Object Hallucination Mitigation in Large Vision-Language Models via Self-Vision Dual Masking and Uncertainty-Triggered Assembly

Ziji Sheng[1,2]*, **Guiyao Tie**[1,2]*, **Weidong Wang**[1,2]*, **Jiawen Shi**[1,2]*, **Junhao Dong**[3], **Xiaoye Qu**[4], **Shanshan Ye**[5], **Daizong Liu**[6]†‡, **Dengpan Ye**[6], **Pan Zhou**[1,2]‡, **Jing Zhang**[6]

[1]Hubei Key Laboratory of Distributed System Security,
Hubei Engineering Research Center on Big Data Security,
School of Cyber Science and Engineering, Huazhong University of Science and Technology
[2]Shenzhen Huazhong University of Science and Technology Research Institute
[3]Nanyang Technological University [4]Shanghai Artificial Intelligence Laboratory
[5]Mohamed bin Zayed University of Artificial Intelligence [6]Wuhan University
{ziji_sheng, tgy, panzhou}@hust.edu.cn, daizongliu@whu.edu.cn

Abstract

Large vision-language models (LVLMs) achieve strong multimodal reasoning performance but remain prone to object hallucination, where generated objects or existence judgments are not faithfully grounded in the image. Existing mitigation methods either require additional training or rely on external visual experts, which limits plug-and-play usability and increases inference overhead. In this paper, we propose **SIGMA**, a training-free framework for hallucination mitigation based on **S**elf-vision **I**nformation-theoretic **G**ating and dual-**M**ask **A**ssembly. SIGMA operates directly on the LVLM's own vision encoder, constructing complementary support and complement masks from image-driven concept relevance, question-aware entity cues, and CLS–patch shortcut suppression. These masks form global, support-masked, and complement-masked visual streams for query-conditioned evidence separation. During decoding, SIGMA activates correction only under high uncertainty, using full-vocabulary and candidate-answer entropy to perform support-first refinement and complement-on-demand suppression. The resulting streams are combined through logit-level assembly with lightweight positive calibration and a plausibility constraint. Experiments across multiple LVLM backbones and hallucination-oriented benchmarks show that SIGMA effectively reduces object hallucination while maintaining competitive perception performance and a selective low-overhead inference path.

1 Introduction

Large Vision-Language Models (LVLMs) have demonstrated remarkable visual understanding capabilities, enabling applications in autonomous driving [52, 4], medical image analysis [22], embodied robotics [14, 48], and various multimodal agents [38, 39, 58, 37]. Despite this progress, LVLMs remain vulnerable to object hallucination, a failure mode in which they generate object descriptions that are not grounded in the input image. This failure stems from statistical biases in the training corpus, which lead LVLMs to predict objects that frequently appear in similar textual or visual contexts [17]. In fact, object hallucination poses a security threat to real-world applications [40, 36, 52].

*Equal Contributions. †Project Lead. ‡Corresponding Authors: Daizong Liu and Pan Zhou.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

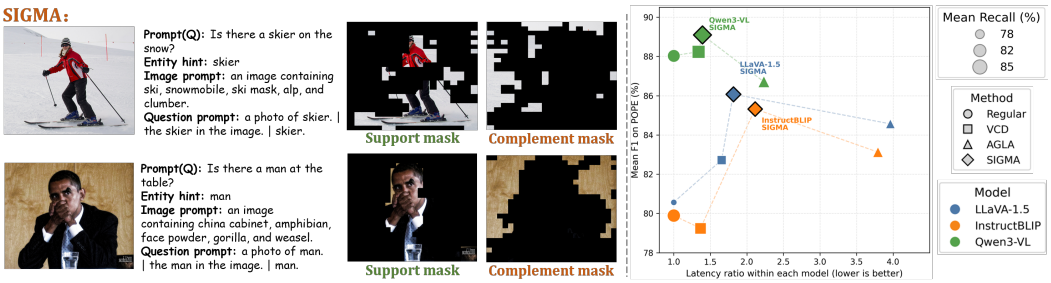

Figure 1: (Left) Visualization of support and complement masks generated by SIGMA for visual evidence separation. (Right) Performance-efficiency comparison on the POPE benchmark. SIGMA achieves superior hallucination mitigation (higher Mean F1) while maintaining a low inference latency ratio across multiple LVLMs.

For example, in autonomous driving, an LVLM may mistake a roadside shadow for a pedestrian or traffic cone, leading to unsafe steering.

Existing hallucination mitigation methods can be organized into three categories according to where additional supervision or computation is introduced [21, 52, 40]: training-stage optimization, inference-stage mitigation with auxiliary vision models, and training-free inference-stage intervention without auxiliary models. Specifically, training-stage methods enhance visual grounding through hallucination-aware data construction, grounding alignment objectives, or model fine-tuning [62, 25, 64, 65], but require additional optimization and often show limited transferability across LVLM backbones [53]. Inference-stage methods with auxiliary vision models avoid updating LVLM parameters, but rely on external detection, segmentation, or grounding models, thereby increasing system complexity and latency [3, 63]. Recent training-free inference-stage methods without auxiliary models improve plug-and-play applicability, yet they often operate on global visual representations or apply uniform decoding interventions across generation steps [53, 24, 26]. As a result, they struggle to disentangle query-relevant visual evidence from shortcut-prone contextual cues, leading to a persistent trade-off between hallucination suppression and inference efficiency [57].

Research Questions. In this paper, we address the above challenges through three research questions.

RQ1: How can query-conditioned visual evidence be constructed within the LVLM architecture without additional training or auxiliary vision modules?

RQ2: How can supportive evidence and misleading context be separated for hallucination correction?

RQ3: How can correction be selectively activated to balance mitigation effectiveness and inference efficiency?

To answer these questions, we propose **SIGMA**, a training-free inference framework based on **S**elf-vision **I**nformation-theoretic **G**ating and dual-**M**ask **A**ssembly. For **RQ1**, SIGMA constructs query-conditioned evidence within the LVLM's own vision encoder by combining image-driven concept relevance and question-aware entity signals, without relying on external vision models. For **RQ2**, SIGMA builds two complementary masks: a support mask that emphasizes patches directly grounding the queried entity, and a complement mask that captures query-inconsistent or shortcut-prone contextual cues. These masks are injected into the upper layers of the self vision encoder to form global, support, and complement visual streams, whose logit residuals are assembled with lightweight positive calibration and a global plausibility constraint. For **RQ3**, SIGMA follows the global stream by default and activates corrective branches only when prediction uncertainty is high, using full-vocabulary and candidate-answer entropy to trigger support-first refinement and complement-on-demand suppression. This design allocates corrective computation to high-uncertainty decoding steps while preserving a low-overhead inference path.

Our main contributions are as follows: (1) We identify insufficient query-conditioned patch discrimination as a key source of object hallucination, and formulate hallucination mitigation as separating supporting visual evidence from shortcut-prone contextual cues. (2) We propose **SIGMA**, a training-free framework that separates supportive visual evidence from misleading context through self-vision dual masking and uncertainty-guided assembly, mitigating object hallucination without auxiliary vision models. (3) We evaluate SIGMA across multiple LVLM backbones and hallucination-oriented

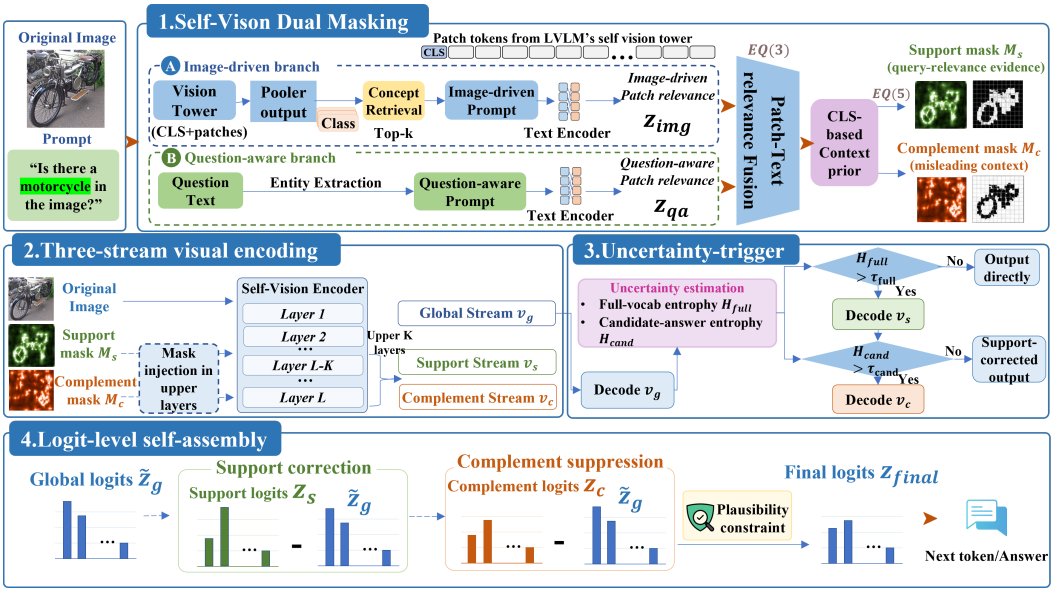

Figure 2: Overview of SIGMA. SIGMA forms global, support, and complement visual streams via self-vision dual masking, and performs uncertainty-triggered logit-level assembly for selective hallucination correction.

benchmarks, showing that it mitigates object hallucination while preserving competitive perception performance and a selective, low-overhead inference path.

2 Method

2.1 Overview

The method consists of three stages. We first perform *self-vision dual masking*, where a support mask and a complement mask are constructed from two complementary prompt families. We then inject the two masks into the upper layers of the self vision encoder to form three visual streams, namely the original global stream, the support stream, and the complement stream. Finally, we apply *uncertainty-triggered three-stream assembly*: the model decodes along the global stream by default, activates the support stream when the decision becomes uncertain, and invokes the complement stream only when the candidate state remains risky. An overview is shown in Fig. 2.

2.2 Self-Vision Dual Masking

The goal of the dual-mask stage is to reorganize visual evidence inside the LVLM's own visual pathway according to the current query. For a queried entity, the model should preserve patches that directly support the decision, while also identifying contextual regions that are not themselves reliable evidence but may bias the prediction through co-occurrence patterns, background semantics, or shortcut features. We therefore construct two complementary evidence views: a support view that emphasizes query-aligned visual evidence, and a complement view that captures anti-query or shortcut-prone context. A key challenge is that neither the image alone nor the question alone provides a sufficiently stable basis for this separation. Image-level semantics can highlight visually dominant scene concepts that are irrelevant to the query, whereas question-only conditioning can overemphasize language priors when the visual evidence is ambiguous. To address this, we build the masks from the interaction of two prompt families: an *image-driven concept prompt*, which provides a visually grounded semantic prior, and a *question-aware entity prompt*, which constrains the evidence toward the queried entity.

Prompt construction. Let $H = [h_{\text{cls}}, h_1, \ldots, h_N]$ denote the final-layer visual tokens, where h_{cls} is the CLS token and $\{h_j\}_{j=1}^{N}$ are patch tokens. After projection into the shared visual–textual space, we obtain the normalized CLS representation c and normalized patch representations $\{p_j\}_{j=1}^{N}$. The

image-driven prompt is obtained by matching the global image representation against a concept vocabulary, retrieving a small set of visually plausible concepts, and verbalizing them into auxiliary concept prompts. These prompts provide a scene-grounded prior over what objects or concepts are likely to appear in the image. In contrast, the question-aware prompt is constructed from an entity-centric phrase extracted from the input question and reformulated as a compact prompt ensemble, so that the mask is explicitly conditioned on the queried object. Together, these two prompt families allow the model to compare each patch against both visual plausibility and query relevance, which forms the basis for subsequent evidence separation.

From prompts to patch-token scores. The retrieved visual concepts are encoded and aggregated into an auxiliary concept embedding z_{img}, while the question-aware prompt ensemble is encoded and averaged into an entity-conditioned embedding z_{qa}. For each patch representation p_j, we first compute raw patch-token prompt scores:

$$s_j^{\text{img}} = p_j^\top z_{\text{img}}, \qquad s_j^{\text{qa}} = p_j^\top z_{\text{qa}}. \tag{1}$$

These scores measure patch–prompt compatibility under the current visual–textual projection, but their numerical direction is not always a reliable indicator of token-level relevance. This distinction is important because patch-token score maps can exhibit a different polarity from global image–text similarity.

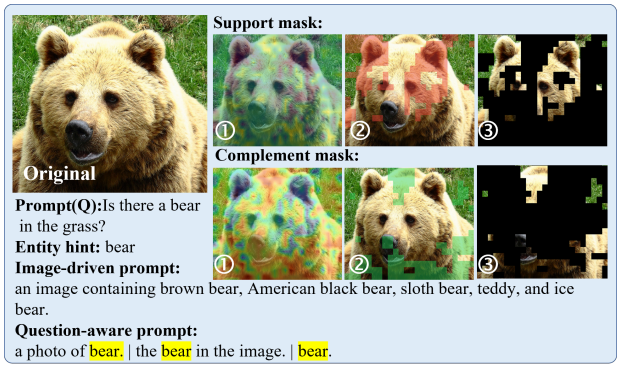

Figure 3: Example of dual masking.

Polarity-corrected relevance fusion. Following the token-level observation that prompt-relevant patch tokens may correspond to lower raw patch–prompt scores under CLIP-style patch representations, we convert the raw scores into calibrated relevance maps with a polarity-corrected normalization:

$$r_j^{\text{img}} = \sigma\big(\text{Std}(\gamma_E s_j^{\text{img}})\big), \qquad r_j^{\text{qa}} = \sigma\big(\text{Std}(\gamma_E s_j^{\text{qa}})\big), \tag{2}$$

where $\text{Std}(\cdot)$ denotes patch-wise standardization, $\sigma(\cdot)$ is the sigmoid function, and $\gamma_E \in \{-1, +1\}$ denotes the patch-token score polarity of the self vision encoder. For ERICT-style CLIP patch-token scoring, $\gamma_E = -1$, which corresponds to extracting concept-relevant tokens from standardized negative scores; for other self-vision encoders, the polarity is specified by the corresponding adapter. The resulting r^{img} and r^{qa} are calibrated relevance maps, rather than raw similarities. We then combine them into an agreement-focused support seed:

$$s_j^{\text{sup}} = (1 - \lambda_q)\sqrt{r_j^{\text{img}} r_j^{\text{qa}}} + \lambda_q r_j^{\text{qa}}, \tag{3}$$

where λ_q controls the strength of the question-aware branch. The geometric agreement term favors patches that are jointly supported by the image-driven prior and the queried entity, while the question-aware term prevents the mask from drifting toward generic scene concepts. We use this fusion as a principled default for query-consistent evidence recovery. Appendix C.2 provides the design rationale for this fusion.

Context-aware dual-mask scoring. Although the support seed identifies query-aligned patches, it does not yet account for globally salient regions that are weakly related to the query but may still influence the model's decision. Such regions are particularly important in object-existence questions, where background patterns or co-occurring objects can induce hallucinated positive judgments. To capture this shortcut-prone context, we introduce a CLS-based context prior b_j, inspired by the LazyStrike intuition that globally dominant visual tokens can reveal context likely to steer the model. The final support and complement scores are defined as

$$b_j = \sigma\big(\text{Std}(p_j^\top c)\big)(1 - s_j^{\text{sup}}), \qquad m_j^{\text{sup}} = s_j^{\text{sup}}(1 - \lambda_b b_j), \qquad m_j^{\text{comp}} = (1 - s_j^{\text{sup}}) + \lambda_b b_j. \tag{4}$$

where λ_b controls the strength of context modeling. Here, b_j is not a generic CLS saliency score; it is a query-inconsistent CLS context prior, because CLS–patch saliency is gated by the anti-query term

Algorithm 1 Tri-stream encoding through masked self-vision injection

Require: Image x, support mask M_s, complement mask M_c, self vision encoder $\mathcal{E}$ with layers $\{\ell_1, \ldots, \ell_L\}$

Require: Number of injected upper layers L_m, minimum keep ratio ρ

$v_g \leftarrow \mathcal{E}(x)$ ▷ native visual stream

for $M \in \{M_s, M_c\}$ **do**

  Register a pre-attention hook on layers $\ell_{L-L_m+1}, \ldots, \ell_L$

  In each hooked layer, reweight patch tokens by $\rho + (1-\rho)M$ and keep the CLS token unchanged

  $v_M \leftarrow \mathcal{E}(x)$ under the hook; then remove all hooks

end for

$v_s \leftarrow v_{M_s}$; $v_c \leftarrow v_{M_c}$

return (v_g, v_s, v_c)

$(1 - s_j^{\mathrm{sup}})$. Thus, globally salient patches are assigned to the complement view only when they are weakly supported by the queried entity. This preserves the intended semantics of the two branches: the support score remains a context-regularized measure of query-aligned evidence, whereas the complement score accumulates anti-query evidence together with shortcut-prone global context. After light spatial smoothing and sparse normalization, m^{sup} and m^{comp} are converted into the final support and complement masks, which are subsequently injected into the self vision encoder to form the support and complement visual streams.

2.3 Three-Stream Visual Encoding

Given the two masks, we do not create external visual experts or separate encoders. Instead, as shown in algorithm 1 , we inject the support and complement masks into the upper layers of the LVLM's own vision encoder to produce three visual streams: the original global stream, the support stream, and the complement stream. This design is important for two reasons. First, all three streams remain in the same latent space because they are produced by the same self-vision encoder. Second, the masks operate as evidence re-weighting rather than hard token deletion, which avoids destroying useful boundary information or low-level visual structure.

Concretely, the global stream preserves the model's native visual path and serves as the default decoding backbone. The support stream emphasizes evidence-consistent regions and is therefore responsible for positive refinement. The complement stream emphasizes anti-query or shortcut-prone context and is therefore responsible for suppressive correction. In this way, the method reorganizes the LVLM's internal visual evidence into a structured triplet rather than forcing all decisions to rely on a single global visual representation.

2.4 Uncertainty-Triggered Three-Stream Assembly

Using all branches at every decoding step would introduce unnecessary computation and would also allow corrective branches to perturb confident predictions. We therefore activate correction only when the current decision is uncertain. At decoding step i, the model first follows the global stream and produces logits $z_g^{(i)}$. We quantify uncertainty from two complementary views: the full-vocabulary entropy, which measures how diffuse the current distribution is over the entire vocabulary, and the candidate-answer entropy, which measures the ambiguity within a compact candidate set associated with the current question. Correction is triggered whenever either uncertainty exceeds its threshold:

$$\mathbb{I}_{\mathrm{trigger}}^{(i)} = \mathbf{1}\left[H_{\mathrm{full}}^{(i)} > \tau_{\mathrm{full}} \ \vee \ H_{\mathrm{cand}}^{(i)} > \tau_{\mathrm{cand}}\right]. \tag{5}$$

For existence questions, the candidate set defaults to the yes/no pair; for more general cases, it falls back to a compact top-k candidate set. This candidate-aware trigger is more faithful to the actual decision boundary than full-vocabulary entropy alone.

Once triggered, the model first invokes the support stream and obtains support-branch logits. Only if the candidate state remains uncertain—for example, because the candidate entropy is still high, the top-margin remains small, or the support branch disagrees with the global branch—does the model further invoke the complement stream. This *support-first, complement-on-demand* schedule keeps the common low-risk path lightweight while concentrating extra computation on high-risk cases.

Table 1: Detailed results on the three POPE subsets. We report Accuracy (Acc.), Precision (Prec.), Recall (Rec.), F1, and average latency per sample (Lat./sample, seconds). For non-Regular methods, the small subscript in the latency column denotes the increase over Regular decoding. Best and second-best performance values within each backbone are highlighted in bold and underlined, respectively.

Model	Method	Random				Popular				Adversarial				Lat./sample ↓
		Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	Acc. ↑	Prec. ↑	Rec. ↑	F1 ↑	
LLaVA-1.5	Regular	83.70	89.34	76.53	82.44	82.00	85.93	76.53	80.96	78.83	80.31	76.40	78.31	**0.1186**
	OPERA	87.27	93.54	80.07	86.28	84.13	88.21	78.80	83.24	81.27	82.17	79.87	81.00	0.1714 +0.05
	VCD	85.80	91.12	79.33	84.82	84.17	87.82	79.33	83.36	80.10	80.40	79.60	80.00	0.1962 +0.08
	AGLA	88.23	93.61	82.07	87.47	86.00	**94.12**	76.80	84.58	81.57	81.20	**82.13**	81.65	0.4697 +0.35
	Ours	**89.27**	**94.15**	**83.73**	**88.64**	**87.17**	90.72	**82.80**	**86.58**	**83.77**	**87.05**	79.33	**83.01**	0.2152 +0.10
InstructBLIP	Regular	82.70	80.33	**86.60**	83.35	77.40	72.86	**87.33**	79.44	74.20	69.64	**85.80**	76.88	**0.1268**
	OPERA	84.90	88.24	80.53	84.21	79.67	77.00	84.60	80.62	76.53	72.77	84.80	78.33	0.1490 +0.02
	VCD	81.83	81.52	82.33	81.92	77.77	74.63	84.13	79.10	74.90	71.53	82.73	76.72	0.1733 +0.05
	AGLA	87.45	88.83	85.68	87.23	82.03	82.71	81.00	81.85	80.70	82.00	78.67	80.30	0.4809 +0.35
	Ours	**88.70**	**95.32**	81.40	**87.81**	**86.17**	**89.98**	81.40	**85.47**	**82.93**	**83.88**	81.53	**82.69**	0.2679 +0.14
Qwen3-VL	Regular	91.83	96.79	86.53	91.38	87.73	89.14	85.93	87.51	85.33	85.91	84.53	85.22	**0.2309**
	OPERA	87.77	92.50	82.20	87.05	85.50	88.01	82.20	85.01	82.80	83.02	82.47	82.74	0.2932 +0.06
	VCD	91.23	96.19	85.87	90.74	88.33	89.14	**87.33**	88.22	85.73	85.59	**85.93**	85.76	0.3091 +0.08
	AGLA	90.00	95.52	83.93	89.35	86.53	88.54	83.93	86.17	84.73	85.25	84.00	84.62	0.5153 +0.28
	Ours	**92.50**	**97.33**	**87.40**	**92.10**	**88.90**	**92.13**	85.07	**88.46**	**87.00**	**88.54**	85.00	**86.73**	0.3213 +0.09
InternVL3.5	Regular	92.90	**95.03**	90.53	92.73	86.23	**83.08**	91.00	86.86	83.37	79.01	90.87	84.53	**0.2068**
	OPERA	90.73	90.79	90.67	90.73	86.40	82.19	92.93	87.23	83.63	78.07	93.53	85.11	0.2271 +0.02
	VCD	91.83	93.73	89.67	91.65	84.27	81.08	89.40	85.03	82.43	78.40	89.53	83.60	0.3337 +0.13
	AGLA	93.07	93.24	92.87	93.05	86.70	82.40	93.33	87.53	84.00	78.21	94.27	85.49	0.5015 +0.29
	Ours	**93.67**	93.78	**93.53**	**93.66**	**87.77**	82.58	**95.73**	**88.67**	**85.47**	**79.65**	**95.27**	**86.76**	0.3525 +0.15

Before assembling the branch logits, we apply two lightweight stabilizers. They are not additional evidence sources, but are used to regularize the residual assembly. Let C_i denote the candidate set at step i, and let $P_i^+ \subseteq C_i$ denote the task-relevant positive subset. We first obtain calibrated global logits by applying an entropy-gated positive shift:

$$\tilde{z}_{g,v}^{(i)} = z_{g,v}^{(i)} + \eta\, \mathbf{1}\big[v \in P_i^+\big]\, \mathbf{1}\Big[H_{\text{cand}}^{(i)} \geq \tau_{\text{cal}}\Big], \tag{6}$$

where η is a small calibration strength. This calibration mildly compensates for conservative suppression of positive candidates and helps preserve recall under uncertain decisions.

The branch logits are then assembled in residual form:

$$z_{\text{asm}}^{(i)} = \tilde{z}_g^{(i)} + \alpha_i \Big[\beta_s\big(z_s^{(i)} - \tilde{z}_g^{(i)}\big) - \beta_c\big(z_c^{(i)} - \tilde{z}_g^{(i)}\big)\Big], \tag{7}$$

where $z_s^{(i)}$ and $z_c^{(i)}$ denote the support- and complement-stream logits, and β_s, β_c are the corresponding fusion strengths. The support stream contributes a positive residual that sharpens predictions with evidence-consistent refinement, whereas the complement stream contributes a subtractive residual that suppresses deviations induced by misleading context. When the complement branch is not activated, Eq. (7) naturally degenerates to support-only correction.

Finally, to prevent residual assembly from drifting toward globally implausible tokens, we constrain the assembled logits using the calibrated global distribution:

$$z_{\text{final},v}^{(i)} = z_{\text{asm},v}^{(i)} - \kappa\, \mathbf{1}\Big[\text{softmax}(\tilde{z}_g^{(i)})_v \leq \tau_p\Big], \tag{8}$$

where τ_p is a plausibility threshold and κ is a penalty. This constraint restricts the final prediction to tokens that remain plausible under the calibrated global stream, limiting logit drift introduced by residual correction. Taken together, the decoding stage yields a selective correction mechanism: the model follows the native global path when it is already confident, and invokes progressively stronger correction only when uncertainty justifies the extra intervention.

3 Experiments

3.1 Experimental Setup

Benchmarks and metrics. Following prior hallucination-mitigation studies, we evaluate SIGMA on both discriminative and generative benchmarks. POPE [29] evaluates object-existence hallucination

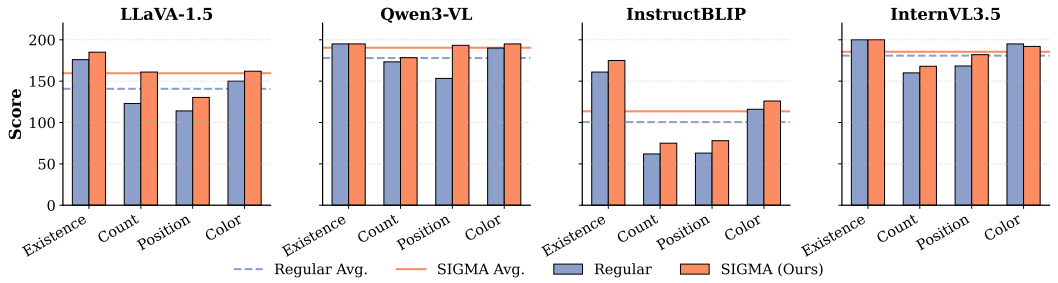

Figure 4: MME object-centric perception results. Bars show subset-level scores, and horizontal lines denote the average scores of Regular decoding and SIGMA. Higher is better.

under Random, Popular, and Adversarial settings; we report Accuracy, Precision, Recall, and F1. AMBER [59] evaluates hallucination in both generative and discriminative settings; we focus on its discriminative benchmark, which covers existence, attribute, and relation hallucination with yes/no-style questions. MME [18] provides yes/no questions across perception and cognition tasks; we focus on object-centric perception subsets when analyzing hallucination mitigation, and report the MME score. CHAIR [47] evaluates object hallucination in open-ended image captioning, where lower CHAIR_s/CHAIR_i and higher object recall are preferred. We additionally report ROPE [11] in Appendix C.1 to examine the transferability of SIGMA to multi-object hallucination settings. Benchmark-specific implementation details for adapting SIGMA are provided in Appendix C.

Evaluated LVLMs. We evaluate representative LVLM backbones, including LLaVA-v1.5-7B-hf [42], InstructBLIP-vicuna-7b [15], Qwen3-VL-8B-Instruct [6], and InternVL3.5-8B [60]. These models cover different visual encoders, multimodal connectors, and language-model backbones, allowing us to examine whether self-vision masking and uncertainty-triggered assembly transfer across heterogeneous LVLM architectures.

Baselines. We compare SIGMA with Regular [26] decoding and representative training-free hallucination mitigation methods, including OPERA [24], VCD [26], and AGLA [3]. We additionally compare with ONLY [57] and EAZY [10] in matched POPE and captioning evaluations; the additional comparisons are reported separately from Table 1. For backbones where a baseline does not provide an official implementation, we use the closest faithful reproduction and specify the adaptation details in Appendix C. All methods are evaluated under the same benchmark-visible prompts whenever prompt rewriting is required by the benchmark protocol.

Implementation details. All methods use identical benchmark-specific preprocessing, prompts, decoding budgets, and answer parsers. SIGMA constructs mask prompts only from test-time visible information and never uses ground-truth labels, reference captions, or image-specific annotations for mask construction. Latency is measured under identical hardware and batch-size settings; detailed protocols and leakage controls are provided in Appendix C.1.

3.2 Main Results

Hallucination mitigation on POPE and AMBER. Table 1 reports detailed results on POPE. Among the methods in Table 1, SIGMA achieves the best F1 on every Random, Popular, and Adversarial subset across all evaluated backbones. Averaged over the three subsets, it improves F1 over Regular decoding by 5.51, 5.43, 1.06, and 1.66 points on LLaVA-1.5, InstructBLIP, Qwen3-VL, and InternVL3.5, respectively. The consistent gains on the Adversarial split further indicate that SIGMA remains effective when hallucination is induced by shortcut-prone contextual co-occurrence. Table 2 shows that SIGMA also generalizes to AMBER discriminative evaluation, improving both Accuracy and F1 across all four backbones. The gains are more pronounced on relatively weaker models, while stronger backbones still obtain consistent improvements. These

Table 2: Overall results on the AMBER discriminative benchmark. We compare Regular decoding and SIGMA using Accuracy (Acc.) and F1; detailed sub-task results are provided in C.1.

Model	Regular		SIGMA (Ours)	
	Acc.	F1	Acc.	F1
LLaVA-1.5	64.0	73.6	**78.9**	**81.2**
InstructBLIP	74.4	78.8	**80.1**	**82.7**
Qwen3-VL	89.3	90.6	**91.7**	**92.3**
InternVL3.5	88.7	90.2	**91.2**	**91.7**

Table 4: POPE results on high-clutter and high-overlap subsets. Each cell reports Accuracy/F1 (%).

Stratum	Split	Regular	ONLY	SIGMA
High clutter ($n = 576$)	Popular	76.39/78.41	80.56/82.05	**84.20/83.60**
	Adversarial	73.09/76.12	76.56/79.13	**81.60/81.40**
High overlap ($n = 348$)	Popular	81.90/83.20	87.36/87.78	**88.79/88.22**
	Adversarial	78.74/80.83	82.76/84.04	**84.77/84.64**

Table 5: End-to-end latency and selective activation analysis of SIGMA on object-centric hallucination benchmarks. Latency ratios are computed as $T_{\text{SIGMA}}/T_{\text{baseline}}$; values below $1.0\times$ indicate that SIGMA is faster than the corresponding baseline.

Benchmark	Backbone	SIGMA Lat./sample ↓	Latency Ratio ↓				Branch Activation Rate (%) ↓	
			vs. Regular	vs. OPERA	vs. VCD	vs. AGLA	Support	Complement
POPE-COCO	LLaVA-1.5	0.2152	1.81×	1.26×	1.10×	0.46×	61.67	58.22
	InstructBLIP	0.2679	2.11×	1.80×	1.55×	0.56×	73.16	65.09
	Qwen3-VL	0.3213	1.39×	1.10×	1.04×	0.62×	42.17	28.89
AMBER-Exist.	LLaVA-1.5	0.2047	1.55×	1.17×	0.95×	0.34×	52.97	47.13
	InstructBLIP	0.2544	1.60×	1.55×	0.92×	0.32×	50.83	49.46
	Qwen3-VL	0.2882	1.52×	1.16×	0.97×	0.25×	41.90	34.03
MME-Obj.	LLaVA-1.5	0.2290	1.59×	1.26×	1.01×	0.36×	66.31	48.70
	InstructBLIP	0.2860	1.89×	1.68×	0.96×	0.34×	64.41	52.10
	Qwen3-VL	0.3380	1.61×	1.31×	1.02×	0.27×	48.50	41.00

results demonstrate that self-vision evidence separation provides robust hallucination mitigation across heterogeneous LVLM architectures and evaluation protocols. Additional comparisons with ONLY and EAZY under matched evaluation settings, together with paired confidence intervals for the SIGMA–ONLY comparison, are provided in Appendix C.4.

Object-centric perception and open-ended generation. Figure 4 shows that SIGMA preserves and often improves object-centric perception on MME. Across the four evaluated backbones, SIGMA raises the average score over Regular decoding, with more visible gains on Count and Position, suggesting that query-conditioned evidence separation improves object localization without suppressing valid object predictions.

Table 3: CHAIR results on 500 COCO images with LLaVA-1.5.

Method	$C_s \downarrow$	$C_i \downarrow$	Rec. ↑
Regular	24.40	8.08	50.69
ONLY	23.80	7.31	52.86
EAZY	21.40	6.33	53.65
SIGMA	**20.20**	**6.13**	**55.80**

Table 3 reports the matched CHAIR evaluation with LLaVA-1.5 on 500 COCO images. SIGMA achieves the lowest CHAIR_s/CHAIR_i and highest recall among Regular, ONLY, EAZY, and SIGMA, using the same images, prompt, 64-token generation cap, and scorer. Average caption lengths are comparable (48.234–49.306 words), indicating that the improvement is not obtained by materially shortening captions. SIGMA takes 3.544 s/sample versus EAZY's 8.758 s/sample, a 59.5% reduction, with additional overhead over Regular and ONLY. Appendix C.7 provides the caption lengths, latency, and original InstructBLIP comparison: OPERA achieves lower CHAIR_i (13.4 vs. 17.1), whereas SIGMA achieves lower CHAIR_s (49.0 vs. 54.0) and higher recall (72.1 vs. 69.3). Additional ROPE results in Appendix C.1 and Table 11 further evaluate SIGMA under multi-object hallucination settings.

Crowded and overlapping scenes. Table 4 evaluates frozen high-clutter and high-overlap strata. SIGMA improves both Accuracy and F1 over Regular and ONLY on Popular and Adversarial, supporting its effectiveness when multiple or overlapping objects compete for visual evidence. COCO annotations are used only after inference to define the evaluation strata.

Latency evaluation. Table 5 analyzes the inference efficiency of SIGMA. Compared with AGLA, SIGMA substantially reduces latency on POPE, from 0.4697s to 0.2152s on LLaVA-1.5, from 0.4809s to 0.2679s on InstructBLIP, and from 0.5153s to 0.3213s on Qwen3-VL. Similar latency advantages over AGLA are observed on AMBER-Exist. and MME-Obj. SIGMA introduces moderate

Table 7: Branch-scheduling ablation on POPE-COCO Adversarial using LLaVA-1.5. Gate, Sup.-first, and Comp.-on-demand denote uncertainty gating, support-first refinement, and complement-on-demand suppression; rates are measured over decoding steps. The ablation is conducted as an independent run with a different random seed.

Variant	Gate	Sup.-first	Comp.-on-demand	Acc. ↑	F1 ↑	Lat./sample ↓	Sup. Rate	Comp. Rate
Ours	●	●	●	83.73	82.84	0.2649	59.57	57.20
Always-on 3-stream	○	○	○	83.73	82.84	0.3141	100.00	100.00
Support-only correction	●	●	○	83.50	82.35	0.2050	59.68	0.00
No uncertainty gating	○	●	●	83.73	82.84	0.2981	100.00	68.87
Global only	○	○	○	78.83	78.31	0.1192	0.00	0.00

overhead over Regular decoding because it constructs dual masks and selectively activates corrective streams, but its latency remains substantially lower than heavier multi-branch correction. The support and complement activation rates further show that the corrective branches are not applied uniformly to all decoding steps. Instead, computation is allocated through support-first refinement and complement-on-demand suppression, which explains how SIGMA achieves hallucination mitigation while maintaining a selective, low-overhead inference path.

3.3 Ablation Study

Effect of core components. Table 6 ablates the core components on POPE-COCO Random using LLaVA-1.5. Removing support, complement, and residual assembly reduces F1 by 0.74, 3.01, and 8.43 points, respectively, identifying the contributions of evidence recovery, shortcut suppression, and structured residual composition. Removing uncertainty triggering preserves Accuracy/F1 but increases latency from 0.217 to 0.251 s/sample. Thus, evidence correction accounts for the quality gain, while triggering primarily reduces computation.

Table 6: Core component ablation on POPE-COCO Random using LLaVA-1.5. Sup., Com., Trig., and Asse. denote the support mask, complement mask, uncertainty trigger, and logit assembly; ●/○ indicate enabled/disabled.

Variant	**Components**				**Acc. ↑**	**F1 ↑**	**Lat.↓**
	Sup.	Com.	Trig.	Asse.			
Ours	●	●	●	●	89.27	88.64	0.217
− Support	○	●	●	●	88.67	87.90	0.175
− Complement	●	○	●	●	87.13	85.63	0.172
− Trigger	●	●	○	●	89.27	88.64	0.251
− Assembly	●	●	●	○	83.33	80.21	0.196
Base (Global)	○	○	○	○	83.70	82.44	0.121

Effect of branch scheduling. Table 7 studies branch scheduling on POPE-COCO Adversarial with LLaVA-1.5. Since this ablation is independently run with a different random seed, its absolute numbers may slightly differ from Table 1. In this controlled setting, the always-on three-stream variant matches SIGMA in Accuracy and F1 but incurs higher latency by activating both corrective branches at every decoding step. Removing uncertainty gating also preserves the final score, yet increases the support activation rate to 100% and raises latency, confirming that gating mainly removes redundant computation. The support-only variant is faster but yields lower F1, showing that support refinement alone is insufficient in harder adversarial cases and that complement-on-demand suppression remains necessary for context-induced hallucination. Overall, SIGMA matches the always-on variant with lower latency, validating the support-first and complement-on-demand schedule. A separate matched 900-example validation gives identical final answers for selective and always-on execution with lower latency (Appendix C.5).

Evidence allocation. Post-hoc spatial tests show support enrichment and complement depletion on visible targets (Appendix C.3).

4 Related Work

Large Vision-Language Models. Large Vision-Language Models (LVLMs) extend language models to multimodal understanding [13, 23, 40, 41, 8, 61, 9, 32, 31, 35, 1, 30, 34, 33] by coupling visual encoders with LLMs through projection layers, query transformers, or modality connectors [46, 20, 2, 28, 56, 67, 43]. Representative models such as BLIP-2 [27], InstructBLIP [15], LLaVA [43, 42], Qwen-VL [5, 6], InternVL [12] et al. have achieved strong performance on visual question answering,

image captioning, and multimodal instruction following [51]. However, their generative ability is still limited by imperfect visual grounding, making them prone to producing object mentions or existence judgments that are not supported by the input image [23, 7, 63].

Hallucination Mitigation in LVLMs. Object hallucination has been widely recognized as a reliability issue in LVLMs, particularly in object-existence judgments [23, 21]. Prior mitigation methods mainly differ in how visual evidence is acquired and corrected. Training-based methods improve grounding through hallucination-aware data, preference alignment, or fine-grained reward modeling [62, 25, 64, 65, 45, 49], but require additional optimization for each target model. Inference-time correction methods often employ auxiliary detectors, segmenters, grounding models, or multi-step verification pipelines [63, 66, 55, 54], introducing extra modules and latency. Training-free decoding methods avoid parameter updates by contrasting visual inputs, adjusting attention, or calibrating logits [26, 24, 19, 44]; however, they usually operate on global visual evidence and do not explicitly separate supportive evidence from context-induced shortcuts. Recent studies on vision-language robustness and ViT representations further show that token-level visual evidence can be manipulated at inference time to reduce spurious correlations or background-dominated shortcuts [16, 50]. VCD [26] contrasts the original image with a query-independent distorted view, whereas AGLA [3] supplements global evidence with a positive local view obtained through an auxiliary matcher. SIGMA constructs native, query-conditioned support and complement views with asymmetric recovery and suppression roles, and invokes complement correction only when support correction leaves residual risk.

5 Conclusion

In this paper, we presented **SIGMA**, a training-free hallucination mitigation framework for LVLMs. SIGMA constructs support and complement masks within the LVLM's own vision encoder, forming query-conditioned visual streams that separate entity-grounded evidence from shortcut-prone contextual cues. During decoding, SIGMA follows the global stream by default and activates corrective branches only under uncertainty, enabling support-first refinement and complement-on-demand suppression through logit-level assembly. Experiments across multiple LVLM backbones and hallucination-oriented benchmarks show that SIGMA reduces object hallucination while preserving competitive object-centric perception performance and maintaining a selective, low-overhead inference path. These results suggest that reorganizing internal visual evidence, without introducing task-specific external visual experts such as detectors, segmenters, or grounding models, is a practical alternative to additional training for improving LVLM reliability. We discuss the limitations of SIGMA and future directions for extending it to more general multi-entity and open-ended generation scenarios in Appendix A.

Acknowledgements

This work was supported by New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0123602) and National Natural Science Foundation of China (NSFC) under grant No.62476107. This work was also supported by the Department of Science and Technology of Hubei Province (No. 2026DJA002). This work was also supported by WHU-Kingsoft Joint Lab. The numerical calculations in this paper have been done in part on the supercomputing system in the Supercomputing Center of Wuhan University.

References

[1] Nian Ai, Guangke Chen, Xiaowen Cai, Zhongliang Guo, Daizong Liu, Pan Zhou, and Ognjen Arandjelović. Attacking hard-label large vision-language models with model-sensitive adversarial patch designs. *Pattern Recognition*, page 114261, 2026.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, and et al. Cabi. Flamingo: a visual language model for few-shot learning. NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc.

[3] Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926, 2025.

[4] Kazuki Atsuta, Kohei Honda, Hiroyuki Okuda, and Tatsuya Suzuki. Lvlm-mpc collaboration for autonomous driving: A safety-aware and task-scalable control architecture. *arXiv preprint arXiv:2505.04980*, 2025.

[5] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[6] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[7] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[8] Xiaowen Cai, Daizong Liu, Runwei Guan, and Pan Zhou. Imperceptible transfer attack on large vision-language models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[9] Xiaowen Cai, Daizong Liu, Xiaoye Qu, Xiang Fang, Jianfeng Dong, Keke Tang, Pan Zhou, Lichao Sun, and Wei Hu. Towards building model/prompt-transferable attackers against large vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[10] Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21635–21644, 2025.

[11] Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *3rd Workshop on Advances in Language and Vision Research (ALVR)*, 2024.

[12] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[13] Zhiyuan Chen, Yuecong Min, Jie Zhang, Bei Yan, Jiahao Wang, Xiaozhen Wang, and Shiguang Shan. A survey of multimodal hallucination evaluation and detection. *International Journal of Computer Vision*, 134(3):131, 2026.

[14] Yao Cong and Hongwei Mo. An overview of robot embodied intelligence based on multimodal models: Tasks, models, and system schemes. *International Journal of Intelligent Systems*, 2025(1):5124400, 2025.

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

[16] Xinpeng Dong, Min Zhang, Didi Zhu, Ye Jun Jian, Zhang Keli, Aimin Zhou, Fei Wu, and Kun Kuang. Erict: Enhancing robustness by identifying concept tokens in zero-shot vision language models. In *Forty-second International Conference on Machine Learning*, 2025.

[17] Zhiyuan Fan, Yumeng Wang, Sandeep Polisetty, and Yi R Fung. Unveiling the lack of lvlm robustness to fundamental visual variations: Why and path forward. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20222–20242, 2025.

[18] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[19] Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7696–7712, 2024.

[20] Yuncheng Guo and Xiaodong Gu. Mmrl: Multi-modal representation learning for vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25015–25025, 2025.

[21] Zheng Yi Ho, Siyuan Liang, and Dacheng Tao. Understanding hallucinations in large visual and language models. *ACM Computing Surveys*, 2026.

[22] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimed-vqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024.

[23] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. 43(2), January 2025.

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

[25] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024.

[26] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[27] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 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, 2023.

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[29] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore, December 2023. Association for Computational Linguistics.

[30] Daizong Liu, Xiaowen Cai, Junhao Dong, Zhongliang Guo, Xiaoye Qu, Runwei Guan, Xiang Fang, and Dengpan Ye. Attacking gray-box large vision-language models with adaptive svd-structured adversarial alignment. In *International Conference on Machine Learning*, 2026.

[31] Daizong Liu, Xiaowen Cai, Pan Zhou, Xiaoye Qu, Lichao Sun, and Wei Hu. Are large vision-language models robust to adversarial visual transformations? *IEEE Transactions on Information Forensics and Security*, 2026.

[32] Daizong Liu, Baoquan Chen, and Wei Hu. Spatial-spectral homogeneous attacks on physical-world large vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7114–7122, 2026.

[33] Daizong Liu, Junhao Dong, Xiang Fang, Hongyang He, Keyan Jin, Zhongliang Guo, Xiaoye Qu, and Keke Tang. Understanding and exploiting phase sensitivity for attacking large vision–language models. In *The 35th International Joint Conference on Artificial Intelligence*, 2026.

[34] Daizong Liu, Junhao Dong, Zhiyuan Ma, Xiaoye Qu, Xiang Fang, Runwei Guan, Keke Tang, Jianfeng Dong, and Yew-Soon Ong. Forbid your attention: Fooling multimodal large language models by selectively removing intrinsic focus in spectral domain. *IEEE Transactions on Multimedia*, 2026.

[35] Daizong Liu, Wangqin Liu, Xiaowen Cai, Pan Zhou, Runwei Guan, Xiaoye Qu, and Bo Du. Generating transferable attacks across large vision-language models using adversarial deformation learning. *Pattern Recognition*, page 113194, 2026.

[36] Daizong Liu, Tianyao Luo, Junhao Dong, Xiaowen Cai, Hao He, Weidong Wang, Runwei Guan, Jianfeng Dong, Pan Zhou, Bo Du, et al. Flexible targeted adversarial alignment with frequency-based visual focus for attacking large vision-language models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–18, 2026.

[37] Daizong Liu, Tianyao Luo, Shuwei Huang, Junyi Yang, Siyuan Liang, Xiaojun Jia, Chun Pong Lau, Jianfeng Dong, Yan Liu, Yang Bai, et al. A survey on ai agent security: Reasoning, acting, and self-evolving. *AI Plus*, 1(1):100004–122, 2026.

[38] Daizong Liu, Xiaoye Qu, Jianfeng Dong, Pan Zhou, Yu Cheng, Wei Wei, Zichuan Xu, and Yulai Xie. Context-aware biaffine localizing network for temporal sentence grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11235–11244, 2021.

[39] Daizong Liu, Xiaoye Qu, Xiao-Yang Liu, Jianfeng Dong, Pan Zhou, and Zichuan Xu. Jointly cross-and self-modal graph attention network for query-based moment localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4070–4078, 2020.

[40] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Yu Cheng, and Wei Hu. A survey of attacks on large vision–language models: Resources, advances, and future trends. *IEEE transactions on neural networks and learning systems*, 2025.

[41] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[42] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024.

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

[44] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024.

[45] Lin Lu, Zhigang Zuo, Ziji Sheng, and Pan Zhou. Merger-as-a-stealer: Stealing targeted PII from aligned LLMs with model merging. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5806–5825, Suzhou, China, November 2025. Association for Computational Linguistics.

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

[47] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[48] Sahar Salimpour, Lei Fu, Kajetan Rachwał, Pascal Bertrand, Kevin O'Sullivan, Robert Jakob, Farhad Keramat, Leonardo Militano, Giovanni Toffetti, Harry Edelman, et al. Towards embodied agentic ai: Review and classification of llm-and vlm-driven robot autonomy and interaction. *arXiv preprint arXiv:2508.05294*, 2025.

[49] Ziji Sheng, Guiyao Tie, Weidong Wang, Pan Zhou, and Daizong Liu. LearnerCoMPASS: Intelligent tutoring system with dynamic cognitive diagnosis and multi-model path planning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9016–9042, San Diego, California, United States, July 2026. Association for Computational Linguistics.

[50] Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. *arXiv preprint arXiv:2602.22394*, 2026.

[51] Daiki Shiono, Shumpei Miyawaki, Ryota Tanaka, and Jun Suzuki. Instruction-following evaluation of large vision-language models. *New Generation Computing*, 44(1):7, 2026.

[52] Gurpreet Singh. Vision–language foundation models and multimodal large language models: A comprehensive survey of architectures, benchmarks, and open challenges. 2026.

[53] Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in lvlms via attention imbalance rectification. *arXiv preprint arXiv:2603.24058*, 2026.

[54] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

[55] Guiyao Tie, Xueyang Zhou, Tianhe Gu, Ruihang Zhang, Chaoran Hu, Sizhe Zhang, Mengqu Sun, Yan Zhang, Pan Zhou, and Lichao Sun. Mmmr: Benchmarking massive multi-modal reasoning tasks. *arXiv e-prints*, pages arXiv–2505, 2025.

[56] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

[57] Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3225–3234, 2025.

[58] Chenyu Wang, Weixin Luo, Sixun Dong, Xiaohua Xuan, Zhengxin Li, Lin Ma, and Shenghua Gao. Mllm-tool: A multimodal large language model for tool agent learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6678–6687. IEEE, 2025.

[59] Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

[60] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[61] Hai Yan, Haijian Ma, Xiaowen Cai, Daizong Liu, Zenghui Yuan, Xiaoye Qu, Jianfeng Dong, Runwei Guan, Xiang Fang, Hongyang He, et al. Fit the distribution: Cross-image/prompt adversarial attacks on multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[62] Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. In *European Conference on Computer Vision*, pages 37–53. Springer, 2024.

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

[64] Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12944–12953, 2024.

[65] Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816, 2024.

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

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

A Limitation and Future Work

SIGMA is primarily designed for entity-conditioned hallucination mitigation, and is most directly applicable when the queried object is explicitly provided or can be reliably derived from the input question, candidate set, or object vocabulary. For multi-object recognition and open-ended generation, benchmark-specific entity interfaces are required, which may limit direct applicability in more unconstrained scenarios. In addition, mask quality depends on the underlying vision encoder and reliable entity resolution. Small-target results demonstrate relative evidence enrichment rather than reliable fine-grained localization, with only 13.75% absolute pointing accuracy. Genuinely ambiguous or unresolvable multi-entity references remain a limitation of the current entity-conditioning interface. Future work will extend SIGMA toward dynamic multi-entity masking, more general object-token-level intervention in free-form generation, and stronger adaptation across heterogeneous LVLM visual encoders.

B Algorithms

Appendix B provides the algorithmic details of SIGMA. Algorithm 2 summarizes the complete inference procedure with uncertainty-triggered three-stream assembly. Given an image-question pair, SIGMA first constructs the support and complement masks and then obtains the global, support, and complement visual streams through masked self-vision encoding. During decoding, the global stream serves as the default path. The support stream is activated only when the full-vocabulary or candidate-answer entropy indicates an uncertain decision, and the complement stream is further invoked only when the support-corrected candidate state remains risky. The final logits are obtained by assembling the calibrated global logits with support and complement residuals, followed by a lightweight plausibility constraint.

Algorithm 3 details the DUALMASK subroutine used in Algorithm 2. It constructs query-conditioned masks from the LVLM's own vision encoder without introducing external detectors, segmenters, or auxiliary visual experts. Specifically, the image-driven branch retrieves visually plausible concepts from the global image representation, while the question-aware branch extracts entity-conditioned prompts from the input question. Their patch-level relevance scores are sign-corrected, normalized, and fused into a support seed. A CLS-based shortcut prior is then used to separate entity-grounded evidence from shortcut-prone contextual regions, producing the support mask and complement mask used for tri-stream visual encoding.

Table 8: Complement-branch scheduler used by NEEDCOMPLEMENT. The default policy activates the complement stream when any flag is true; a stricter variant may require at least two flags.

Flag	**Activation condition**
Disagreement	$\mathrm{top}(S_g) \neq \mathrm{top}(S_s)$
High candidate entropy	$H(S_s) \geq \tau_{\mathrm{bg-ent}}$
Low candidate margin	$\Delta(S_s) \leq \tau_{\mathrm{bg-mar}}$
Unsupported positive	$\pi^+(S_g) \geq \tau_g^+$ and $\pi^+(S_s) \leq \tau_s^+$

B.1 Single object mask visualization

Figure 5 provides qualitative examples of the single-object self-vision dual masks constructed by SIGMA. For each image, the queried entity is used to build an entity-conditioned support mask and a complementary context mask. The support mask highlights visual regions that are directly aligned with the queried object, while the complement mask preserves regions that are weakly aligned, query-inconsistent, or potentially shortcut-prone. These examples illustrate that the two masks are not intended to form a foreground–background segmentation pair. Instead, they provide two complementary evidence views for subsequent three-stream encoding and logit-level assembly.

We visualize results on two representative LVLM backbones with different self vision encoders. For LLaVA-1.5, the masks are generated from its CLIP-based visual representation, whereas for Qwen3-VL, the masks are generated from its SigLIP-based visual representation. Across examples, the support masks tend to concentrate on the queried object or its discriminative parts, while the

Algorithm 2 SIGMA inference with uncertainty-triggered three-stream assembly

Require: Image x, question q, initial tokens y_0, LVLM $\mathcal{M}_\theta$, self vision encoder $\mathcal{E}$, maximum length T
Require: Uncertainty thresholds $\tau_{\text{full}}, \tau_{\text{cand}}$, calibration threshold τ_{cal}, calibration strength η
Require: Fusion weights α, β_s, β_c, plausibility threshold τ_p, plausibility penalty κ

$(M_s, M_c) \leftarrow \textsc{DualMask}(x, q)$
$(v_g, v_s, v_c) \leftarrow \textsc{TriEncode}(x, M_s, M_c)$
$y \leftarrow y_0$; initialize branch caches $K_g, K_s, K_c \leftarrow \emptyset$
for $i = 1$ **to** T **do**
  $(z_g, K_g) \leftarrow \mathcal{M}_\theta(y, v_g, K_g)$ ▷ default global stream
  $C_i \leftarrow \textsc{CandidateSet}(q, z_g)$; $P_i^+ \leftarrow \textsc{PositiveSubset}(q, C_i)$
  $H_{\text{cand}} \leftarrow -\sum_{a \in C_i} \text{softmax}(z_g|_{C_i})_a \log \text{softmax}(z_g|_{C_i})_a$
  $H_{\text{full}} \leftarrow -\sum_v \text{softmax}(z_g)_v \log \text{softmax}(z_g)_v$
  $u_i \leftarrow \mathbf{1}[H_{\text{full}} > \tau_{\text{full}} \ \vee \ H_{\text{cand}} > \tau_{\text{cand}}]$
  if $u_i = 0$ **then**
    $z_{\text{final}} \leftarrow z_g$ ▷ keep the native global prediction
  else
    $\widetilde{z}_{g,v} \leftarrow z_{g,v} + \eta\, \mathbf{1}[v \in P_i^+]\, \mathbf{1}[H_{\text{cand}} \geq \tau_{\text{cal}}]$ for all tokens v
    $(z_s, K_s) \leftarrow \mathcal{M}_\theta(y, v_s, K_s)$ ▷ support-first refinement
    $\alpha_i \leftarrow \alpha\big(\frac{1}{2} + \frac{1}{2} \max\{H_{\text{full}}/8,\ H_{\text{cand}}/\log|C_i|\}\big)$
    $z_{s\text{-only}} \leftarrow \widetilde{z}_g + \alpha_i \beta_s (z_s - \widetilde{z}_g)$
    $S_g \leftarrow \textsc{CandidateState}(\widetilde{z}_g, C_i, P_i^+)$; $S_s \leftarrow \textsc{CandidateState}(z_{s\text{-only}}, C_i, P_i^+)$
    if $\textsc{NeedComplement}(S_g, S_s)$ **then**
      $(z_c, K_c) \leftarrow \mathcal{M}_\theta(y, v_c, K_c)$ ▷ complement-on-demand suppression
      $z_{\text{asm}} \leftarrow \widetilde{z}_g + \alpha_i\, [\beta_s (z_s - \widetilde{z}_g) - \beta_c (z_c - \widetilde{z}_g)]$
    else
      $z_{\text{asm}} \leftarrow z_{s\text{-only}}$
    end if
    $z_{\text{final},v} \leftarrow z_{\text{asm},v} - \kappa\, \mathbf{1}[\text{softmax}(\widetilde{z}_g)_v \leq \tau_p]$ for all tokens v
  end if
  $y_i \leftarrow \arg\max_v z_{\text{final},v}$; $y \leftarrow [y; y_i]$
  if $y_i =$ `<eos>` **then break**
  end if
end for
return y

complement masks retain surrounding context and alternative salient regions. This behavior supports the motivation of SIGMA: hallucination mitigation benefits from explicitly separating entity-grounded evidence from contextual cues that may bias object-existence judgments.

B.2 Multi-objects mask visualization

Figure 6 shows examples of multi-object self-vision dual masking. Unlike the single-object setting, where one queried entity corresponds to one support–complement pair, the multi-object setting constructs entity-conditioned masks separately for each candidate object. For example, in the fruit image, SIGMA builds individual masks for *banana* and *grape*; in the wildlife image, it builds separate masks for *cheetahs* and *antelopes*. The entity-specific support masks highlight visual evidence aligned with each queried object, while the corresponding complement masks retain weakly aligned or shortcut-prone regions that may bias the existence decision.

When a benchmark requires simultaneous reasoning over multiple candidate objects, these entity-level masks can be combined into a multi-object support view and a multi-object complement view. This combination is only an interface-level adaptation for multi-entity inputs: the underlying self-vision dual masking, tri-stream encoding, and uncertainty-triggered assembly remain unchanged. During decoding or per-object existence evaluation, each entity can still trigger support refinement and complement suppression independently.

Algorithm 3 Self-vision dual-mask construction

Require: Image x, question q, self vision encoder $\mathcal{E}$, concept vocabulary $\mathcal{E}$, top-k
Require: Question-aware weight λ_q, CLS-context weight λ_b, mask temperature τ_m
Require: Patch-token score polarity $\gamma_E \in \{-1, +1\}$ specified by the self-vision adapter

$[h_{\mathrm{cls}}, h_1, \ldots, h_N] \leftarrow \mathcal{E}(x)$
$c \leftarrow \mathrm{Norm}(W_v h_{\mathrm{cls}})$; $\quad p_j \leftarrow \mathrm{Norm}(W_v h_j)$ for $j = 1, \ldots, N$
$\mathcal{C}_k \leftarrow \mathrm{TopK}_{e \in \mathcal{E}}\ c^\top e$ ▷ image-driven concepts
$z_{\mathrm{img}} \leftarrow \mathrm{Agg}\{\mathrm{TextEnc}(\text{“a photo of } c'\text{”}) : c' \in \mathcal{C}_k\}$
$\mathcal{Q} \leftarrow$ EntityPrompts(q); $\quad z_{\mathrm{qa}} \leftarrow |\mathcal{Q}|^{-1} \sum_{r \in \mathcal{Q}} \mathrm{TextEnc}(r)$
for $j = 1$ **to** N **do**
    $s_j^{\mathrm{img}} \leftarrow p_j^\top z_{\mathrm{img}}$; $\quad s_j^{\mathrm{qa}} \leftarrow p_j^\top z_{\mathrm{qa}}$
end for
$r^{\mathrm{img}} \leftarrow \sigma\big(\mathrm{Std}(\gamma_E s^{\mathrm{img}})\big)$; $\quad r^{\mathrm{qa}} \leftarrow \sigma(\mathrm{Std}(\gamma_E s^{\mathrm{qa}}))$
$s^{\mathrm{sup}} \leftarrow (1 - \lambda_q)\sqrt{r^{\mathrm{img}} \odot r^{\mathrm{qa}}} + \lambda_q r^{\mathrm{qa}}$
$b \leftarrow \sigma\big(\mathrm{Std}([p_1^\top c, \ldots, p_N^\top c])\big) \odot (1 - s^{\mathrm{sup}})$ ▷ query-inconsistent CLS context prior
$m^{\mathrm{sup}} \leftarrow \mathrm{Smooth}\left(s^{\mathrm{sup}} \odot (1 - \lambda_b b)\right)$
$m^{\mathrm{comp}} \leftarrow \mathrm{Smooth}\left((1 - s^{\mathrm{sup}}) + \lambda_b b\right)$
$M_s \leftarrow \mathrm{NormMax}(\mathrm{Sparsemax}(m^{\mathrm{sup}}/\tau_m))$; $\quad M_c \leftarrow \mathrm{NormMax}(\mathrm{Sparsemax}(m^{\mathrm{comp}}/\tau_m))$
return (M_s, M_c)

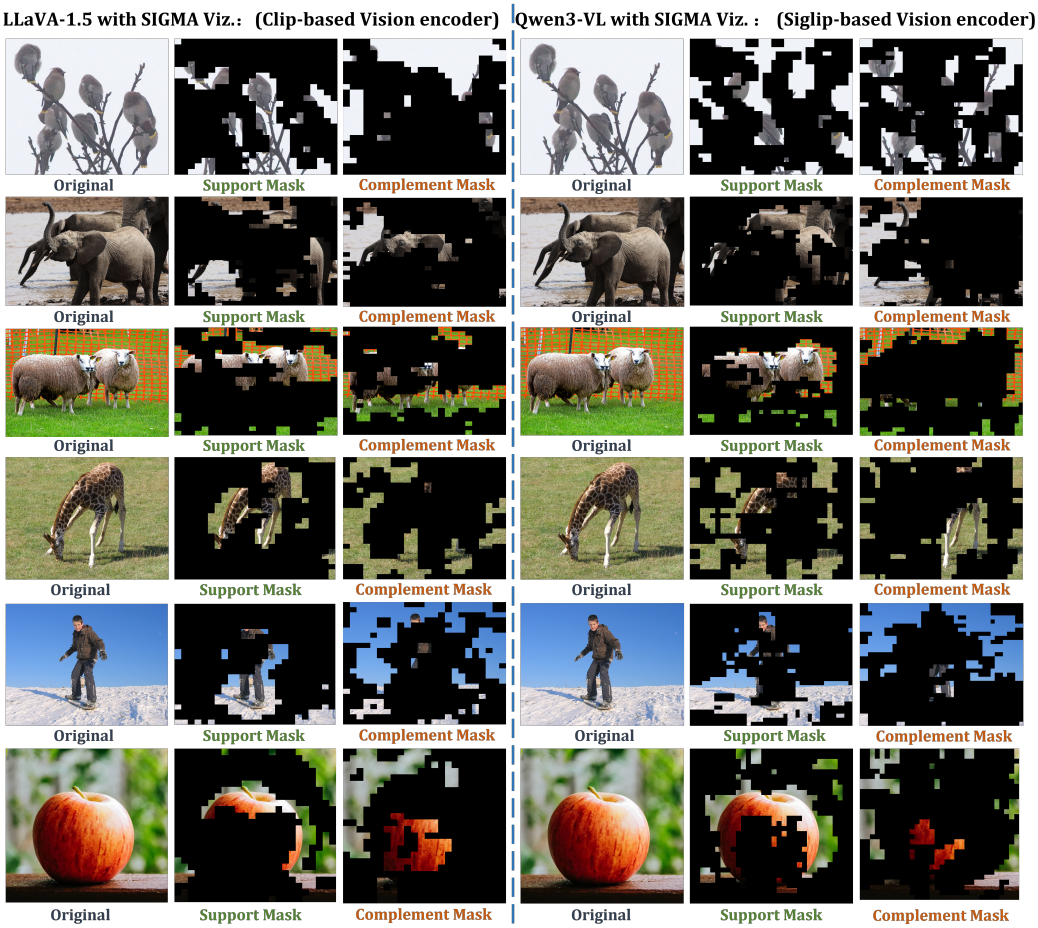

Figure 5: Single-object self-vision dual-mask visualization. Each group shows the original image, the support mask, and the complement mask generated by SIGMA. The left panel visualizes masks produced with the CLIP-based vision encoder of LLaVA-1.5, while the right panel visualizes masks produced with the SigLIP-based vision encoder of Qwen3-VL. The support mask emphasizes query-aligned object evidence, whereas the complement mask retains query-inconsistent or shortcut-prone contextual regions for suppressive correction during three-stream assembly.

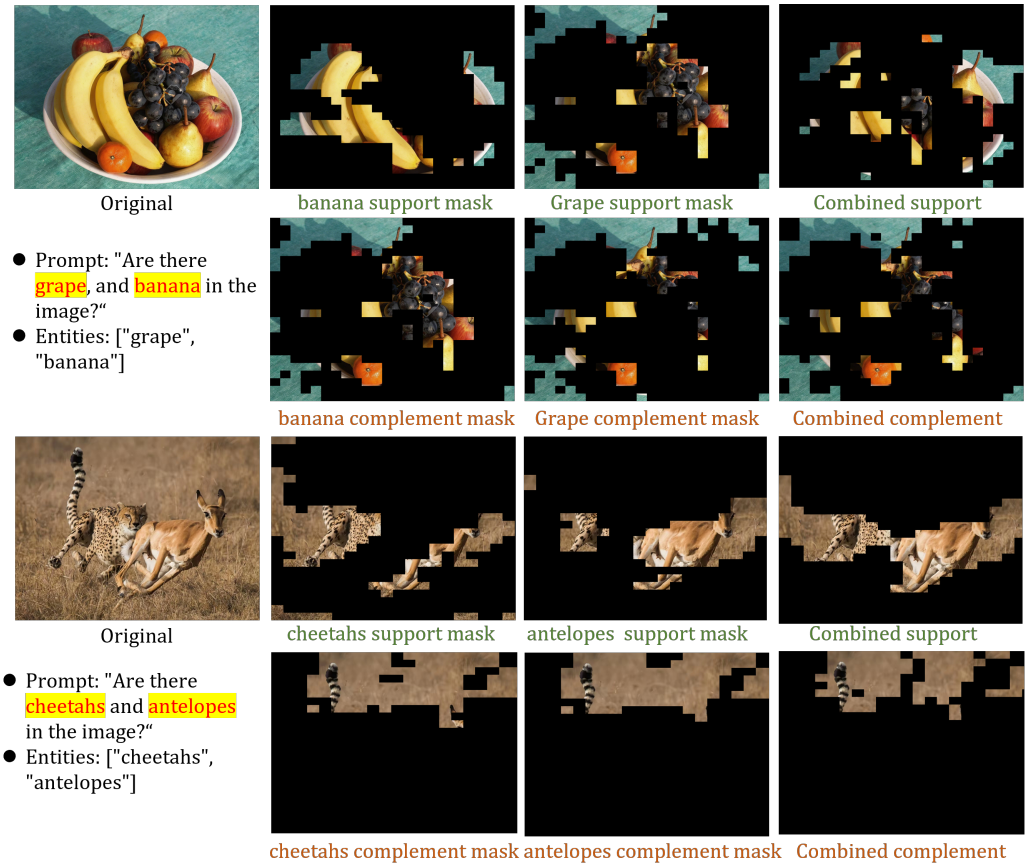

Figure 6: Multi-object self-vision dual-mask visualization. For each image, SIGMA constructs entity-conditioned support and complement masks for different queried objects, such as *banana*/*grape* and *cheetahs*/*antelopes*. The combined support mask aggregates object-aligned evidence across entities, while the combined complement mask preserves weakly aligned or shortcut-prone contextual regions. This visualization illustrates how SIGMA extends the same dual-mask mechanism from single-object queries to multi-object candidate settings without changing the core inference procedure.

C Experiment and Implementation details

C.1 Detailed Results of experiment

Detailed results on AMBER. Table 9 reports detailed results on the AMBER discriminative benchmark, including Overall performance and the Existence, Attribute, and Relation sub-tasks. For the Overall row, Accuracy, Precision, and Recall are computed as the macro average over the three sub-tasks, while Overall F1 is computed from the Overall Precision and Overall Recall. SIGMA improves Overall Accuracy and F1 across all four backbones, consistent with the main results in Table 2. Across sub-tasks, SIGMA improves F1 on Existence, Attribute, and Relation for every evaluated backbone, showing that the proposed self-vision evidence separation is not limited to object-existence questions. The improvement is often driven by higher precision, while recall remains competitive in most settings but may decrease on some attribute-oriented cases where the suppressive branch becomes more conservative.

Detailed results on ROPE. We further evaluate SIGMA on ROPE to examine its behavior under multi-object hallucination settings. Unlike single-entity existence benchmarks such as POPE, ROPE requires the model to reason over multiple candidate objects under adversarial or heterogeneous visual prompts. We adapt SIGMA only at the entity-conditioning interface by constructing entity-specific masks for candidate objects while keeping the self-vision masking, tri-stream encoding, and uncertainty-triggered assembly unchanged. As shown in Table 11, SIGMA consistently improves over Regular decoding across all five ROPE settings, indicating that the proposed evidence reorganization can extend to multi-object hallucination scenarios.

Parameter and component sensitivity. Table 10 reports all tested values and the F1 ranges from 11 unique one-factor configurations. Continuous support/complement weights and positive calibration change F1 by at most 0.40 points across the tested values. Masked-layer depth has a wider range (87.08–89.82 F1), indicating a structural sensitivity distinct from the continuous parameters. All additional matched analyses use frozen settings; clutter, overlap, and test subsets are not used for configuration selection.

C.2 Fusion rationale

Let I and Q denote normalized image-driven and query-aware relevance maps. A support-seed fusion should preserve query consistency, cross-source agreement, and controlled recovery when the image prior is weak. The geometric mean $G(I,Q) = \sqrt{I \odot Q}$ satisfies $G(I,0) = 0$ and $G(x,x) = x$, but gives $G(0,Q) = 0$. Equation (3) uses a query-directed residual relaxation:

$$S(I,Q) = (1-\lambda_q)\sqrt{I \odot Q} + \lambda_q Q. \tag{9}$$

For $0 < \lambda_q < 1$, it is bounded and monotonic, and satisfies

$$S(I,0) = 0, \qquad S(0,Q) = \lambda_q Q, \qquad S(x,x) = x. \tag{10}$$

The residual therefore retains query-supported evidence under a weak image prior. In contrast, arithmetic fusion $A(I,Q) = (I+Q)/2$ gives $A(I,0) = I/2 > 0$ for $I > 0$, while pure geometric fusion loses query-supported evidence as I approaches zero. These properties provide a principled design rationale, rather than a theorem establishing necessity, uniqueness, or universal optimality.

C.3 Mask reliability and difficult scenes

We evaluate spatial evidence allocation post hoc using COCO target boxes, which are never accessed by SIGMA during inference. Across 371 visible targets, support-mass lift is 1.691× (95% CI: [1.564, 1.838]). Support-pointing lift exceeds one and complement lift falls below one, supporting their asymmetric evidence roles. Label-blind spatial randomization gives $p < 10^{-5}$ for the reported enrichment/depletion effects.

The large small-target pointing lift is relative to a very small target area; the 13.75% absolute pointing accuracy indicates a remaining fine-grained localization limitation. The masks are evidence views rather than a foreground/background segmentation pair. Independently, the frozen high-clutter and high-overlap results in Table 4 test whether the correction remains useful when multiple or overlapping objects compete for attention. Annotations are used only for post-hoc stratification.

Table 9: Experimental results on the AMBER discriminative benchmark. We report the Overall discriminative performance alongside its three primary sub-tasks. The best results between the Base model and our SIGMA are highlighted in bold.

Model	Task	Regular				SIGMA (Ours)			
		Acc.	Prec.	Rec.	F1	Acc.	Prec.	Rec.	F1
LLaVA-1.5	***Overall***	64.0	70.4	**77.2**	73.6	**78.9**	**85.8**	77.0	**81.2**
	Existence	71.5	**100.0**	71.5	83.3	**80.0**	**100.0**	**80.0**	**88.9**
	Attribute	51.8	51.0	**88.0**	64.6	**77.7**	**87.7**	64.3	**74.2**
	Relation	68.7	60.2	72.0	65.6	**79.0**	**69.8**	**86.6**	**77.3**
InstructBLIP	***Overall***	74.4	77.5	80.2	78.8	**80.1**	**83.2**	**82.3**	**82.7**
	Existence	80.2	**100.0**	80.2	89.0	**85.0**	**100.0**	**85.0**	**91.9**
	Attribute	76.1	75.9	76.7	76.3	**79.8**	**81.2**	**77.5**	**79.3**
	Relation	66.8	56.7	83.6	67.6	**75.5**	**68.5**	**84.5**	**75.7**
Qwen3-VL	***Overall***	89.3	90.6	**90.6**	90.6	**91.7**	**95.4**	89.3	**92.3**
	Existence	90.1	**100.0**	90.1	94.7	**92.5**	**100.0**	**92.5**	**96.1**
	Attribute	87.5	87.5	**87.5**	87.5	**90.3**	**94.2**	85.8	**89.8**
	Relation	90.4	84.3	**94.3**	89.0	**92.1**	**92.1**	89.5	**90.8**
InternVL3.5-8B	***Overall***	88.7	90.3	**90.1**	90.2	**91.2**	**95.0**	88.6	**91.7**
	Existence	89.1	**100.0**	89.1	94.2	**92.2**	**100.0**	**92.2**	**95.9**
	Attribute	86.6	87.1	**86.1**	86.6	**89.4**	**93.5**	84.5	**88.8**
	Relation	90.4	83.8	**95.2**	89.1	**91.8**	**91.4**	89.2	**90.3**

Table 10: One-factor sensitivity across 11 unique configurations. All tested values and F1 ranges (%) are reported.

Parameter	Default	Tested values	F1 range
Support weight β_s	0.35	0.25, 0.35, 0.45	89.29–89.29
Complement weight β_c	0.80	0.60, 0.80, 1.00	89.29–89.68
Positive calibration η	0.12	0.08, 0.10, 0.12, 0.14, 0.16	88.89–89.29
Masked upper layers L_m	2	1, 2, 3	87.08–89.82

C.4 Paired uncertainty and recent POPE comparisons

For the additional POPE comparisons, predictions are aligned by question ID. Accuracy and nonlinear F1 are recomputed over 10,000 paired bootstrap resamples for each official 3,000-question split. These intervals quantify evaluation-sampling uncertainty rather than training-run variance.

SIGMA exceeds ONLY by 3.87 Accuracy points and 2.15 F1 points on the full Adversarial split. In a separate matched Random run, EAZY obtains 88.81 F1 and SIGMA obtains 88.31. The recent-baseline advantage is therefore setting-dependent.

C.5 Matched scheduling validation

A separate matched 900-example evaluation gives identical final answers for selective and always-on execution on all 900 examples. Selective complement activation is 55.8–56.8%, and mean latency is reduced by 14.75–17.37% relative to always-on execution. Together with Table 7, this supports conditional complement correction rather than indiscriminate suppression.

C.6 Entity interfaces and configuration policy

POPE and AMBER provide entity-bearing questions, ROPE provides object candidates and visual prompts, and CHAIR uses image-driven object candidates without requiring an explicit target noun. Once an entity or candidate is defined, the same support/complement construction and conditional assembly are applied. Explicit nouns, synonyms/paraphrases, and resolvable references can be handled by the entity interface; genuinely ambiguous or unresolvable references do not show a uniform gain and remain a limitation. For non-binary object-centric VQA, candidate entropy can use the compact top-k fallback described in the method.

Table 11: Detailed results on ROPE for multi-object hallucination mitigation.

Model	Setting	Decoding	Accuracy ↑	Precision	Recall	F1 Score ↑
LLaVA-1.5	*Adversarial-A*	Regular	10.69	38.51	14.82	21.40
		Ours	**22.34**	**40.80**	**24.20**	**30.38**
	Adversarial-B	Regular	9.88	37.09	13.92	20.24
		Ours	**20.66**	**39.50**	**22.70**	**28.83**
	Heterogeneous	Regular	2.80	9.46	4.00	5.62
		Ours	**7.07**	**10.20**	**7.60**	**8.71**
	Homogeneous	Regular	18.45	59.59	25.15	35.37
		Ours	**33.51**	**61.80**	**36.40**	**45.82**
	Mixed	Regular	7.02	23.54	9.82	13.85
		Ours	**15.00**	**25.80**	**16.80**	**20.35**

Table 12: Post-hoc mask reliability (95% CIs in brackets). Pointing accuracy is in percent; lift is area-normalized. Dashes indicate unreported values.

Metric	Visible targets ($n = 371$)	Small targets ($n = 80$)
Pointing accuracy (%)	43.13	13.75
Support-mass lift	1.691 [1.564, 1.838]	–
Support-pointing lift	3.201 [2.734, 3.763]	46.248 [23.607, 73.014]
Complement lift	0.6210 [0.5250, 0.7080]	0.0565 [0.0123, 0.1167]
Complement depletion	–	0.002805 [0.002164, 0.003605]

Table 13: Additional matched POPE comparisons. Differences are percentage points with 95% paired-bootstrap confidence intervals.

Comparison	Split	Δ Acc. [95% CI]	Δ F1 [95% CI]
SIGMA−Regular	Random	+4.43 [3.20, 5.70]	+3.37 [2.14, 4.64]
SIGMA−Regular	Popular	+5.90 [4.57, 7.20]	+4.33 [3.06, 5.62]
SIGMA−Regular	Adversarial	+7.40 [6.00, 8.77]	+4.96 [3.68, 6.24]
SIGMA−ONLY	Adversarial	+3.87 [2.77, 5.00]	+2.15 [1.14, 3.18]

SIGMA shares the native evidence-construction and global/support/complement routing policy across backbones. Encoder polarity, injected layer indices, and token interfaces are deterministic architecture mappings rather than benchmark-selected mechanisms. All new matched analyses use frozen settings; no clutter, overlap, or test subset is used to select configurations. Table 14 separates the shared mechanism from the interfaces described in this paper.

C.7 Additional CHAIR results

Caption length and inference cost. Table 15 supplements the main CHAIR table with the matched caption lengths and per-sample latency.

Comparison with OPERA. Table 16 preserves the direct InstructBLIP comparison. OPERA has lower CHAIR_i than SIGMA (13.4 vs. 17.1), whereas SIGMA has lower CHAIR_s (49.0 vs. 54.0) and higher object recall (72.1 vs. 69.3).

C.8 Implementation details

SIGMA constructs query-conditioned visual evidence through entity-aware self-vision masking. For single-target yes/no benchmarks such as POPE and AMBER, the mechanism described in the main paper can be applied directly: the queried object or attribute-entity phrase is extracted from

Table 14: Shared mechanism and benchmark interfaces. Architecture adapters map encoder polarity, layer indices, and token interfaces to each backbone; benchmark interfaces supply entities or candidates. Numerical settings from the controlled sensitivity study are given in Table 10.

Benchmark	Entity/candidate interface	Shared policy
POPE/AMBER	Entity from the visible question	Native dual masks; global default, support first, complement on demand
MME-Obj.	Object or attribute-entity phrase from the question	Same correction policy
ROPE	Provided candidates and visual prompts	Same correction policy
CHAIR	Image-driven object candidates	Same correction policy

Table 15: LLaVA-1.5: caption length and latency on CHAIR-500.

Method	Avg. words	Latency (s)
Regular	48.670	2.370
ONLY	48.234	2.480
EAZY	48.934	8.758
SIGMA	49.306	3.544

Table 16: InstructBLIP: CHAIR results (%).

Method	$C_s \downarrow$	$C_i \downarrow$	Rec.↑
Regular	54.0	18.1	71.1
OPERA	54.0	13.4	69.3
SIGMA	49.0	17.1	72.1

the benchmark-visible question, and one pair of support/complement masks is constructed for the corresponding entity. No benchmark-specific supervision, ground-truth label, reference caption, or image-specific annotation is used for mask construction or branch activation.

For ROPE, the main difference lies in the entity-conditioning interface rather than the core mechanism. ROPE provides candidate object classes and visual prompts, and the natural-language input may not always correspond to a single explicit queried object. We therefore combine the benchmark-provided candidate classes with the visual prompt to construct entity-specific masks for different candidate objects. During assembly, each object's existence prediction is treated as an independent entity-conditioned decision: the uncertainty trigger, support refinement, complement suppression, and logit fusion are applied separately for each queried object. The self-vision dual masking and uncertainty-triggered assembly mechanisms remain unchanged. Table 17 summarizes the benchmark-specific interfaces used in our experiments.

Leakage prevention. For all benchmarks, SIGMA constructs masks only from information available at inference time: the input image, the benchmark-visible question, candidate classes explicitly provided by the benchmark, visual prompts, or fixed public object vocabularies. Ground-truth labels, reference captions, and image-specific object annotations are never used for mask construction or branch activation. When prompt reformulation is required by a benchmark protocol, the same visible prompt is applied to all compared methods.

Scope of benchmark adaptation. The purpose of adaptation is to define the entity-conditioning interface required by SIGMA, not to change the benchmark objective or answer space. POPE and AMBER are evaluated in their native single-target yes/no formats. MME is evaluated on object-centric perception subsets, where the target noun phrase or attribute-entity phrase can be derived from the question. ROPE is evaluated with a candidate-aware entity interface derived from its provided candidate classes and visual prompts. In all cases, the decoding rule, evaluation target, and benchmark-visible information are kept consistent with the original benchmark setting.

Adaptation for baselines without official implementations. For Qwen3-VL and InternVL3.5, OPERA, VCD, and AGLA do not provide official implementations. We therefore reproduce them by preserving each method's original intervention mechanism while adapting only the model-specific interfaces. For AGLA, we keep its external image–prompt matching module and implement the corresponding logit-assembly stage for Qwen3-VL and InternVL3.5. For VCD, we port the official open-source implementation to the two backbones without modifying the visual contrastive decoding mechanism. OPERA is the most difficult to transfer because its implementation depends on model-specific transformer internals and beam-search decoding. To make it compatible with Qwen3-VL

Table 17: Benchmark-specific entity-conditioning protocol for SIGMA. POPE and AMBER require no special mechanism adaptation beyond extracting the queried entity from the benchmark-visible question. ROPE only adapts the entity-conditioning interface for multi-object visual-prompted evaluation. Ground-truth labels, reference captions, and image-specific annotations are never used for mask construction or branch activation.

Benchmark	Original task format	Relation to SIGMA	Entity-conditioning interface	Mask and assembly protocol	Leakage control
POPE	Single-object existence question with yes/no ground truth under Random, Popular, and Adversarial splits.	Directly matches SIGMA's single-entity query setting.	Extract the queried object from the question, e.g., *"Is there a [object] in the image?"*.	Construct one support mask and one complement mask for the queried entity. The candidate-answer set is fixed to {yes, no}, and uncertainty-triggered assembly is applied during answer decoding.	Only the image and question are used. The yes/no label is used only for evaluation.
AMBER	Discriminative yes/no questions covering existence, attribute, and relation hallucination.	Directly matches SIGMA when the queried entity or attribute-entity phrase is explicit in the question.	Extract the target object, attribute-object phrase, or relation-centered entity phrase from the benchmark-visible question.	Construct one entity-conditioned support/complement mask pair for each question. The candidate-answer set remains {yes, no}; no benchmark-specific mechanism change is introduced.	Ground-truth answers and annotations are used only for metric computation.
MME	Yes/no questions across perception and cognition tasks, including existence, count, position, color, OCR, and reasoning-related subsets.	Matches SIGMA for object-centric perception tasks, but not all MME questions require object-level patch discrimination.	For object-centric questions, extract the target noun phrase or attribute-entity phrase from the question, e.g., *"red bus"* or *"dog on the left"*.	Apply dual-mask construction to object-centric perception subsets. For non-object-centric or unparseable questions, the global stream is used or results are reported separately. The candidate-answer set remains {yes, no}.	The ground-truth yes/no answer and official annotations are not used to build masks.
ROPE	Multi-object hallucination benchmark with candidate object sets and visual prompts.	Requires an explicit multi-entity interface, but the core SIGMA mechanism is unchanged.	Use the benchmark-provided candidate classes together with the visual prompt to define entity-conditioned queries for different candidate objects.	Construct masks for different candidate objects using the candidate class set and prompt information. During assembly, each object's existence decision separately triggers uncertainty estimation, support refinement, complement suppression, and logit fusion.	Ground-truth target objects are not used for mask construction or branch activation. All compared methods use the same visible prompt.
CHAIR	Open-ended image caption generation evaluated by CHAIR_s, CHAIR_i, and object recall.	Extends SIGMA to generation without an explicit question-aware entity.	Uses image-retrieved concepts to define object candidates without requiring an explicit queried entity.	Constructs support and complement evidence views using image-driven concept information, retaining the same self-vision injection and uncertainty-triggered assembly.	Reference captions, ground-truth object labels, and image-specific annotations are used only for post-hoc CHAIR evaluation and never for mask construction.

and InternVL3.5, we reproduce OPERA under greedy decoding rather than its original beam-search setting. This may slightly underestimate OPERA's performance, but its over-trust penalty and retrospection-allocation mechanism are kept unchanged.

C.9 Compute Resources

All experiments are conducted on a single NVIDIA A800 GPU with CUDA 12.2 and 64GB system RAM. SIGMA is a training-free inference-time method, and therefore does not involve model training, optimizer updates, or training epochs. All benchmark evaluations are performed in a sample-wise inference setting rather than batched training-style computation. For each benchmark and backbone, we report the average latency per sample in the main paper, and use the same hardware setting to compute latency ratios against Regular decoding and representative training-free baselines.

C.10 Existing Asset Licenses and Terms of Use

We use only publicly available benchmarks, LVLM backbones, and baseline implementations. All datasets, models, and baseline methods are credited by citing their original papers in the main text and references. We follow the corresponding licenses and terms of use of each asset, including the

benchmark datasets used for evaluation, the publicly released LVLM checkpoints, and the available baseline codebases. No existing dataset is re-packaged or redistributed as a new dataset in this work, and all ground-truth labels, reference captions, and image-specific annotations are used only for evaluation.

D Broader Impacts.

This work aims to improve the reliability of LVLMs by reducing object hallucination, which may benefit safety-sensitive applications such as autonomous driving, robotics, and medical image analysis where visually ungrounded object predictions can lead to harmful decisions. At the same time, SIGMA should not be viewed as a complete safety solution: incorrect masks, entity extraction errors, or failures of the underlying LVLM may still produce misleading outputs, and improved hallucination mitigation could encourage over-reliance on LVLMs in high-stakes scenarios. We therefore recommend that SIGMA be used together with task-specific validation, human oversight, and deployment-time safety checks when applied beyond benchmark evaluation.

E Declaration of LLM Usage

This work studies LVLMs as the target models for hallucination mitigation and evaluates SIGMA on publicly available LVLM backbones, as described in the experimental setup. During manuscript preparation, LLM-based tools were used only for language polishing, editing, and formatting assistance. They were not used to generate datasets, labels, experimental results, core methodological ideas, algorithmic designs, or scientific claims. All technical content, experiments, analyses, and conclusions were designed, conducted, and verified by the authors.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions: a training-free SIGMA framework that uses self-vision dual masking, uncertainty-triggered assembly, and logit-level correction for object hallucination mitigation. These claims are supported by the method design and experiments across multiple LVLM backbones and hallucination-oriented benchmarks, while the scope is framed around object hallucination and entity-conditioned visual evidence separation.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitation and Future Work" section in Appendix discussing the main scope limitations of SIGMA, including its dependence on explicitly provided or reliably extracted entities, the need for benchmark-specific interfaces in multi-object and open-ended settings, and the sensitivity of mask quality to the underlying vision encoder and entity extraction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present formal theoretical results such as theorems, lemmas, or propositions. The equations and algorithms are used to define the proposed masking, uncertainty-triggering, and logit-assembly procedures, and the claims are supported empirically rather than through formal proof.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the main information needed to reproduce the experimental results, including the SIGMA inference procedure, dual-mask construction, tri-stream encoding, uncertainty-triggered assembly, evaluated benchmarks, LVLM backbones, baselines, evaluation metrics, and unified prompt/preprocessing/decoding/parsing protocols. Additional algorithmic and benchmark-specific implementation details are provided in the appendix to clarify how entity-conditioned masking and baseline adaptations are applied.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not currently provide a public code repository at the submission stage. The experiments are conducted on public benchmarks and publicly available LVLM backbones, and the paper provides algorithmic details, evaluation protocols, and benchmark-specific implementation descriptions; upon acceptance, we will release the source code and reproduction scripts through a publicly accessible GitHub repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the experimental settings needed to understand the results, including the evaluated benchmarks and metrics, LVLM backbones, baselines, shared preprocessing, prompt templates, decoding length, answer parsing protocol, and latency measurement procedure. Since SIGMA is a training-free inference method, optimizer and training-schedule details are not applicable; more detailed benchmark-specific entity-conditioning protocols, implementation details, and baseline adaptation procedures are provided in Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Appendix C.4 reports 95% paired-bootstrap confidence intervals for additional POPE performance differences, using 10,000 resamples aligned by question ID. Appendix C.3 reports confidence intervals and spatial-randomization evidence for mask reliability. These analyses quantify evaluation-sampling uncertainty for selected comparisons; they do not establish training-run variability or provide uncertainty estimates for every benchmark and latency measurement.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports average latency per sample and latency ratios for each benchmark and backbone in the main experimental tables. We additionally provide the compute-resource details in the appendix, including that all experiments are run on a single NVIDIA A800 GPU with CUDA 12.2 and 64GB system RAM, under a sample-wise inference setting without training or optimizer-based computation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and believe that the research conforms to it. The work focuses on mitigating object hallucination in LVLMs using public benchmarks and publicly available model backbones, does not involve human-subject experiments or private/sensitive data collection, and uses ground-truth annotations only for evaluation rather than for mask construction or inference.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a broader-impact discussion in Appendix covering the positive impact of reducing object hallucination in reliability- and safety-sensitive LVLM applications, as well as potential negative impacts such as over-reliance on corrected LVLM outputs and remaining failure cases caused by imperfect masks, entity extraction, or underlying model errors.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: The paper does not introduce or release new high-risk pretrained models, image generators, scraped datasets, or other data assets that would require dedicated release safeguards. The work proposes a training-free inference-time hallucination mitigation method evaluated on public benchmarks and publicly available LVLM backbones; any future code release will contain implementation and reproduction scripts rather than newly trained model weights or newly collected sensitive datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper credits the existing benchmarks, LVLM backbones, and baseline methods used in the experiments through citations, including the evaluated datasets, publicly released LVLM checkpoints, and training-free baseline methods. We additionally provide an appendix subsection on existing asset licenses and terms of use, stating that we use only publicly available assets, follow their corresponding licenses and terms, do not re-package or redistribute existing datasets, and use ground-truth annotations only for evaluation.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce or release new datasets, model checkpoints, pretrained models, scraped data, or benchmark assets. The work proposes a training-free inference-time method evaluated on existing public benchmarks and publicly available LVLM backbones; upon acceptance, we will release the implementation code and reproduction scripts with documentation through a publicly accessible GitHub repository.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing experiments or research with human subjects. All experiments are conducted through offline evaluation on existing public benchmarks and publicly available LVLM backbones, so there are no participant instructions, screenshots, consent procedures, or compensation details to report.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments are conducted through offline evaluation on existing public benchmarks and publicly available LVLM backbones, so there are no study participants, participant risks, disclosure procedures, consent procedures, or IRB approvals to report.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper describes the use of publicly available LVLMs as the target models and evaluation backbones for the proposed hallucination-mitigation framework. We additionally provide an appendix declaration stating that any LLM-based assistance during manuscript preparation was limited to language polishing, editing, and formatting, and did not affect the core methodology, experiments, scientific claims, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.