



HALLUCINATION SALIENCY DROPS

BEGINS WHERE

**Xiaofeng Zhang^{*1}, Yuanchao Zhu^{*1}, Chaochen Gu^{1‡}, Xiaosong Yuan², Qiyan Zhao¹
Jiawei Cao¹, Feilong Tang³, Sinan Fan², Yaomin Shen¹, Chen Shen², Hao Tang⁴**

¹Shanghai Jiaotong University ²Alibaba Group ³Monash University

⁴Peking University

^{*}Equal contribution [‡]Corresponding author

framebreak@sjtu.edu.cn

ICLR 2026

Background

- 视觉语言模型（LVLMs）在跨模态任务中取得了显著进展。
- 关键挑战：在视觉问答和图像描述中经常出现“幻觉”现象（即模型输出与事实不符）。
- 现状：一部分研究者通常依赖前向传播的注意力权重来分析模型。
- 局限性：注意力图仅反映模型在“看”哪里，无法区分正确输出与幻觉输出。

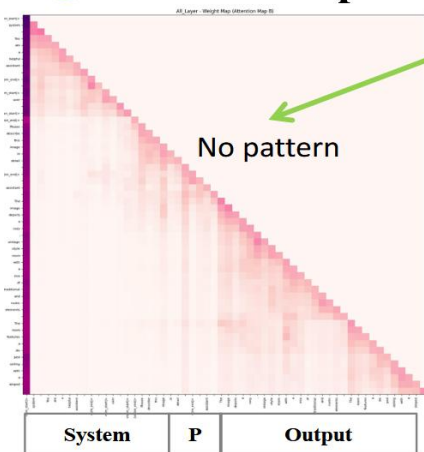


Please describe this image in detail

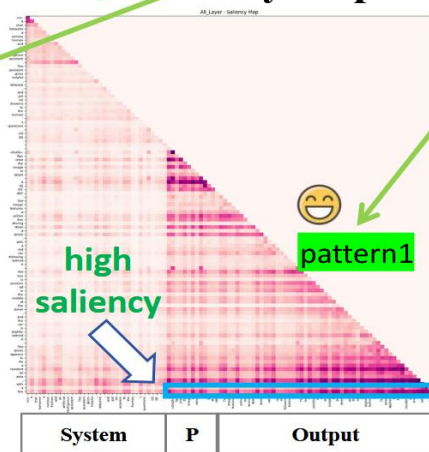


The image depicts a cozy, vintage-style room with a mix of traditional and rustic elements. The room features a sloped ceiling with a striped wallpaper pattern in shades of blue, white, and red. The walls are also adorned with similar striped wallpaper.

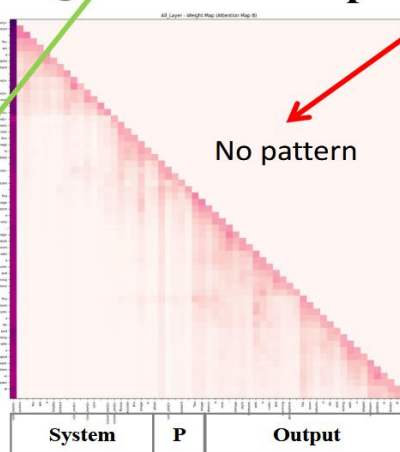
✓ Attention map



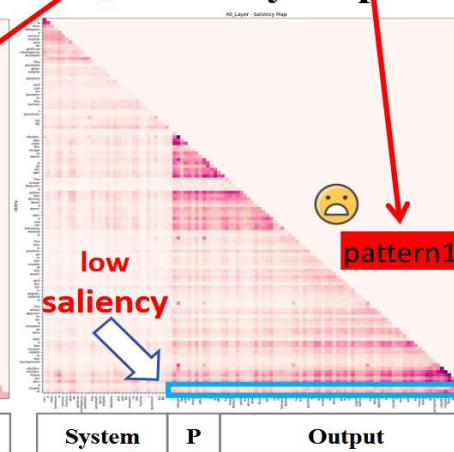
✓ Saliency map



✗ Attention map



✗ Saliency map



Background

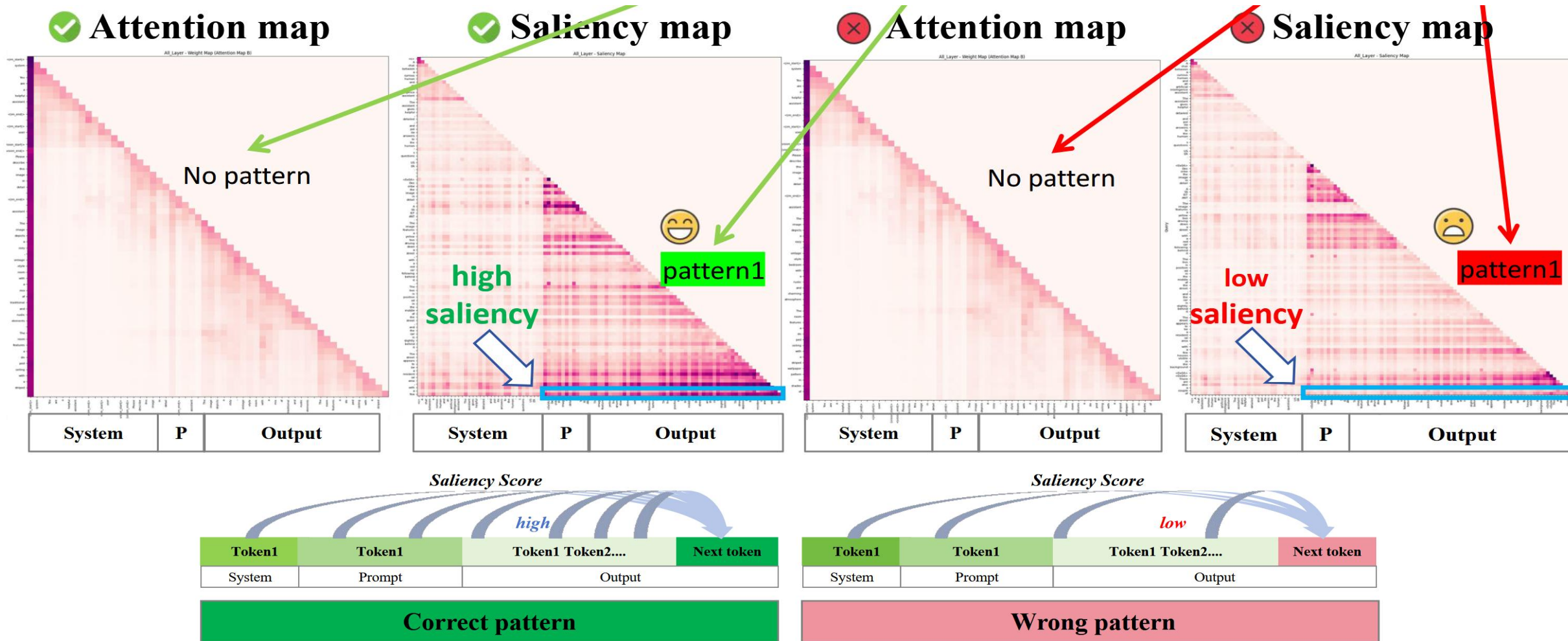


Figure 1: Attention vs. Saliency Maps for Correct and Hallucinated Tokens (Qwen2-VL-7B). Left (correct token **wallpaper**): Attention maps show no distinctive pattern, while our LVLMS-Saliency maps reveal strong, structured grounding to prior outputs. Right (hallucinated token **blue**): Attention maps remain visually similar, but saliency maps collapse, signaling loss of contextual dependency.

前向传播与损失函数:

$$(y, \{\mathbf{A}^{(l,h)}\}_{l=1, h=1}^{L,H}, s) = \mathcal{M}(x)$$

$$\mathcal{L}(y, s) = - \sum_{t=1}^T y_t \log \sigma(s_t)$$

显著性计算 (核心创新):

$$\nabla \mathbf{A}^{(l,h)} = \frac{\partial \mathcal{L}}{\partial \mathbf{A}^{(l,h)}} \in \mathbb{R}^{n \times n}$$

$$\mathbf{S}^{(l,h)} = \text{tril} \left(\left| \mathbf{A}^{(l,h)} \odot \nabla \mathbf{A}^{(l,h)} \right| \right)$$

层级聚合:

$$\bar{\mathbf{S}}^{(l)} = \frac{\sum_{h=1}^H \mathbf{S}^{(l,h)}}{\left\| \sum_{h=1}^H \mathbf{S}^{(l,h)} \right\|_2}$$

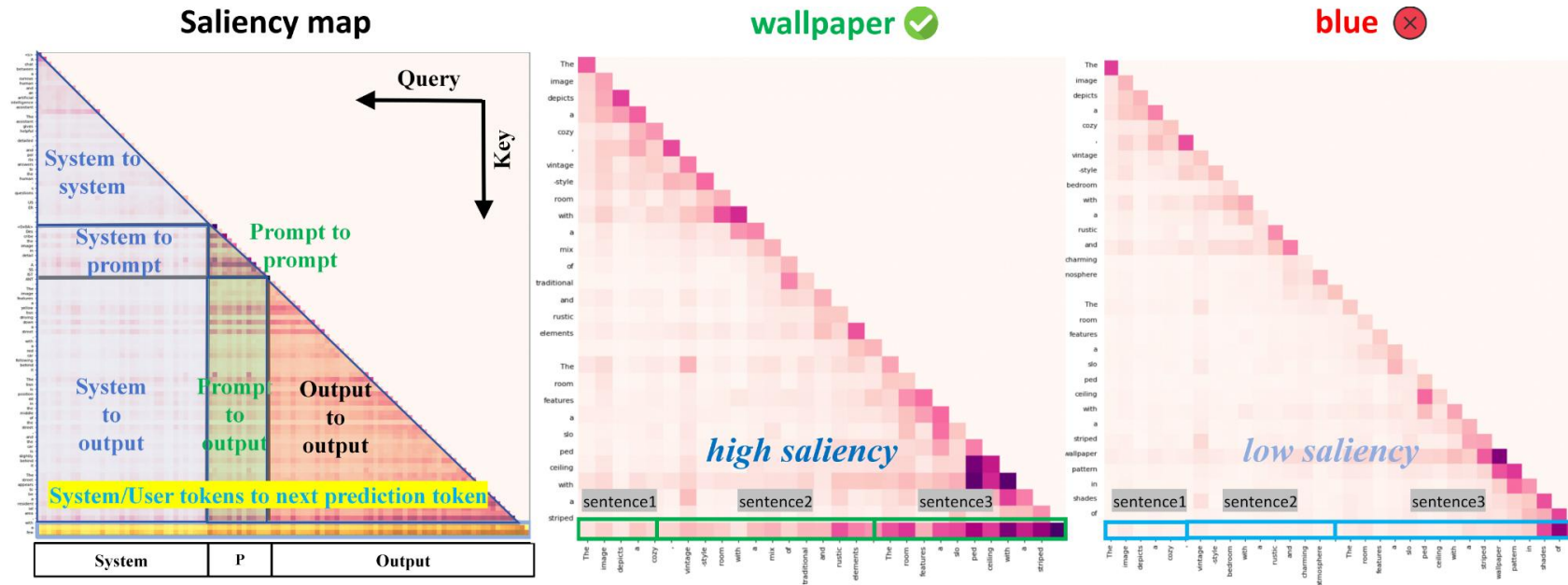
A (Attention) : 前向传播输出的“静态观察权重”。

∇A (Gradient): 损失函数对注意力的偏导数, 代表“实际影响”。

S (Saliency): 结合“观察”与“影响”的综合得分。

\odot 与 tril : 哈达玛乘积与下三角掩码。

Pattern: Hallucinations occur when prior output tokens shows low saliency to the next token.



sentence1 The image depicts a cozy,
 sentence2 vintage-style room with a mix of traditional and rustic elements.
 sentence3 The image features a sloped ceiling with a striped **wallpaper** pattern in shades of **blue**.

high saliency *low saliency*

正确输出：对近期 Token 保持高显著性（高依赖）。

幻觉输出：模型对过去上下文的依赖关系崩溃。

Figure 2: **Output Token Saliency Patterns in Qwen2-VL-7B.** When generating a correct token (e.g., **wallpaper**), the current token assigns high saliency to recent output tokens, typically decaying with distance. In contrast, when generating a hallucinated token (e.g., **blue**), saliency toward all prior outputs collapses — signaling contextual disconnection.

Algorithm 1 SGRS

Require: $\mathcal{M}, \mathbf{x}, K, R, \alpha, W, \mathcal{L}, S=35, I=576, H$
Ensure: x_P : accepted token at position P

- 1: $\text{logits} \leftarrow \mathcal{M}(\mathbf{x}_{\text{input}}, \mathbf{KV})[:, -1, :]$
- 2: $\mathcal{C} \leftarrow \text{TopK}(\text{softmax}(\text{logits}), K)$, $\text{accepted} \leftarrow \text{False}$
- 3: **for** $r = 1$ to R **do**
- 4: $c \sim \text{Sample}(\mathcal{C})$
- 5: $\mathcal{S}(c) \leftarrow \text{SALIENCY}(\mathcal{M}, c, \mathcal{L}_{\text{target}}, P, S, I) \triangleright$
 Eq. (1)
- 6: $\mathcal{J}_P \leftarrow \{j \mid S + I \leq j < P\}$ \triangleright Output token positions
- 7: $\mathcal{H}_P \leftarrow \{j \in \mathcal{J}_P \mid (P - 1) - j \leq W\}$ \triangleright
 Recent W outputs
- 8: $\tau \leftarrow \alpha \cdot \frac{1}{|\mathcal{H}_P|} \sum_{j \in \mathcal{H}_P} H[j]$ \triangleright Eq. (2)
- 9: **if** $\mathcal{S}(c) \geq \tau$ **then**
- 10: $x_P \leftarrow c$, $H.\text{append}(\mathcal{S}(c))$, $\text{accepted} \leftarrow$
 True, **break**
- 11: **else**
- 12: $\mathcal{C} \leftarrow \mathcal{C} \setminus \{c\}$
- 13: **end if**
- 14: **end for**
- 15: **if not accepted then**
- 16: $x_P \leftarrow \arg \max_{c \in \text{original } \mathcal{C}} \mathcal{S}(c)$ \triangleright Fallback:
 best saliency
- 17: **end if**
- 18: **return** x_P

定位: 在模型正式输出 Token 前进行拦截、评估与过滤。

逻辑: 计算每个候选 Token 与上下文的显著性得分。

动作: 如果得分低于动态阈值, 则视为与上下文脱节 (可能是幻觉), 直接拒绝并重新采样。

$$\mathcal{S}(c_i) = \frac{1}{|\mathcal{L}_{\text{target}}| \cdot |\mathcal{J}|} \sum_{l \in \mathcal{L}_{\text{target}}} \sum_{j \in \mathcal{J}} \bar{\mathcal{S}}_{P,j}^{(l)},$$

$$\tau^{(P)} = \alpha \cdot \frac{1}{|\mathcal{H}|} \sum_{j \in \mathcal{H}} \mathcal{S}(x_j), \quad \mathcal{H} = \{j \in \mathcal{J} \mid (P - 1) - j \leq W\},$$

Algorithm 2 LOCORE

Require:

- 1: $\mathbf{A}^{(P+1)} \in \mathbb{R}^{B \times n_h \times (P+1) \times (P+1)}$: attention weights for step $P + 1$
- 2: $S = 35, I = 576$: system and image token lengths
- 3: w_s : local window size, $\beta \geq 0$: gain strength

Ensure: $\mathbf{A}^{(P+1)}$: modified attention weights for step $P + 1$

- 4: $P \leftarrow$ current position \triangleright Last generated token position
 - 5: $t \leftarrow P - (S + I)$
 - 6: **if** $t \leq 0$ **then return** $\mathbf{A}^{(P+1)}$
 - 7: **end if** \triangleright No output yet
 - 8: $\mathcal{J}_P \leftarrow \{j \mid S + I \leq j < P\}$ \triangleright Historical output positions
 - 9: **if** $\mathcal{J}_P = \emptyset$ **then return** $\mathbf{A}^{(P+1)}$
 - 10: **end if**
 - 11: **for all** $j \in \mathcal{J}_P$ **do**
 - 12: $d_j \leftarrow P - j$ \triangleright Distance to current position
 - 13: $\gamma_j \leftarrow 1 + \beta \cdot \mathbb{I}(d_j \leq w_s)$ \triangleright Eq. (3)
 - 14: **for all** $b \in [B], h \in [n_h]$ **do**
 - 15: $\mathbf{A}^{(P+1)}[b, h, P + 1, j] \leftarrow \mathbf{A}^{(P+1)}[b, h, P + 1, j] \cdot \gamma_j$ \triangleright Eq. (4)
 - 16: **end for**
 - 17: **end for**
 - 18: **return** $\mathbf{A}^{(P+1)}$
-

逻辑：人为地、强制性地给当前词与最近输出词之间的注意力权重，乘上一个大于 1 的增益系数

$$\mathcal{J}_P = \{j \in \mathbb{N} \mid \text{Sys}_L + \text{Img}_L \leq j < P\}$$

$$\gamma_j^{(P)} = 1 + \beta \cdot \mathbb{I}((P - j) \leq w_s) \quad j \in \mathcal{J}_P$$

$$\mathbf{A}^{(P+1)}[b, h, P + 1, j] \leftarrow \mathbf{A}^{(P+1)}[b, h, P + 1, j] \cdot \gamma_j^{(P)}$$

$$\forall b \in [B], h \in [n_h], j \in \mathcal{J}_P$$

$$\mathbf{A}_{P+1, \mathcal{J}_P}^{(P+1)} \leftarrow \mathbf{A}_{P+1, \mathcal{J}_P}^{(P+1)} \odot \gamma^{(P)}$$

Table 1: Compare results of LocoRE with other SOTA methods on POPE, CHAIR and MME datasets. The best performances within each setting are **bolded**, baseline: LLaVA-1.5-7B.

Method	Venue	POPE		CHAIR				MME			Total↑	
		F1↑	Acc↑	C _S ↓	C _I ↓	Recall↑	length	Exist.↑	Count↑	Pos.↑		Color↑
Beam Search	-	85.4	84.0	51.0	15.2	75.2	102.2	175.67	124.67	114.00	151.00	565.34
Dola Chuang et al. (2023)	ICLR 2024	80.2	83.1	57.0	15.2	78.2	97.5	180.10	127.40	119.30	154.60	594.10
VCD Leng et al. (2024)	CVPR 2024	85.3	85.0	51.0	14.9	77.2	101.9	184.66	137.33	128.67	153.00	603.66
OPERA Huang et al. (2024)	CVPR 2024	84.2	85.2	47.0	14.6	78.5	95.3	180.67	133.33	111.67	123.33	549.00
DOPRA Wei & Zhang (2024)	MM 2024	84.6	84.3	46.3	13.8	78.2	96.1	185.67	138.33	120.67	133.00	577.67
HALC Chen et al. (2024c)	ICML 2024	83.9	84.0	50.2	12.4	78.4	97.2	190.00	143.30	128.30	160.00	621.60
CCA-LLaVA Xing et al. (2024)	NeurIPS 2024	86.4	86.5	43.0	11.5	80.4	96.6	190.00	148.33	128.33	153.00	641.66
RITUAL Woo et al. (2024)	Arxiv 2024	85.2	84.3	45.2	13.2	78.3	99.2	187.50	139.58	125.00	164.17	616.25
EAH Zhang et al. (2024a)	EMNLP 2025	85.7	86.0	36.4	9.9	74.9	97.7	190.00	108.33	145.00	160.66	603.99
SID Huo et al. (2025)	ICLR 2025	85.6	85.8	44.2	12.2	73.0	99.4	183.90	132.20	127.80	155.90	599.80
TAME Tang et al. (2025a)	ICLR 2025	85.4	85.7	41.3	12.2	74.4	98.8	193.00	137.33	139.00	164.67	634.00
Vissink Kang et al. (2025)	ICLR 2025	86.0	86.5	52.4	14.5	79.1	103.0	190.00	148.33	138.33	155.00	631.33
CausalLLM Zhou et al. (2025)	ICLR 2025	86.0	86.5	-	-	-	-	195.00	156.00	135.00	170.00	656.00
AGLA An et al. (2024)	CVPR 2025	84.6	85.5	43.0	14.1	78.9	98.8	195.00	153.89	129.44	161.67	640.00
FarsightTang et al. (2025b)	CVPR 2025	-	-	41.6	13.2	75.5	100.6	-	-	-	-	-
MemVR Zou et al. (2024)	ICML 2025	87.1	<u>87.4</u>	46.6	13.0	80.8	99.6	190.00	155.00	133.33	170.60	648.30
ONLY Wan et al. (2025)	ICCV 2025	85.5	85.1	49.8	14.3	75.9	99.7	191.67	145.55	136.66	161.66	635.55
Reverse-VLM Wu et al. (2025b)	NeurIPS 2025	-	-	35.3	9.3	75.2	70.4	-	-	-	-	-
LocoRE	-	86.9	87.3	38.4	11.2	75.4	98.2	190.00	158.33	133.33	175.00	656.66
SGRS + LocoRE	-	<u>87.0</u>	87.5	35.6	8.2	75.4	98.2	195.00	158.33	140.00	175.00	668.33

Experiments — GENERATION BENCHMARK

Table 2: Comparison of different LVLMs and LocoRE across all image benchmarks. Notably, in the Hallucination Benchmark, lower scores on CHAIR_I and CHAIR_S indicate better performance, while higher scores are preferable for other metrics.

Method	Comprehensive Benchmark		General VQA		Hallucination Benchmark				
	LLaVA ^W	MM-Vet [↑]	VizWiz [↑]	SQA [↑]	CHAIR _S ↓	CHAIR _I ↓	POPE-R [↑]	POPE-F1 [↑]	POPE-A [↑]
LLaVA-1.5-7B	72.5	30.5	48.5	65.5	48.0	13.9	87.0	85.4	84.0
+ICD	69.7	30.4	46.9	62.8	47.7	13.6	87.9	84.9	84.0
+VCD	70.9	29.5	43.4	63.3	46.8	13.2	87.0	85.3	85.0
+OPERA	72.0	31.4	50.0	64.9	45.2	12.7	88.8	84.2	85.2
+SID	73.4	31.2	50.9	67.8	44.2	14.0	89.4	85.6	85.8
+TAME	73.9	30.5	51.6	66.0	41.3	12.2	88.9	85.4	85.7
+Vissink	74.1	33.5	53.8	67.0	52.4	14.5	87.7	84.9	85.8
+FarSight	74.7	32.5	50.8	67.4	41.6	13.2	90.5	85.5	85.8
+LocoRE	74.8 (+2.3)	33.8 (+3.3)	54.8 (+6.3)	67.5 (+2.0)	38.4 (+9.6)	10.2 (+3.7)	89.5 (+2.5)	86.9 (+1.5)	87.3 (+3.3)
+SGRS+LocoRE	76.7 (+4.2)	36.0 (+5.5)	54.9 (+6.4)	67.8 (+2.3)	35.6 (+12.4)	8.2 (+5.7)	89.8 (+2.8)	87.0 (+1.6)	87.5 (+3.5)
LLaVA-1.5-13B	72.5	36.1	60.5	71.6	47.2	13.6	82.5	86.6	87.2
+ LocoRE	74.0 (+1.5)	38.4 (+2.3)	62.1 (+1.6)	72.5 (+0.9)	43.8 (+3.4)	12.8 (+0.8)	87.8 (+5.3)	87.7 (+1.1)	87.4 (+0.2)
SGRS + LocoRE	76.8 (+4.3)	42.0 (+5.9)	64.0 (+3.5)	75.5 (+3.4)	39.8 (+7.4)	8.8 (+4.8)	88.0 (+5.5)	88.1 (+1.5)	87.6 (+0.4)
Intern-VL-7B	51.6	31.2	51.7	66.2	46.6	12.4	80.0	85.3	86.2
+ LocoRE	52.8 (+1.2)	33.7 (+2.5)	54.5 (+2.8)	66.4 (+0.2)	40.2 (+6.4)	10.5 (+1.9)	85.8 (+5.8)	87.2 (+1.9)	87.3 (+1.1)
SGRS + LocoRE	55.5 (+3.9)	35.0 (+5.0)	56.2 (+4.5)	67.9 (+1.7)	34.4 (+12.2)	7.5 (+3.9)	86.0 (+6.0)	87.6 (+2.3)	87.7 (+1.5)
Intern-VL-13B	53.2	33.7	47.4	70.1	45.4	12.7	82.8	86.4	86.9
+ LocoRE	54.1 (+0.9)	35.4 (+1.7)	50.1 (+2.7)	70.4 (+0.3)	43.6 (+1.8)	12.5 (+0.2)	86.3 (+3.5)	87.2 (+0.8)	87.3 (+0.4)
SGRS + LocoRE	56.8 (+3.6)	37.3 (+3.6)	52.0 (+4.6)	71.0 (+0.9)	45.2 (+3.4)	14.0 (+2.7)	87.0 (+4.2)	88.1 (+1.7)	88.8 (+1.9)
Qwen2-VL-7B	75.6	63.2	57.3	74.1	25.0	7.3	79.1	86.6	87.6
+ LocoRE	77.8 (+2.2)	64.8 (+1.6)	59.4 (+2.1)	74.2 (+0.1)	23.5 (+1.5)	6.8 (+0.5)	81.3 (+2.2)	87.5 (+0.9)	88.2 (+0.6)
SGRS + LocoRE	79.7 (+4.1)	67.7 (+4.5)	60.3 (+3.0)	75.3 (+1.2)	19.3 (+5.7)	5.1 (+2.2)	82.6 (+3.5)	88.0 (+1.4)	89.0 (+1.4)
Qwen2.5-VL-7B	76.8	62.2	60.9	79.0	27.2	9.0	80.4	87.4	88.4
+LocoRE	77.9 (+1.1)	64.8 (+2.6)	61.6 (+0.7)	80.8 (+1.8)	23.0 (+4.2)	8.5 (+0.5)	80.9 (+0.5)	87.8 (+0.4)	88.7 (+0.3)
SGRS +LocoRE	80.0 (+3.2)	66.2 (+4.0)	62.7 (+1.8)	82.1 (+3.1)	21.0 (+6.2)	6.5 (+2.5)	81.5 (+0.5)	88.3 (+0.9)	89.5 (+1.1)
Qwen2.5-VL-32B	81.2	72.2	70.8	89.0	43.6	9.5	79.1	86.7	87.8
+LocoRE	82.7 (+0.5)	73.1 (+0.9)	71.2 (+0.4)	89.3 (+0.3)	41.8 (+1.8)	8.5 (+1.0)	79.5 (+0.4)	86.9 (+0.2)	88.0 (+0.2)

ABLATION STUDY

使用 LocoRE 后，显著性图重新出现了清晰的结构化模式
LocoRE 单独使用时，延迟增加小于 2%，是性价比极高的选择

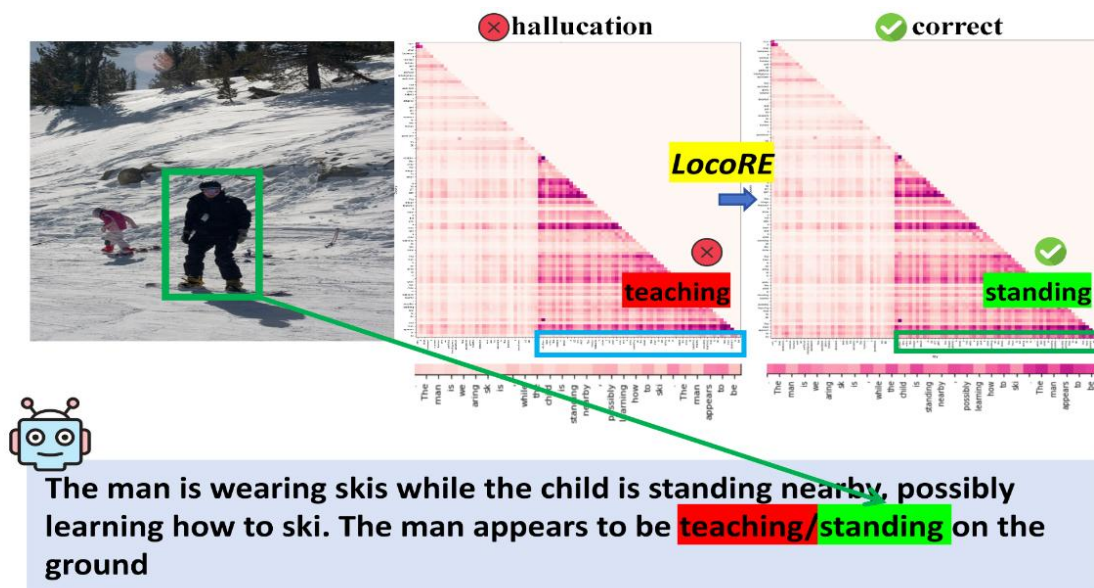
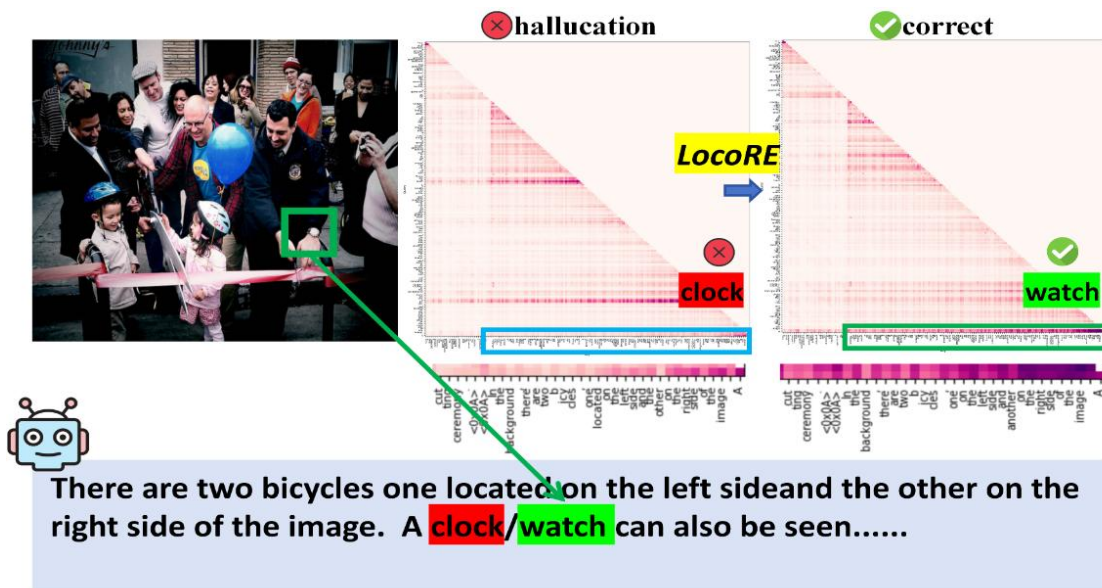


Figure 3: **Effect of LocoRE on output token saliency map (Qwen2-VL-7B).** *Without LocoRE:* When generating an incorrect token(**clock**), saliency scores assigned to prior output tokens are low — indicating weak contextual grounding. *With LocoRE:* The same position now generates a correct token(**watch**), accompanied by significantly higher saliency scores to recent outputs — demonstrating LocoRE’s ability to restore contextual coherence and prevent hallucination via attention reinforcement.

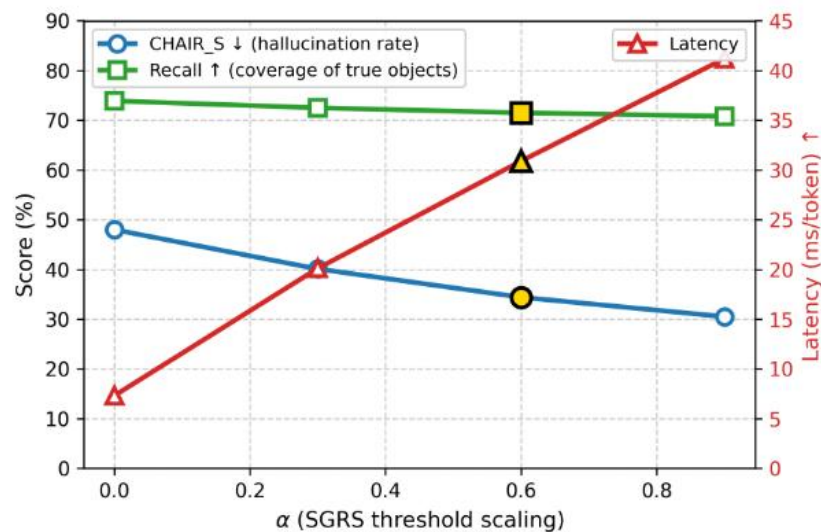


Figure 4: Ablation study of α : trade-offs between hallucination rate, recall, and latency.

α	β	LLaVA-1.5				Qwen2-VL-7B			
		CHAIR		POPE		CHAIR		POPE	
		S \downarrow	I \downarrow	F1 \uparrow	Acc \uparrow	S \downarrow	I \downarrow	F1 \uparrow	Acc \uparrow
0.0	0.0	48.0	13.9	85.4	84.0	25.0	7.3	86.6	87.6
0.0	0.15	38.4	10.2	86.9	87.3	—	—	—	—
0.0	0.20	—	—	—	—	23.5	6.8	87.5	88.2
0.6	0.0	36.5	9.0	86.9	87.4	20.5	5.6	87.9	88.9
0.6	0.15	35.6	8.2	87.0	87.5	—	—	—	—
0.6	0.20	—	—	—	—	19.3	5.1	88.0	89.0
0.6	1.0	50.2	20.9	60.3	57.8	37.5	18.5	55.3	54.6

Table 3: Ablation study on α (SGRS) and β (LocoRE). Best in bold. β : 0.15 (LLaVA-1.5), 0.20 (Qwen2-VL).

Thanks