CausalLens: Sensitivity-Guided Multi-Head Causal Intervention for Hallucination Mitigation in Large Vision-Language Models

CVPR 2026

Junyang Ji[1,2], Qifan Liu[3], Wenming Yang[1,†], Zhihai He[2,†]

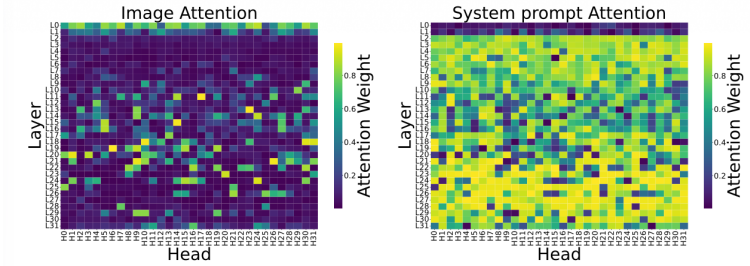
[1]Tsinghua University, [2]Southern University of Science and Technology, [3]Shenzhen Polytechnic University

Contributions

We reveal a structural causal imbalance inside LVLM decoders: **visual signals are strong only in early layers but rapidly fade.**

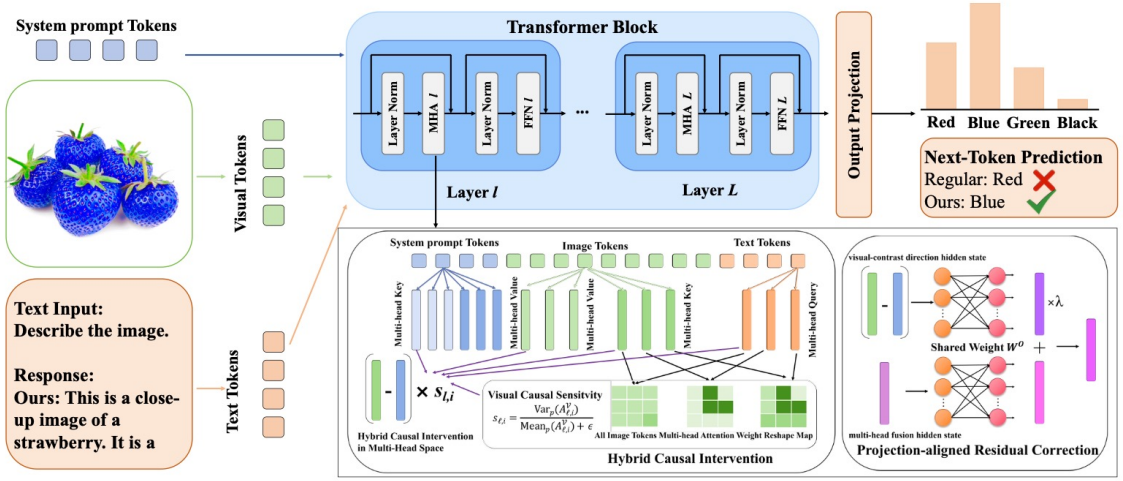
We propose **CausalLens**: a training-free, single-pass intervention modulating decoder hidden states via **sensitivity-guided multi-head decomposition** + **projection-aligned correction.**

Extensive experiments demonstrate **state-of-the-art hallucination mitigation** with negligible overhead — making CausalLens ideal for latency-sensitive applications.



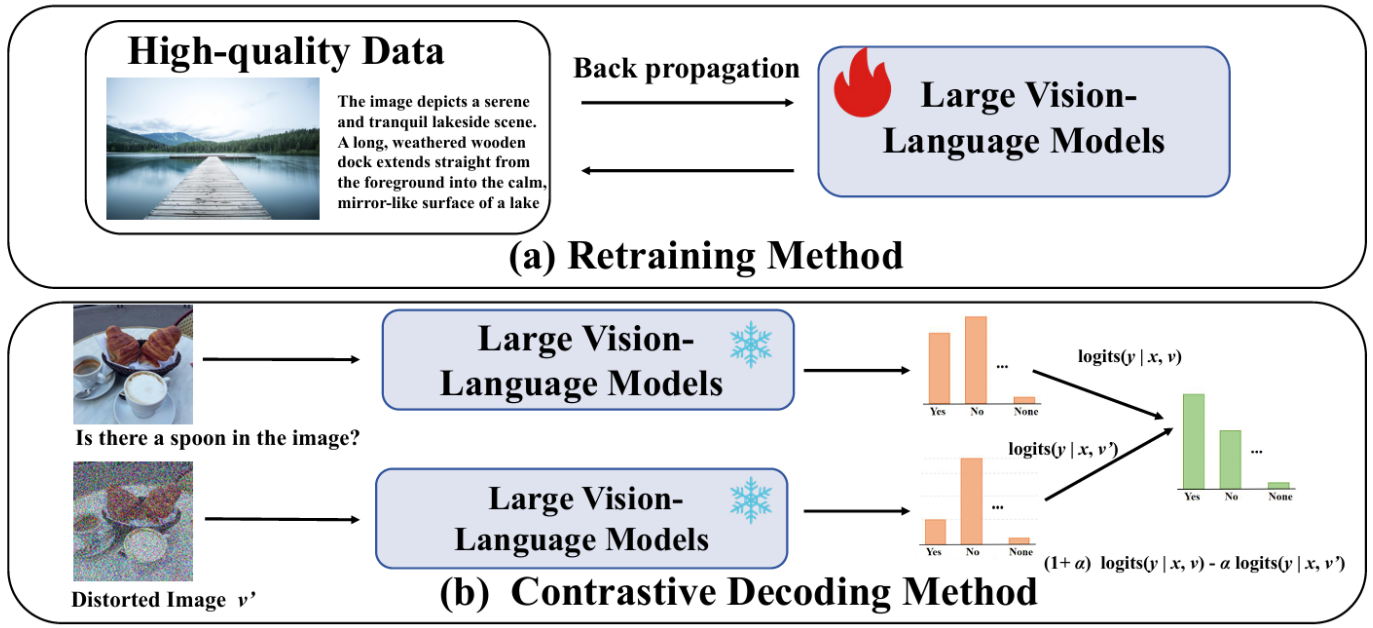
(a) Image Attention across layers and heads.

(b) System Prompt Attention across layers and heads.



The Hallucination Dilemma

- High-performance LVLMs are essential for visual reasoning.
- Critical Fault: Models generate visually ungrounded facts.



- Paradigm 1: Retraining
- High data/compute cost
- Requires gradient access

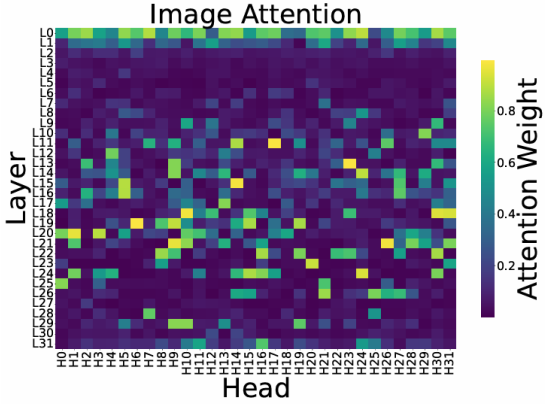
- Paradigm 2: Contrastive decoding
- Multi-pass model queries
- Severe inference latency

Traditional methods require heavy training **multiple forward passes** during decoding.

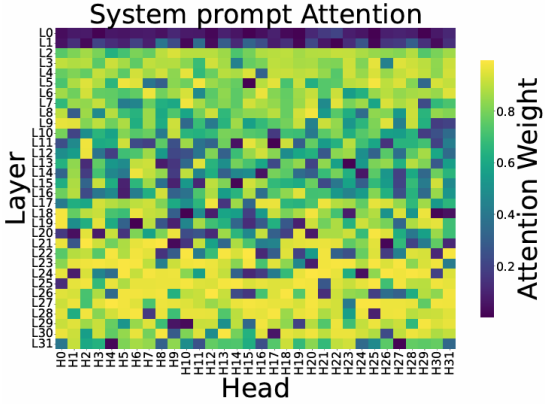
Why Do LVLMs Hallucinate?
An Internal Causal Perspective on Attention Distribution

CVPR JUNE 3-7, 2026 DENVER COLORADO

- Layer-wise Attention Decay
- Textual priors dominate deep layers (60%-80% weight concentration)



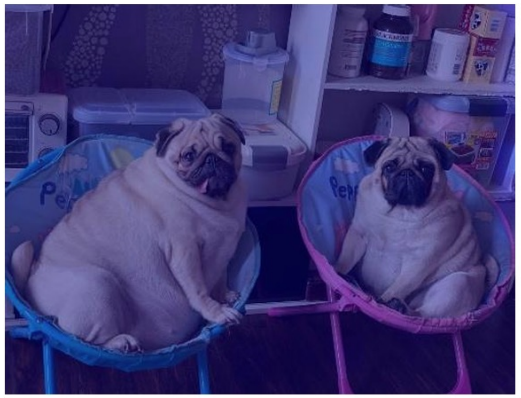
(a) Image Attention across layers and heads.



(b) System Prompt Attention across layers and heads.

- Head-level Sensitivity Divergence
- High-s heads are vital causal carriers (Ablation triggers severe drop)

(a) Head A: concentrated visual attention on target objects (chairs).

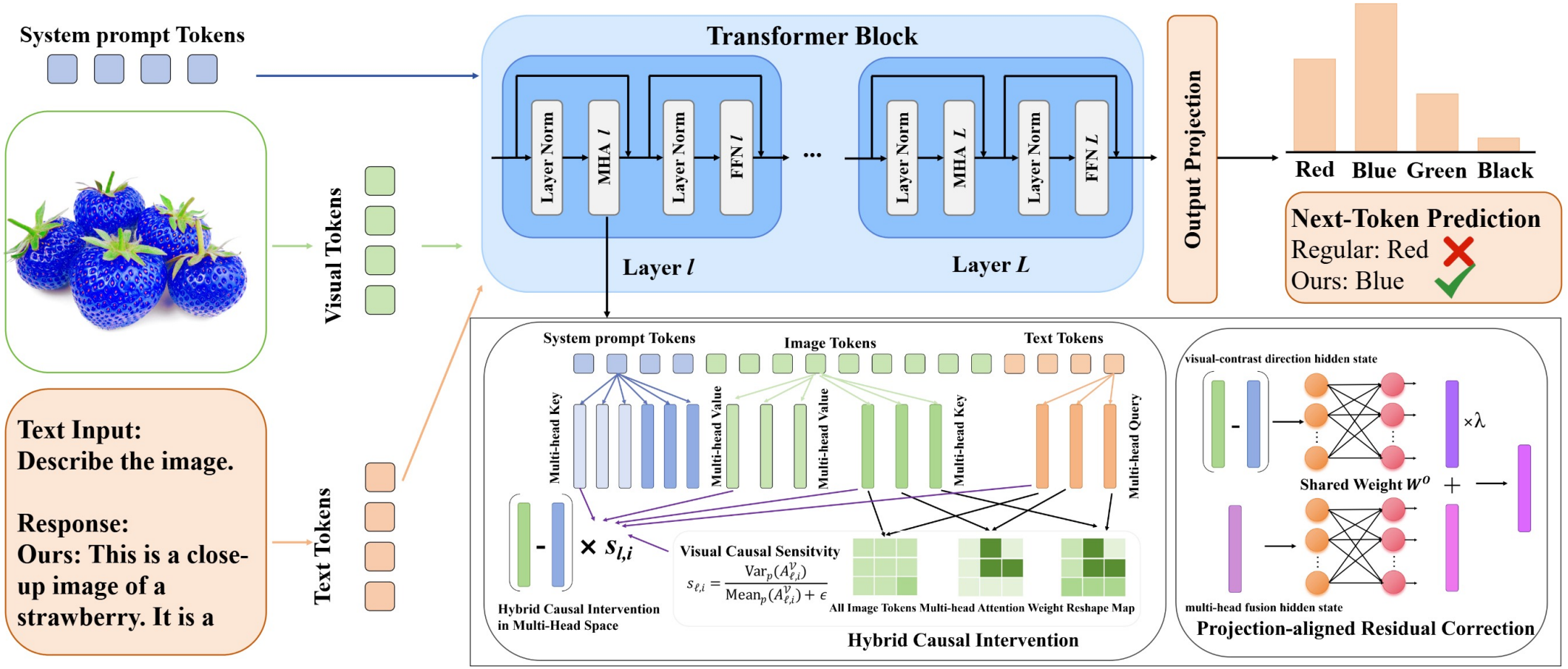


(b) Head B: diffuse attention uniformly spread across the image.

💡 Conclusion: Modulating these specific mid-layer hidden states restores the visual pathway.

- Visual information is *strongly expressed* only in carly layers and *quickly weakens* during decoding.
- Mid-to-late layers become *dominated by system prompts* and textual priors.
- This imbalance *suppresses the causal pathway* from visual tokens to generated outputs, leading to hallucination.

Methodology: CausalLens Framework



✓ Decompose each attention head into *system*, *visual*, and *textual* pathways.

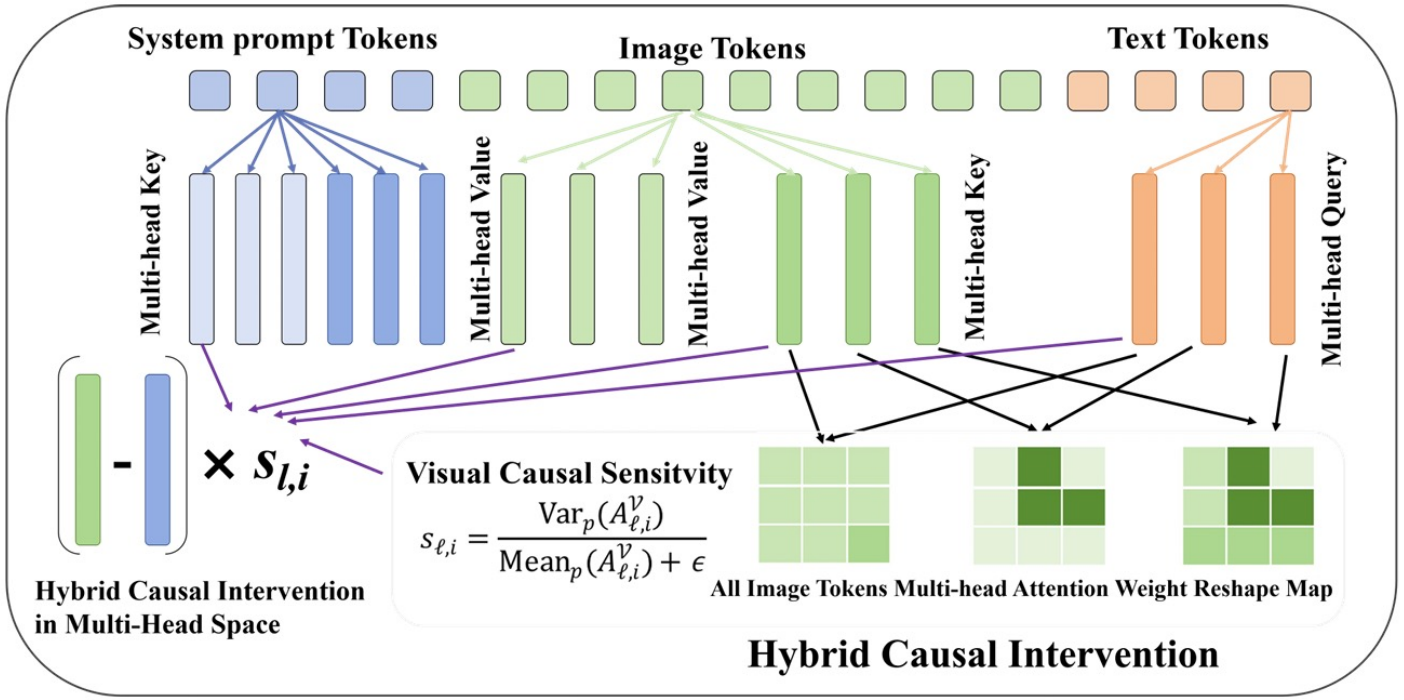
✓ Amplify *visually reliable* heads using a sensitivity-guided hybrid causal intervention.

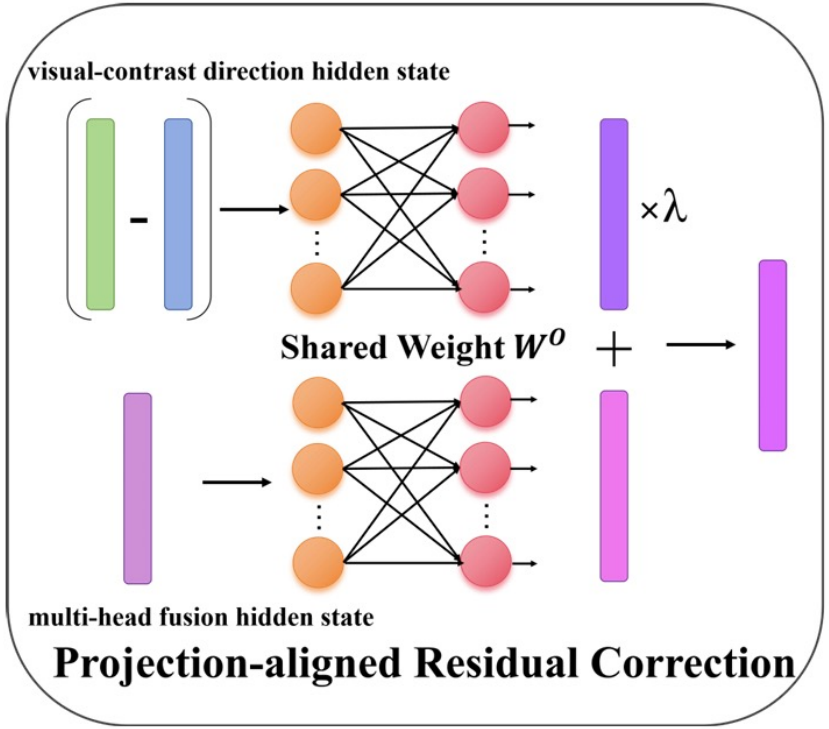
✓ Preserve the intervention after head fusion using *projection-aligned residual correction*.

Sensitivity-Weighted Mid-layer Hybrid Causal Intervention

Restoring the visual pathway by modulating pathway-specific hidden states.

- Multi-Head Pathway Decomposition
- Visual Causal Sensitivity Scoring
- Mid-Layer Hybrid Intervention (L10-L20)
- Adaptive Contrast Shifting (Visual vs. Prompt)





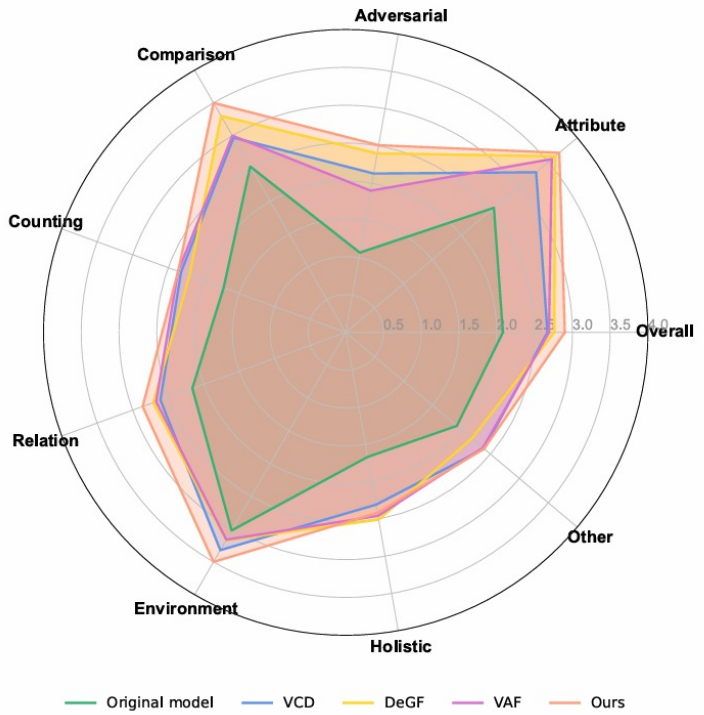
💡 Conclusion: This single-pass intervention **restores the causal chain from visual tokens to outputs**.

Experimental Results & Analysis

CausalLens generalises robustly across diverse benchmarks and different LVLM backbones.

- **State-of-the-Art Mitigation Performance**
- **Consistently Beats VCD, DeGF, and VAF**
- **Broad Spectrum Hallucination Suppression**

Category	Method	LLaVA-v1.5-7B		LLaVA-v1.5-13B		Qwen2-VL-7B	
		Accuracy	F1-score	Accuracy	F1-score	Accuracy	F1-score
Random	Regular	85.9	85.2	87.3	87.1	88.8	88.1
	VCD(CVPR2024)	88.8	88.7	88.8	88.7	90.6	90.0
	DeGF (ICLR2025)	89.3	89.2	90.1	90.3	90.1	89.9
	VAF(CVPR2025)	89.6	89.3	90.1	89.9	90.5	89.8
	Ours	**90.6**	**90.4**	**90.9**	**90.5**	**91.4**	**91.0**
Popular	Regular	82.3	82.1	84.0	83.6	86.6	86.1
	VCD(CVPR2024)	84.2	84.6	87.0	87.1	88.2	87.8
	DeGF (ICLR2025)	84.6	84.2	86.7	87.1	87.9	87.6
	VAF(CVPR2025)	84.8	84.9	87.3	87.1	88.7	88.1
	Ours	**86.5**	**86.8**	**88.8**	**88.6**	**89.0**	**88.9**
Adversarial	Regular	77.9	78.6	79.7	80.0	82.6	82.7
	VCD (CVPR2024)	79.1	81.2	81.7	81.9	84.7	84.8
	DeGF (ICLR2025)	80.3	80.2	82.4	82.1	83.6	83.4
	VAF (CVPR2025)	80.1	81.0	82.7	82.5	**84.9**	84.8
	Ours	**81.6**	**82.8**	**83.9**	**84.5**	84.7	**85.2**



CausalLens consistently improves hallucination benchmarks across multiple LVLMs with negligible overhead.

Qualitative Results and Efficient Analysis

A **plug-and-play, lightweight solution** perfectly suited for **latency-sensitive scenarios.**

Visual Input

Query:
Describe this photo in detail.

Regular Decoding GPT-4V-Aided Evaluation Accuracy: 4 Detailedness: 3

The image features a dining table with a variety of food items, including a large bowl of soup, a plate of noodles, and a bowl of vegetables. The soup is placed in the center of the table, while the noodles and vegetables are positioned around it. \n\nIn addition to the main dishes, there are several bowls and cups scattered across the table, possibly containing other food items or beverages. A spoon is also visible on the table, likely used for eating the soup. The table setting creates an inviting atmosphere for a meal.

Our Approach GPT-4V-Aided Evaluation Accuracy: 7 Detailedness: 5

The image features a dining table with a variety of food items and bowls. There is a large bowl of soup, possibly ramen, placed in the center of the table. The soup is accompanied by a spoon. \n\nIn addition to the soup, there are several bowls containing different types of food, including a bowl of vegetables. A cup can also be seen on the table, possibly containing a beverage to accompany the meal. The table setting creates an inviting atmosphere for a delicious meal.

- **Recovers Key Semantic Cues (e.g., "Ramen")**

Method	Avg. Latency ↓	GPU Memory ↓
Regular	0.281 s (×1.00)	15898 MB (×1.00)
VCD	0.482 s (×2.01)	16744 MB (×1.05)
DeGF	1.143 s (×4.07)	19327 MB (×1.22)
VAF	0.292 s (×1.03)	15916 MB (×1.00)
Ours	0.293 s (×1.04)	16111 MB (×1.01)

- **Training-Free & Single-Forward Pass**
- **Negligible Memory Increase (<1.3)**

- **Latency: On par with VAF, 1.6x faster than VCD**
- **Deployment: Perfect for real-time scenarios**

Conclusion & Future Work

CVPR JUNE 3-7, 2026 DENVER COLORADO

Key Summary:

- CausalLens provides a **paradigm-shifting, training-free solution** for grounded and efficient LVLM reasoning.
- Proposed **CausalLens**, a **plug-and-play, training-free framework** utilizing **sensitivity-guided multi-head intervention and projection-aligned correction**.
- Achieved state-of-the-art hallucination mitigation across major benchmarks while maintaining a **near-zero latency and memory overhead** (only 1.04x latency).

Future Horizons:

- Exploring the extension of CausalLens to a **wider range of next-generation Video-LLMs** and Large Multimodal Models (LMMs).
- Adapting the lightweight single-pass mechanism for resource-constrained edge devices and **real-time robotics vision**.