**	**10**	78.46	77.23
+ IRI w/ Last 10 Layers	0.20	0.14	**3**	**10**	80.42	79.87
Ablation of Layer Nums.						
+ IRI w/ Less Source Layers	0.26	0.16	1	**10**	82.92	83.13
+ IRI w/ More Source Layers	0.24	0.14	5	**10**	84.76	83.94
+ IRI w/ Less Target Layers	0.22	0.12	**3**	5	83.32	82.71
+ IRI w/ More Target Layers	0.26	0.16	**3**	15	84.49	83.56
Ablation of RRs						
+ IRI w/o RRs	0.26	0.14	**3**	**10**	84.42	83.92
Ablation of Normalization						
+ IRI w/o Norm.	0.24	0.14	**3**	**10**	85.07	83.98
+ IRI	0.26	0.16	**3**	**10**	**85.17**	**84.18**

Table 6: Result of ablation study on MS-COCO POPE. For each experiment, the parameter λ_a and λ_m is individually optimized to ensure fair comparison.

Ablation Study

As shown in Table 6, to validate the effectiveness of each component within the proposed IRI method and the key findings of FCCT framework, we conducted a comprehensive and systematic set of ablation experiments. We specifically focus on addressing the following five questions:

(1) Why is it necessary to intervene in both MHSA and MLP, but not directly in hidden states? Experimental results show that removing either MHSA or MLP results in a slight performance decrease. Combined aggregation of two modules yields greater improvements, which demonstrates that both MHSAs and MLPs play a critical role in enriching the high-level representations in the middle layers. Furthermore, we also apply interventions to the hidden states based on IRI and observe a performance drop. We believe that hidden states are the cumulative result of MHSA, MLP, and the states from the previous layer. Directly intervening in these highly integrated and functionally specialized hidden states can disrupt the hierarchically constructed flow of semantic information. Consequently, this approach is less effective than precisely enhancing the individual key components: the MHSA, which is responsible for information aggregation, and the MLP, which handles representation processing.

(2) Do the intermediate layers really play a crucial role? Specifically, when IRI's source and target layers are both limited in the first ten or the last ten layers, accuracy drops to 78.46% and 80.42%, respectively. These results align well with our FCCT findings: intermediate layers carry the strongest causal effect for perceiving and aggregating crucial visual&textual object information, whereas shallow layers lack sufficient crucial visual object perception and deep layers already focus on final output generation.

(3) How does the number of source and target layers affect performance? We find that the number of intervention layers affects the effectiveness of IRI to some extent. It is necessary to select a sufficient number of intermediate layers with strong causal effects for visual information perception and inject them into a sufficient number of target layers to make IRI effective. Notably, even when injecting from too many source layers or into too many target layers, IRI's performance does not drop significantly compared to its peak, which demonstrates strong robustness. Nevertheless, selecting too many source layers may introduce noise, while including too many target layers may inject mid-level representations into task-relevant representation for final prediction, preventing optimal performance.

(4) Why is Recovery Rate necessary for more precise injection? The Recovery Rate adaptively controls how much each layer's information is amplified in line with its estimated causal effect. By re-weighting restored activations, it highlights components across layers that contribute more strongly to visual object perception.

(5) Why is normalization necessary for more stable injection? The normalization strategy ensures that the scale of the vector remains consistent before and after injection, preventing undesired magnitude shifts that may distort downstream representations. This mechanism stabilizes the effect of representation injection, thereby enhancing the robustness of the IRI method to distributional shifts.

Inference Latency

As shown in Table 7, IRI achieves the best hallucination mitigating performance while preserves the inference speed.

Method	TTFT(ms)	TPOT(ms)	Acc(%)
LLaVA-1.5-7b	99.8 1.0×	36.0 1.0×	78.96
+ VCD	160.1 1.6×	96.8 2.7×	80.88
+ OPERA	109.8 1.1×	69.5 1.9×	81.24
+ PAI	156.3 1.6×	93.6 2.6×	83.17
+ IRI(ours)	102.2 1.0×	36.5 1.0×	85.17

Table 7: Inference latency (Time to First Token, Time Per Output Token) and the accuracy on adversarial POPE.

Conclusion

In this paper, we introduce the **Fine-grained Cross-modal Causal Tracing (FCCT)** framework and the **Intermediate Representation Injection (IRI)** technique to improve the interpretability and performance of large vision-language models (LVLMs). Our FCCT framework provides a comprehensive, fine-grained causal analysis of the internal components of LVLMs, uncovering key insights into the cross-modal aggregation and hierarchical representation formation, particularly through MHSA and MLP mechanisms. Building on these insights, IRI proves to be a robust and training-free inference-time method, significantly mitigating object hallucinations across various LVLM architectures while maintaining inference speed and foundational model capabilities. Experimental results not only demonstrate the superiority of IRI, but also validate the findings of FCCT.

Acknowledgements

Xiaocheng Feng and Xiachong Feng are the co-corresponding authors of this work. We thank the anonymous reviewers for their insightful comments. This work was supported by the National Natural Science Foundation of China (NSFC) (grant 62522603, 62276078, U22B2059), the Key R&D Program of Heilongjiang via grant 2022ZX01A32, and the Fundamental Research Funds for the Central Universities (XNJKKGYDJ2024013).

References

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

Basu, S.; Grayson, M.; Morrison, C.; Nushi, B.; Feizi, S.; and Massiceti, D. 2024. Understanding information storage and transfer in multi-modal large language models. *arXiv preprint arXiv:2406.04236*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12): 220101.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Qiu, Z.; Lin, W.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; and Ji, R. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *ArXiv*, abs/2306.13394.

Golovanevsky, M.; Rudman, W.; Palit, V.; Singh, R.; and Eickhoff, C. 2024. What Do VLMs NOTICE? A Mechanistic Interpretability Pipeline for Gaussian-Noise-free Text-Image Corruption and Evaluation. *arXiv preprint arXiv:2406.16320*.

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 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*, 13418–13427.

Huo, J.; Yan, Y.; Hu, B.; Yue, Y.; and Hu, X. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 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*, 13872–13882.

Li, Q.; Ye, Z.; Feng, X.; Zhong, W.; Qin, L.; Chen, R.; Li, B.; Jiang, K.; Wang, Y.; Liu, T.; et al. 2025a. CAI: Caption-Sensitive Attention Intervention for Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2506.23590*.

Li, Y.; Cao, Y.; He, H.; Cheng, Q.; Fu, X.; Xiao, X.; Wang, T.; and Tang, R. 2025b. M²IV: Towards Efficient and Fine-grained Multimodal In-Context Learning via Representation Engineering. In *Second Conference on Language Modeling*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J.-R. 2023. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Li, Y.; Yang, J.; Shen, Z.; Han, L.; Xu, H.; and Tang, R. 2025c. CATP: Contextually Adaptive Token Pruning for Efficient and Enhanced Multimodal In-Context Learning. *arXiv preprint arXiv:2508.07871*.

Li, Y.; Yang, J.; Yun, T.; Feng, P.; Huang, J.; and Tang, R. 2025d. Taco: Enhancing multimodal in-context learning via task mapping-guided sequence configuration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 736–763.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge.

Liu, Q.; Mao, J.; and Wen, J.-R. 2025. How do Large Language Models Understand Relevance? A Mechanistic Interpretability Perspective. *arXiv preprint arXiv:2504.07898*.

Liu, S.; Ye, H.; and Zou, J. 2024. Reducing Hallucinations in Vision-Language Models via Latent Space Steering. *arXiv preprint arXiv:2410.15778*.

Liu, S.; Zheng, K.; and Chen, W. 2024. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, 125–140. Springer.

Lyu, X.; Chen, B.; Gao, L.; Song, J.; and Shen, H. T. 2024. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *arXiv preprint arXiv:2405.15356*.

Meng, K.; Bau, D.; Andonian, A.; and Belinkov, Y. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35: 17359–17372.

Neo, C.; Ong, L.; Torr, P.; Geva, M.; Krueger, D.; and Barez, F. 2024. Towards interpreting visual information processing in vision-language models. *arXiv preprint arXiv:2410.07149*.

Palit, V.; Pandey, R.; Arora, A.; and Liang, P. P. 2023. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2856–2861.

Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; and Saenko, K. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.

Salin, E.; Farah, B.; Ayache, S.; and Favre, B. 2022. Are vision-language transformers learning multimodal representations? a probing perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11248–11257.

Sun, Z.; Shen, S.; Cao, S.; Liu, H.; Li, C.; Shen, Y.; Gan, C.; Gui, L.-Y.; Wang, Y.-X.; Yang, Y.; et al. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.

Tang, K.; You, J.; Ge, X.; Li, H.; Guo, Y.; and Huang, X. 2025. Mitigating Hallucinations via Inter-Layer Consistency Aggregation in Large Vision-Language Models. *arXiv preprint arXiv:2505.12343*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Ye, Z.; Li, Q.; Feng, X.; Qin, L.; Huang, Y.; Li, B.; Jiang, K.; Xiang, Y.; Zhang, Z.; Lu, Y.; et al. 2025. CLAIM: Mitigating Multilingual Object Hallucination in Large Vision-Language Models with Cross-Lingual Attention Intervention. *arXiv preprint arXiv:2506.11073*.

You, H.; Zhang, H.; Gan, Z.; Du, X.; Zhang, B.; Wang, Z.; Cao, L.; Chang, S.-F.; and Yang, Y. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*.

Yu, T.; Zhang, H.; Yao, Y.; Dang, Y.; Chen, D.; Lu, X.; Cui, G.; He, T.; Liu, Z.; Chua, T.-S.; et al. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Zhang, Y.; Shen, G.; Ning, K.; Ren, T.; Qiu, X.; Wang, M.; and Kong, X. 2025. Improving Region Representation Learning from Urban Imagery with Noisy Long-Caption Supervision. *arXiv preprint arXiv:2511.07062*.

Zhong, W.; Feng, X.; Zhao, L.; Li, Q.; Huang, L.; Gu, Y.; Ma, W.; Xu, Y.; and Qin, B. 2024. Investigating and Mitigating the Multimodal Hallucination Snowballing in Large Vision-Language Models. arXiv:2407.00569.