HT-SCD: SELF-CONTRASTIVE DECODING VIA HALLUCINATION TRAJECTORIES FOR MITIGATING OBJECT HALLUCINATIONS IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) demonstrate strong capabilities in image captioning, visual question answering, and multimodal reasoning, yet remain prone to object hallucinations. Existing mitigation methods often require additional training or external detectors, or incur substantial GPU memory overhead. We introduce HT-SCD, a training-free self-contrastive decoding method that uses hallucination trajectories to detect and mitigate object hallucinations online during autoregressive generation, without external models. Whereas conventional contrastive decoding methods construct a generic hallucination-prone negative branch in advance, HT-SCD first identifies specific hallucinations in the generated content and then uses the resulting hallucination trajectories to perform targeted rollback and correction through contrastive decoding. HT-SCD addresses the limited specificity and high computational cost of conventional contrastive decoding, and substantially outperforms existing contrastive decoding methods on MSCOCO-CHAIR and AMBER-G, two benchmarks for hallucination evaluation in open-ended visual generation.

1 INTRODUCTION

Recent advances in large vision-language models, including LLaVA (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2024), and Qwen-VL (Bai et al., 2023), have enabled strong open-ended generation capabilities in image captioning, visual question answering, and multimodal reasoning. However, numerous studies have shown that these models frequently generate content inconsistent with the input image, a phenomenon known as multimodal hallucination. Prior work generally attributes hallucinations in LVLMs to language priors and co-occurrence biases in the training data, which favor frequently co-occurring objects or descriptions when visual evidence is insufficient. Although existing methods alleviate object hallucinations to some extent, they struggle to jointly provide training-free operation, independence from external detectors, online interruption of hallucination accumulation, and targeted correction.

To address these challenges, we propose Self-Contrastive Decoding via Hallucination Trajectories (HT-SCD), a new decoding strategy. Our experiments demonstrate that HT-SCD detects object hallucinations online during autoregressive generation and uses their trajectories to perform immediate, targeted correction. It avoids several limitations of conventional contrastive decoding: a mismatch between the negative branch and the hallucinations actually generated, degradation of useful visual information, and high computational overhead. We also develop a visual evidence localization method that quantifies the contribution of visual tokens to the predicted word for object hallucination detection. Our main contributions are as follows:

(1) We introduce contrastive decoding based on hallucination trajectories. The predictive distributions at hallucinated object positions in the current generation guide sentence-level rollback and regeneration, directly linking the correction signal to the observed generation history.

(2) We develop an internal hallucination detector based on single-layer gradient weighting. It weights image–text representation similarity by visual sensitivity scores conditioned on the target word, without additional training or external detection models.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(3) We evaluate HT-SCD across three vision-language models and two object hallucination benchmarks, analyzing correction effectiveness under different negative branches, visual localization quality, inference overhead, and failure modes.

2 Related Work

Existing hallucination mitigation methods broadly fall into three categories: training-time intervention, decoding-time intervention, and post-hoc verification. Training-time methods typically modify the model distribution through higher-quality image–text data and training (Sun et al., 2024; Yu et al., 2024; Li et al., 2023; Zhang et al., 2024; Yang et al., 2025b; Chen et al., 2026). Post-hoc methods verify the output after autoregressive generation and revise it when hallucinations are detected (Zhou et al., 2024; Lee et al., 2024; Sahu et al., 2024; Wu et al., 2026). Decoding-time intervention has become a particularly active research direction in recent years. Within this category, contrastive decoding has emerged as a dominant approach because it requires no additional training and transfers readily across models.

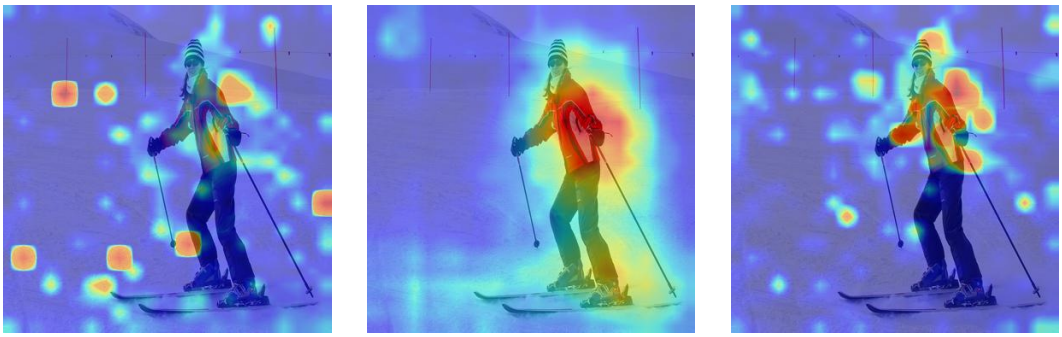
Contrastive decoding compares the predictive distribution under reliable visual conditioning with a hallucination-prone distribution, promoting visually grounded candidate tokens while suppressing content driven primarily by language priors. Most existing methods focus on constructing a generic hallucination-prone negative branch and can be grouped into three main categories. First, visual perturbation methods construct negative branches using diffusion noise, visual token sparsification, or frequency-domain information removal; examples include VCD, VASparse, and MFCD (Leng et al., 2024; Zhuang et al., 2025; Liu et al., 2025a). Second, methods based on multiscale visual evidence obtain finer-grained contrastive signals from local image regions, view trees, or multiscale visual representations, as in HALC, Game on Tree, AGLA, and SECOND (Chen et al., 2024; Zhuang et al., 2024; An et al., 2025; Park et al., 2025). Third, visual evidence calibration methods characterize the discrepancy between visual evidence and language priors through modality ablation, mutual information modeling, or positive and negative visual pathways; examples include M3ID, PAI, CMI-VLD, and PND (Favero et al., 2024; Liu et al., 2024; Fang et al., 2026; Jiang et al., 2026). Other approaches, such as SID (Huo et al., 2025) and DCD (Chen et al., 2026), construct generic hallucination-prone negative branches from internal model states and additionally learned visual projections, respectively. Octopus (Suo et al., 2025) and FuzzyCD (Kim et al., 2025) improve decoding adaptivity by dynamically selecting contrastive pathways or adjusting the contrastive strength.

Although these methods reduce hallucinations, their negative branches do not directly correspond to specific hallucinations arising in the current generation. Moreover, continuously contrasting multiple branches incurs additional computational cost. HT-SCD instead triggers correction only after detecting a hallucinated object word, avoiding the cost of continuously constructing a hallucination-prone negative branch. Its trajectory-based contrastive signal directly corresponds to the detected hallucination, enabling targeted mitigation.

3 Preliminary Analysis of Object Hallucination Detection

The contrastive decoding mechanism in HT-SCD requires a hallucination detector that operates without external models. We therefore investigate whether internal model representations can reveal the visual evidence underlying a target object word. Human judgments offer an intuitive motivation: to determine whether an object mention is hallucinated, a person first examines a local image region and then compares it with the object word. This suggests that object hallucination detection should begin by identifying the image regions on which an LVLM relies when predicting an object word.

Prior work (Guidotti et al., 2018; Ribeiro et al., 2016; Fong & Vedaldi, 2017) uses random masking or image perturbations to estimate how different image regions contribute to a prediction. Inspired by these approaches, we first investigate perturbations of visual representations to localize the image regions most relevant to LVLM inference. To reduce the substantial cost of propagating perturbations through all Transformer layers, we restrict perturbation analysis to a single layer. Specifically, we perturb the hidden states corresponding to visual tokens at an intermediate Transformer layer and use the final hidden state in that layer's output sequence as the output representation for assessing token importance. Comparing perturbations across layers reveals that localization is weakest in early layers,



Attention **Perturbation** **Gradient**

Figure 1: Visualizations of image-region attribution in LLaVA-1.5-7B using attention scores from all layers, visual representation perturbation, and gradient attribution.

followed by late layers, and strongest in middle and moderately late layers. These findings suggest that the middle and moderately late Transformer layers already encode useful spatial information for localizing task-relevant visual evidence, consistent with prior observations (Chen et al., 2025; Yang et al., 2025a; Yin et al., 2025). Figure 1 shows the localization results obtained using different methods.Appendix A provides a detailed analysis and layer-wise visualization results.

Perturbation-based localization remains costly even within a single Transformer layer.To obtain an efficient attribution signal conditioned on the target word, we instead use the gradient of the target-word probability read out from the selected layer. Under a first-order approximation, the gradient norm is proportional to the root mean squared probability change under small isotropic Gaussian perturbations. This motivates gradient-based localization without repeated perturbations. Appendix B derives this relation and distinguishes it from hidden-representation perturbation analysis.

4 METHODOLOGY

HT-SCD follows an online, closed-loop procedure consisting of generation, detection, rollback, and regeneration. After generating each sentence, the model checks it for hallucinated object words. If none are detected, the sentence is committed as a reliable output. Otherwise, decoding is restored to the state preceding that sentence, and the logits associated with the hallucinated object words serve as the corrective signal for contrastive regeneration. This procedure enables targeted hallucination mitigation.

4.1 OBJECT HALLUCINATION DETECTION

Hallucinations in LVLMs primarily arise when autoregressive generation produces descriptions or objects unsupported by visual evidence. Accordingly, our detector comprises two stages: localization and comparison. Localization identifies the visual regions on which the model relies when generating an object word. Comparison determines whether the generated object word is consistent with those regions and, consequently, whether it is hallucinated.

4.1.1 SINGLE-LAYER TRANSFORMER GRADIENT ATTRIBUTION

Building on the analysis in Section 3, we introduce a visual evidence localization method based on gradient attribution within a single Transformer layer. The method localizes the visual regions supporting an object-word prediction online. Gradients obtained through backpropagation measure the local sensitivity of the target-word probability to the input representation of each visual token at the selected layer.

Selecting the Scalar Target During autoregressive generation, once an object word has been identified, it is appended to the input sequence to predict the next token, as illustrated on the right of Figure 2. The hidden state corresponding to the object word is then at position t, whereas the

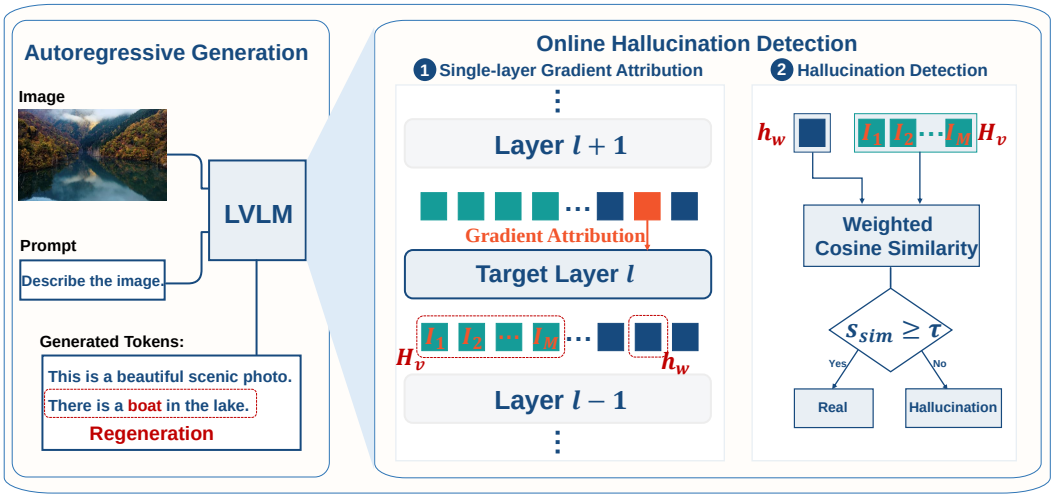


Figure 2: Overview of HT-SCD. Hallucinations are detected sentence by sentence during autoregressive generation. Reliable sentences are committed; sentences containing detected hallucinations trigger rollback to the preceding decoding state and regeneration through trajectory-based contrastive decoding. Object words in generated sentences are identified by vocabulary matching; see Appendix E.

hidden state at position $t-1$ produced the object-word prediction in the preceding forward pass. To identify the visual tokens supporting that prediction, gradient attribution must therefore use the vector at position $t-1$. Because attribution requires a scalar reference, we do not use the norm of the hidden vector as the objective. Instead, we construct a semantically meaningful probability P as the scalar target for backpropagation, giving the attribution an explicit probabilistic interpretation.

Previous studies (Halawi et al., 2024; Chuang et al., 2024; Elhoushi et al., 2024; Chen et al., 2023; Huo et al., 2024; Liang et al., 2024) suggest that intermediate hidden states can also predict the next token. Following this observation, we pass $h^l_{out,t-1}$, the output hidden vector at position $t-1$ of the target Transformer layer, through LayerNorm and the language modeling head (LM_Head). The resulting probability of the target word is used as the scalar target P. Here, w_t denotes the object word and id_{target} its token ID:

$$P\left(w_t|\text{prefix}\right) = \text{Softmax}\left(\text{LM_Head}\left(\text{Norm}\left(h^l_{out,t-1}\right)\right)\right)\left[id_{target}\right]$$

Gradient Attribution Within a Single Transformer Layer Given the scalar target P, we use backpropagation to compute its partial derivatives with respect to the input hidden states H^l_{in} of layer l. Unlike conventional gradient computation through the full model, gradient propagation here is restricted to the selected Transformer layer:

$$\nabla H^l_{in} = \frac{\partial P}{\partial H^l_{in}} = \left[\frac{\partial P}{\partial h^l_{in,1}}, \frac{\partial P}{\partial h^l_{in,2}}, \ldots, \frac{\partial P}{\partial h^l_{in,N}}\right]$$

Here, N is the total number of input tokens at the layer, including text and visual tokens. We retain only the gradients over the visual token interval $[1, M]$ to assess the local sensitivity of the target-word probability to these visual token representations. Backpropagation assigns each visual token a gradient vector $g^l_i \in R^d$. Taking the L_2 norm of each vector yields a scalar importance score I_i:

$$I_i = \|g^l_i\|_2 = \sqrt{\sum_{j=1}^{d}\left(g^l_{i,j}\right)^2}$$

The attribution map for the image is represented by the sequence of scores across all visual tokens: $\mathcal{S}_{\text{attr}} = [I_1, I_2, \ldots, I_M]$.

4.1.2 SIMILARITY SCORING AND HALLUCINATION DETECTION

We first define the object-word representation h_w and the visual token representations $H_v = \{v_1, v_2, \ldots, v_M\}$. After the object word is appended to the input sequence to predict the next token, h_w is the input hidden state at the object-word position in the target layer l, encoding the word's abstract features. The visual representations H_v are the input hidden states corresponding to the image and encode its abstract features. As shown on the right of Figure 2, both h_w and H_v are drawn from the input hidden states of the target layer.

The gradient attribution scores I_i from Section 4.1.1 measure the local sensitivity of the target-word probability P to the visual token representations H_v. We compute the similarity score s_{sim} as the weighted average of the cosine similarities between h_w and the visual representations H_v in feature space. A higher score indicates stronger support for the object word in the underlying visual features:

$$s_{sim} = \sum_{i=1}^{M} \frac{I_i}{\sum_{j=1}^{M} I_j} \cdot \frac{h_w^\top v_i}{\|h_w\|_2 \|v_i\|_2}$$

Finally, each object word is checked online against a predefined detection threshold τ . A word is classified as hallucinated if $s_{sim} < \tau$.

4.2 CONTRASTIVE DECODING VIA HALLUCINATION TRAJECTORIES

Most existing contrastive decoding methods mitigate hallucinations by constructing a generic hallucination-prone negative branch L_{hallu}. Such a branch L_{hallu} may not match the hallucinated content actually generated. Our detect-then-correct architecture instead uses the predictive distributions associated with detected hallucinated words directly as the negative branch L_{hallu}, eliminating the mismatch between the constructed negative branch L_{hallu} and the observed hallucination.

When an object word satisfies the hallucination criterion $s_{sim} < \tau$, the framework first rolls back the decoding state to the beginning of the sentence. Trajectory-based contrastive decoding then performs targeted online correction without modifying model parameters. If the regenerated sentence passes the hallucination check in Section 4.1.2, it is committed and generation proceeds to the next sentence. Otherwise, regeneration is repeated for up to n attempts. If hallucinations persist after n attempts, the final regenerated sentence is committed.

During regeneration, let L_{norm} denote the model's original predictive logits and L_{hallu} the sum of the logit vectors associated with all hallucinated object words in the preceding hallucinated sentence. We define the adjusted generation logits L_{new} as

$$\mathbf{L_{new}} = (\mathbf{1} + \lambda)\, \mathbf{L_{norm}} - \lambda \mathbf{L_{hallu}}$$

where λ controls the contrastive strength. The historical negative branch penalizes vocabulary preferences associated with detected hallucinations, while the current logits remain conditioned on the image and the evolving prefix. Applying this contrast throughout sentence regeneration can change the prefix and subsequent predictions, helping the model avoid hallucinated continuations. If hallucinations persist, the negative branch is updated from the regenerated sentence.

5 EXPERIMENTS

5.1 HALLUCINATION DETECTION

Evaluation. To assess whether the proposed detector distinguishes real object mentions from hallucinated ones, we formulate detection as object-level binary classification, with real objects as positive examples and hallucinated objects as negative examples. Following CHAIR, objects are

Table 1: Comparison with hallucination detection baselines in terms of AUROC and AUPR.

Metric	NLL	Entropy	SVAR	Contextual Lens	Internal Confidence	attention	Ours
AUROC	63.7	64.0	74.7	75.4	72.9	74.3	81.2
AUPR	84.9	85.0	91.2	90.7	89.3	89.6	92.4

Table 2: Comparison of hallucination mitigation methods on MSCOCO-CHAIR. C_s and C_i denote CHAIR_s and CHAIR_i, respectively; R denotes Recall.

Method	LLaVA-1.5-7B				Qwen2.5-VL-7B				InstructBLIP Vicuna-7B			
	$C_s \downarrow$	$C_i \downarrow$	$R \uparrow$	Len	$C_s \downarrow$	$C_i \downarrow$	$R \uparrow$	Len	$C_s \downarrow$	$C_i \downarrow$	$R \uparrow$	Len
Base	46.8	12.6	76.1	90.7	28.2	7.8	58.9	145.4	46.2	12.7	73.6	101.4
OPERA	45.2	13.2	78.9	88.2	18.8	5.9	49.0	119.5	43.8	10.3	70.9	100.6
VTI	34.2	10.6	70.8	85.5	35.4	9.2	62.4	185.8	39.0	10.6	69.4	106.1
ONLY	57.6	17.8	75.6	87.7	19.4	6.6	52.1	113.7	55.4	17.4	71.5	99.9
MemVR	48.0	13.0	80.0	91.4	27.2	8.5	58.7	200.3	45.6	12.5	75.2	100.6
CMAC	36.0	9.2	76.3	95.1	27.2	8.5	57.1	144.6	59.4	16.5	76.1	101.2
EAZY	41.6	11.6	77.9	91.2	21.0	6.6	53.5	137.1	37.6	11.1	65.9	106.1
VCD	51.4	16.1	76.3	92.9	29.6	8.6	59.2	150.5	52.4	14.9	75.5	98.9
PAI	48.8	13.2	79.0	98.2	23.4	9.4	51.8	120.1	56.8	16.0	75.5	113.2
SID	41.4	12.1	76.5	79.2	33.4	10.5	61.2	129.5	53.4	16.6	72.1	94.8
VASparse	39.6	12.1	77.6	87.4	25.8	8.5	59.6	135.0	44.8	12.9	75.2	104.3
FuzzyCD	47.2	13.5	75.6	88.6	29.4	8.6	58.4	128.7	46.9	13.1	72.2	100.3
AGLA	38.4	11.8	77.1	86.4	21.6	5.8	57.4	134.9	44.6	14.0	72.7	104.3
Ours	**29.4**	**9.2**	**75.9**	**85.9**	**18.4**	**5.7**	**55.8**	**129.7**	**30.6**	**8.8**	**71.8**	**101.5**

extracted from generated descriptions and matched by exact string matching against the ground-truth object categories and their synonyms for each image. We report two widely used, threshold-independent classification metrics: the area under the receiver operating characteristic curve (AUROC) and the area under the precision–recall curve (AUPR). The evaluation set consists of 5,000 images randomly sampled from MSCOCO (Lin et al., 2014).

Baselines. We conduct experiments on LLaVA-1.5-7B and we compare our detector with representative approaches based on (1) token probabilities: negative log-likelihood (NLL) (Zhou et al., 2024) and entropy (Malinin & Gales, 2020); (2) logit-lens probabilities: Internal Confidence (Jiang et al., 2025a); (3) attention: Summed Visual Attention Ratio (SVAR) (Jiang et al., 2025b); (4) embedding similarity: Contextual Lens (Phukan et al., 2025) (5) We replace gradient attribution scores with attention scores while keeping the rest of the detection pipeline unchanged.

Results and discussion. AUROC measures the overall ability to distinguish hallucinated from non-hallucinated examples across thresholds, whereas AUPR summarizes the trade-off between precision and recall. As shown in Table 1, our method achieves the best results on both metrics. Among the baselines, Contextual Lens yields the highest AUROC of 75.4, and SVAR yields the highest AUPR of 91.2. Our method improves these values to 81.2 and 92.4, respectively. These results indicate more discriminative hallucination risk scores and a better precision–recall trade-off for hallucination detection.

5.2 Hallucination Mitigation

Evaluation. To assess whether HT-SCD reduces hallucinations in LVLM outputs, we use two widely adopted evaluation benchmarks: (1) the CHAIR metric (Rohrbach et al., 2018) on MSCOCO (Lin et al., 2014), which measures object hallucinations; and (2) AMBER (Wang et al., 2023), which comprises generative and discriminative components and covers both common objects and objects that frequently induce hallucinations.

Table 3: Comparison of hallucination mitigation methods on AMBER-G. C and Cov denote CHAIR and Cover, respectively.

Method	LLaVA-1.5-7B				Qwen2.5-VL-7B				InstructBLIP Vicuna-7B			
	$C\downarrow$	Cov↑	Hal↓	Cog↓	$C\downarrow$	Cov↑	Hal↓	Cog↓	$C\downarrow$	Cov↑	Hal↓	Cog↓
Base	6.7	50.2	28.6	4.3	4.0	56.1	21.4	1.6	7.1	54.0	34.6	3.7
OPERA	7.5	47.3	29.2	4.1	4.4	47.0	17.7	2.5	8.7	47.8	34.6	3.7
VTI	5.8	47.0	21.7	3.5	4.1	62.2	24.4	2.5	6.6	50.8	27.2	2.9
ONLY	10.1	47.5	38.5	4.9	4.1	56.1	23.6	2.1	10.7	49.6	43.4	5.2
MemVR	6.7	50.1	28.6	4.2	3.6	55.1	17.5	1.2	7.5	53.5	35.9	3.8
CMAC	6.3	53.5	32.1	4.0	3.5	53.6	17.1	1.3	8.8	55.7	44.6	6.7
EAZY	5.6	49.8	22.9	3.7	3.7	54.4	19.0	1.6	9.9	54.0	33.6	5.3
VCD	6.9	50.6	31.0	4.4	4.7	54.0	21.4	2.0	8.1	54.2	40.2	3.8
PAI	6.6	51.3	31.2	4.3	4.6	51.0	20.1	1.4	7.8	52.1	35.2	3.8
SID	4.7	50.7	21.2	2.6	5.4	54.3	24.3	1.7	9.9	51.4	43.3	6.0
VASparse	5.5	50.0	24.7	3.5	3.4	53.7	16.6	1.5	6.6	54.2	32.7	5.1
FuzzyCD	7.1	51.5	31.4	4.3	4.4	53.5	20.7	1.5	6.7	54.1	32.3	4.5
AGLA	5.7	50.6	25.4	3.5	3.4	54.5	16.6	1.7	6.9	53.5	33.3	4.9
Ours	**4.1**	**49.2**	**18.8**	**2.4**	**2.6**	**53.9**	**10.7**	**0.9**	**4.8**	**52.7**	**22.6**	**2.8**

Models and baselines. To evaluate applicability across LVLM architectures, we conduct experiments on LLaVA-1.5-7B, Qwen2.5-VL-7B, and InstructBLIP-Vicuna-7B. We compare HT-SCD with a range of state-of-the-art baselines: OPERA (Huang et al., 2024), VTI (Liu et al., 2025b), ONLY (Wan et al., 2025), MemVR (Zou et al., 2024), CMAC (Li et al., 2026), EAZY (Che et al., 2025), VCD (Leng et al., 2024), PAI (Liu et al., 2024), SID (Huo et al., 2025), VASparse (Zhuang et al., 2025), FuzzyCD (Kim et al., 2025), and AGLA (An et al., 2025).

MSCOCO-CHAIR results and discussion. As shown in Table 2, relative to the unmodified models, HT-SCD reduces CHAIRs by 37.2%, 34.8%, and 33.8%, and CHAIRi by 27.0%, 26.9%, and 30.7%, across the three models, respectively. Most baselines perform well on only one or two models, indicating limited generalization across architectures. By comparison, HT-SCD achieves the best CHAIRs and CHAIRi results across all three models when evaluated against the strongest baseline for each metric, demonstrating consistent object hallucination mitigation and cross-model generalization. Recall and output length remain broadly comparable to those of the baselines and the unmodified models.

AMBER-G results and discussion. Table 3 reports the cross-benchmark transfer performance of HT-SCD on AMBER-G. All hyperparameters are transferred directly from MSCOCO without retuning on AMBER-G. Relative to the unmodified models, HT-SCD reduces CHAIR by 38.8%, 35%, and 32.5%; Hal by 34.3%, 50%, and 34.7%; and Cog by 44.2%, 43.8%, and 24.3%, across the three models, respectively. HT-SCD also achieves substantially better results than the strongest baselines for CHAIR, Hal, and Cog across all three models. Its consistent reductions across multiple hallucination metrics without dataset-specific adaptation demonstrate effective cross-dataset transfer.

5.3 Ablation Studies

5.3.1 Effectiveness of Hallucination Trajectories and Detection

Table 4 evaluates the contributions of trajectory-based contrastive correction and the proposed hallucination detector through two groups of controlled ablations. We first vary the negative branch or correction strategy while retaining the proposed detector and rollback procedure. We then replace the detector while retaining trajectory-based contrastive correction and the same rollback procedure.

Hallucination trajectory negative branch. With the detector and rollback procedure fixed, HT-SCD achieves CHAIRs/CHAIRi of 29.4/9.2, outperforming the negative branches from VCD (38.4/11.9), SID (38.0/11.0), and AGLA (37.6/10.9), with comparable recall and output length. It also outperforms Resample (sampling from the original distribution), TokenOnly (penalizing

Table 4: CHAIR ablations of correction strategies and detectors in HT-SCD.

Method	$C_s \downarrow$	$C_i \downarrow$	$R \uparrow$	**Len**
Base	46.8	12.6	76.1	90.7
VCD	38.4	11.9	75.8	87.3
SID	38.0	11.0	75.9	86.9
AGLA	37.6	10.9	76.1	86.8
Resample	38.8	11.8	73.8	85.7
TokenOnly	39.6	12.1	77.1	87.1
Mean	31.0	9.3	76.5	86.1
RandomPos	38.2	11.2	73.8	85.7
SVAR	36.4	10.6	75.4	86.4
Contextual Lens	37.8	10.4	73.8	87.1
Ours	**29.4**	**9.2**	75.9	85.9

Table 5: Localization results across object size.

Method	PG-Box↑	EIB↑	R@5%↑
Small (Count: 166)			
Gradient	0.428	0.166	0.611
Attention	0.211	0.058	0.356
Perturbation	0.265	0.053	0.714
Medium (Count: 168)			
Gradient	0.482	0.309	0.358
Attention	0.339	0.146	0.279
Perturbation	0.631	0.183	0.514
Large (Count: 166)			
Gradient	0.699	0.607	0.132
Attention	0.542	0.474	0.111
Perturbation	0.795	0.506	0.159

Table 6: Results of the three visual attribution methods. Time is measured in seconds per object.

Method	**PG-Box↑**	**EIB↑**	**R@5%↑**	**Time↓**
Gradient	0.536	0.360	0.367	0.118
Attention	0.364	0.226	0.249	0.001
Perturbation	0.564	0.247	0.462	45.893

Table 7: Target-region coverage at different visual token selection ratios K/N.

Method	**R@1%↑**	**R@5%↑**	**R@20%↑**
Gradient	0.195	0.367	0.564
Attention	0.107	0.249	0.471
Perturbation	0.176	0.462	0.701

detected object words), and RandomPos (using logits from a random non-hallucinated position), supporting the use of full predictive distributions from detected hallucinated positions. Averaging rather than summing these logits (Mean) retains most gains (31.0/9.3), suggesting that the historical distributional signal is more important than the additional penalty strength from summation.

Hallucination detector. With trajectory-based correction and rollback unchanged, replacing our detector with SVAR or Contextual Lens yields CHAIRs/CHAIRi of 36.4/10.6 and 37.8/10.4, respectively. Our detector achieves lower hallucination rates and higher recall (75.9 versus 75.4 and 73.8), with similar output lengths. These results show that detector selection also contributes to downstream mitigation while preserving object coverage.

5.3.2 Effectiveness of Visual Localization

To evaluate single-layer gradient attribution, we conduct a controlled localization experiment at layer 16 of LLaVA-1.5-7B using ground-truth bounding boxes for 500 MSCOCO objects spanning all 80 categories. With human-written reference captions and teacher forcing, we fix the image, text prefix, target object word, and prediction position, comparing Gradient, Attention, and Perturbation on the same 24×24 visual grid. We measure localization accuracy, attribution concentration, and region coverage using PG-Box, EIB, and Recall@5%, respectively. Appendix D details the protocol.

Table 6 shows that Gradient outperforms Attention from the same layer in localization accuracy, attribution concentration, and region coverage: PG-Box, EIB, and Recall@5% increase from 0.364, 0.226, and 0.249 to 0.536, 0.360, and 0.367, respectively. Compared with Perturbation, Gradient achieves higher EIB, comparable PG-Box, and slightly lower Recall@5%, while substantially reducing computational cost. Perturbation takes an average of 45.893 s per object, approximately 389 times the runtime of Gradient. Gradient thus provides competitive localization quality at a much lower cost.

Table 5 shows that the advantage in attribution concentration holds across object scales. Gradient achieves the highest EIB in the Small, Medium, and Large groups, as well as the best PG-Box in the Small group. Perturbation, by contrast, leads in PG-Box for medium and large objects and in

Table 8: Computational overhead of HT-SCD and six other contrastive decoding methods.

Method	Latency (s/image)	Relative to base model	Throughput (tokens/s)	Peak GPU memory (GB)	LLM forward passes/token	Visual encodings per image
LLaVA-1.5-7B	2.87	1.00×	35.42	14.07	1.03	1.0
VASparse	3.06	1.06×	33.68	14.07	1.03	1.0
Ours	**4.38**	**1.53×**	**22.94**	**14.86**	**1.57**	**1.0**
PAI	5.34	1.86×	18.55	14.35	2.06	1.0
SID	5.59	1.95×	17.01	14.76	2.06	1.0
AGLA	5.78	2.01×	17.72	14.58	2.06	2.0
VCD	5.79	2.02×	17.73	14.45	2.06	2.0
FuzzyCD	15.60	5.43×	6.97	16.25	5.13	5.0

Recall@5% across all three groups. These results suggest that Gradient concentrates attribution more effectively within the target, whereas Perturbation provides better coverage of the target region.

Table 7 further shows that the relative performance of the two methods depends on the token selection budget. When only 1% of visual tokens are selected, Gradient achieves higher recall than Perturbation (0.195 versus 0.176), whereas Perturbation performs better at 5% and 20%. This suggests that Gradient ranks target-relevant tokens more effectively at the top, while Perturbation covers more of the target region as the selection budget increases. Gradient also consistently outperforms Attention at all three ratios, indicating that its advantage is not specific to a particular K/N setting.

Taken together, these results show that Gradient consistently outperforms Attention from the same layer and offers more concentrated attribution and greater computational efficiency than Perturbation. Although its region coverage remains lower at larger selection budgets, its balance between localization quality and computational cost supports its use in online hallucination detection.

5.3.3 Inference Overhead

We evaluate the computational overhead of HT-SCD, the unmodified model, and six training-free contrastive decoding methods under an identical protocol: LLaVA-1.5-7B, the same samples from MSCOCO val2014, greedy decoding, a maximum of 256 new tokens, and a single RTX A6000 GPU. The evaluation results are presented in Table 8. Among the contrastive decoding methods, HT-SCD has the second-lowest computational cost: 4.38 s per image, or 22.9 tokens/s, corresponding to 1.53× the latency of the original decoding procedure. In comparison, PAI, SID, AGLA, and VCD incur 1.86–2.02× latency, while FuzzyCD incurs 5.43×. This advantage stems from the on-demand computation in HT-SCD. It requires 1.57 decoder forward passes per token: one standard pass plus an amortized 0.57 passes for sentence-level rollback and regeneration triggered only by detected hallucinations. This is lower than all other contrastive branch methods except VASparse. Hallucination detection accounts for only a small fraction of the total computational cost, and peak GPU memory increases by just 0.79 GB.

6 Conclusion

We introduced HT-SCD, a self-contrastive decoding method based on hallucination trajectories, to mitigate object hallucinations in LVLMs. Without additional training or external detectors, HT-SCD integrates object hallucination detection and generation-time mitigation into a single autoregressive process, forming an online sentence-level loop of generation, detection, rollback, and regeneration. Overall, HT-SCD substantially outperforms existing methods in both object hallucination detection and mitigation. The results further demonstrate that internal hidden states and gradients in LVLMs can serve not only as detection signals but also as corrective evidence for contrastive decoding. This offers a new direction for multimodal hallucination mitigation without additional training or external models.

REFERENCES

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

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)*, pp. 21635–21644. IEEE, 2025.

Haoran Chen, Junyan Lin, Xinghao Chen, Yue Fan, Jianfeng Dong, Xin Jin, Hui Su, Jinlan Fu, and Xiaoyu Shen. Multimodal language models see better when they look shallower. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 6688–6706, 2025.

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models. *Advances in Neural Information Processing Systems*, 38:166734–166756, 2026.

Yanxi Chen, Xuchen Pan, Yaliang Li, Bolin Ding, and Jingren Zhou. Ee-llm: Large-scale training and inference of early-exit large language models with 3d parallelism. *arXiv preprint arXiv:2312.04916*, 2023.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, et al. Layerskip: Enabling early exit inference and self-speculative decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12622–12642, 2024.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *Advances in Neural Information Processing Systems*, 38:105233–105258, 2026.

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

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 3449–3457. IEEE, 2017.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

Danny Halawi, Jean-Stanislas Denain, and Jacob Steinhardt. Overthinking the truth: Understanding how language models process false demonstrations. In *International Conference on Learning Representations*, volume 2024, pp. 42749–42787, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427. IEEE, 2024.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pp. 24272–24295, 2025.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6801–6816, 2024.

Nick Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 63582–63605, 2025a.

Yubo Jiang, Yitong An, Xin Yang, Abudukelimu Wuerkaixi, Xuxin Cheng, Fengying Xie, Zhiguo Jiang, Cao Liu, Ke Zeng, and Haopeng Zhang. Breaking the illusion: When positive meets negative in multimodal decoding. *arXiv preprint arXiv:2605.06679*, 2026.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 25004–25014. IEEE, 2025b.

Jieun Kim, Jinmyeong Kim, Yoonji Kim, and Sung-Bae Cho. Fuzzy contrastive decoding to alleviate object hallucination in large vision-language models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20572–20581. IEEE, 2025.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404, 2024.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882. IEEE, 2024.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-modal attention calibration for lvlm hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40186–40196, 2026.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023.

Yaoyuan Liang, Zhuojun Cai, Jian Xu, Guanbo Huang, Yiran Wang, Xiao Liang, Jiahao Liu, Ziran Li, Jingang Wang, and Shao-Lun Huang. Unleashing region understanding in intermediate layers for mllm-based referring expression generation. *Advances in Neural Information Processing Systems*, 37:120578–120601, 2024.

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

Bingqian Liu, Fu Zhang, Guoqing Chen, and Jingwei Cheng. Multi-frequency contrastive decoding: Alleviating hallucinations for large vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28556–28572, 2025a.

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

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pp. 72402–72419, 2025b.

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*, pp. 125–140. Springer, 2024.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. *arXiv preprint arXiv:2506.08391*, 2025.

Anirudh Phukan, Divyansh, Harshit Kumar Morj, Vaishnavi, Apoorv Saxena, and Koustava Goswami. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 9661–9675, 2025.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

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*, pp. 4035–4045, 2018.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-llms via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pp. 13088–13110, 2024.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29904–29914. IEEE, 2025.

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)*, pp. 3225–3234. IEEE, 2025.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Tsung-Han Patrick Wu, Heekyung Lee, Jiaxin Ge, Joseph Gonzalez, Trevor Darrell, and David Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. *Advances in Neural Information Processing Systems*, 38:65749–65777, 2026.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29968–29978. IEEE, 2025a.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10610–10620. IEEE, 2025b.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9382–9391, 2025.

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*, pp. 13807–13816, 2024.

Mengxi Zhang, Wenhao Wu, Yu Lu, Yuxin Song, Kang Rong, Huanjin Yao, Jianbo Zhao, Fanglong Liu, Haocheng Feng, Jingdong Wang, et al. Automated multi-level preference for mllms. *Advances in Neural Information Processing Systems*, 37:26171–26194, 2024.

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. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Xianwei Zhuang, Zhihong Zhu, Zhanpeng Chen, Yuxin Xie, Liming Liang, and Yuexian Zou. Game on tree: Visual hallucination mitigation via coarse-to-fine view tree and game theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17984–18003, 2024.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4189–4199. IEEE, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

A LOCALIZATION VISUALIZATIONS

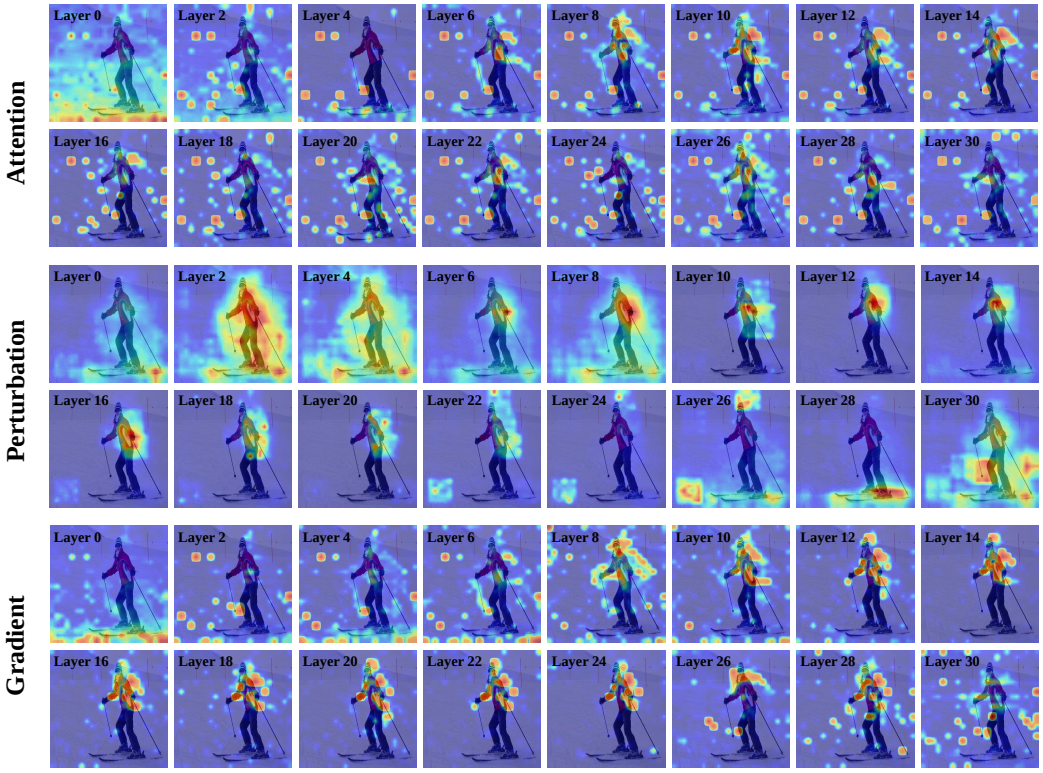


Figure 3: Visualizations of image-region attribution using different localization methods in LLaVA-1.5-7B. Compared with attention scores and visual representation perturbation, gradient attribution focuses more consistently on local regions relevant to the question, supporting its suitability for localizing visual evidence associated with target object words.

A.1 ANALYSIS OF LOCALIZATION ACROSS LAYERS

To further investigate how LVLMs use visual information across Transformer layers, we visualize salient image regions using attention scores, visual representation perturbation, and gradient attribution. The prompt is "What color is the person's top in the image?" Answering this question requires identifying the person and then localizing the region directly relevant to the question: the person's upper-body clothing. This task therefore provides a useful test of whether an attribution method captures visual evidence relevant to the target answer.

Figure 3 reveals clear differences in localization across the three methods. Attention-based attribution performs worst overall: its high-response regions are often dispersed across the background or other objects unrelated to the question. Thus, Transformer attention weights do not necessarily reflect contributions to visual evidence. Visual representation perturbation, by contrast, more accurately highlights question-relevant local regions. A large change in the output representation after perturbing a visual patch indicates that the corresponding region strongly influences the current prediction, giving this approach a stronger causal interpretation. In Figure 3, perturbation-based importance concentrates to some extent around the person's top, suggesting that LVLMs do rely on local visual evidence associated with the queried attribute when answering color questions. Gradient attribution performs best in this qualitative comparison, with salient regions focused most consistently on the person's top. This suggests that gradients more directly characterize the local sensitivity of the target-word probability to individual visual tokens. A large gradient norm indicates that a small change in the corresponding token has a greater effect on the target-word prediction, making that token more likely to provide visual evidence supporting the generated word.

Across Transformer layers, visual focus exhibits a clear progression. In early layers, salient regions are relatively dispersed: the model primarily retains low-level visual features and coarse spatial information, without a strong semantic focus on the specific question. In middle layers, salient regions increasingly concentrate around the person's top, suggesting that semantic constraints such as "person," "top," and "color" are being aligned with local image regions. Visual representations at this stage are more task-relevant. In later layers, the salient regions become somewhat more diffuse, consistent with a shift from integrating local visual evidence toward global semantic modeling and language generation decisions. Middle layers therefore tend to exhibit the strongest local visual localization, while deeper layers contribute more to cross-modal semantic integration and the final textual output.

These visualizations provide an intuitive basis for the layer selection study in Appendix C. Early layers lack sufficient semantic alignment to provide stable target-relevant visual evidence. Although very deep layers contain richer global semantic information, their attributions are more susceptible to language priors and the generation context, and no longer correspond consistently to local image regions. A suitable detection layer should therefore lie in the middle or moderately late portion of the model, where strong local visual localization coexists with meaningful cross-modal semantic alignment. The quantitative experiments in Appendix C corroborate this observation: middle and moderately late layers yield better hallucination mitigation than early and late layers, with layer 16 providing the most consistent overall balance among CHAIRs, CHAIRi, recall, and output length.

B Derivation Relating Visual Perturbation to Gradient Attribution

Let $f(\cdot)$ be a nonlinear scalar-valued function with input $x \in \mathbb{R}^d$ and output $y = f(x)$. In our attribution method, x is the input representation of a visual token at the selected Transformer layer, and $f(x)$ is the target-word probability read out from that layer, with all other layer-input representations held fixed. Perturbation analysis measures the change in y induced by a small change in x. Let the perturbation be $\varepsilon\Delta x$, where Δx specifies the perturbation vector and $\varepsilon > 0$ is a small scalar controlling its scale. The resulting output change is

$$\Delta y = f(x + \varepsilon\Delta x) - f(x). \tag{1}$$

By Taylor's theorem, assuming sufficient local smoothness, expanding $f(x + \varepsilon\Delta x)$ around x gives

$$\begin{aligned} f(x + \varepsilon\Delta x) = f(x) + \varepsilon\nabla f(x)^\top \Delta x \\ + \frac{\varepsilon^2}{2}\Delta x^\top H(x)\Delta x + \mathcal{O}(\varepsilon^3). \end{aligned}$$

Here, $\nabla f(x)$ is the gradient of the scalar output with respect to the input, and $H(x)$ is the Hessian matrix of second derivatives. For sufficiently small perturbations, neglecting higher-order terms yields the first-order approximation

$$f(x + \varepsilon\Delta x) \approx f(x) + \varepsilon\nabla f(x)^\top \Delta x.$$

Subtracting $f(x)$ from both sides and substituting into Equation (1) yields

$$\Delta y \approx \varepsilon\nabla f(x)^\top \Delta x.$$

Using the inner-product identity, the output change can be expressed as

$$\Delta y \approx \varepsilon\|\nabla f(x)\|_2\|\Delta x\|_2 \cos(\theta),$$

where θ is the angle between the gradient $\nabla f(x)$ and the perturbation vector Δx. Thus, an individual perturbation response depends on both the gradient magnitude and the perturbation's magnitude and direction.

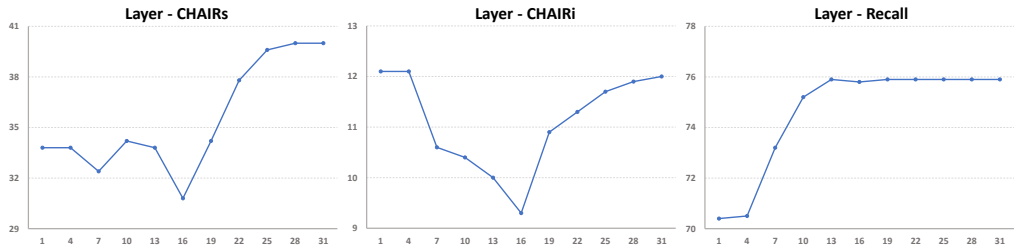


Figure 4: Effect of the gradient attribution layer on hallucination mitigation with HT-SCD.

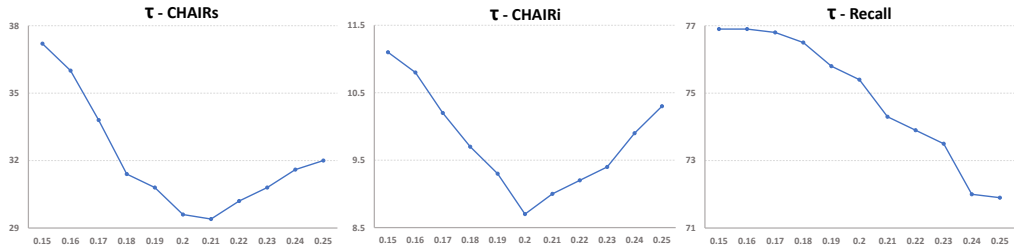


Figure 5: Effect of the detection threshold τ on hallucination mitigation with HT-SCD.

To characterize the response statistically, assume isotropic Gaussian perturbations, $\Delta x \sim \mathcal{N}(0, I_d)$. Then $\mathbb{E}[\Delta x] = 0$ and $\mathbb{E}[\Delta x \Delta x^\top] = I_d$. Although the signed first-order output change has zero mean, its mean squared magnitude satisfies

$$\begin{aligned}\mathbb{E}\big[(\Delta y)^2\big] &\approx \varepsilon^2 \nabla f(x)^\top \mathbb{E}[\Delta x \Delta x^\top] \nabla f(x) \\ &= \varepsilon^2 \|\nabla f(x)\|_2^2.\end{aligned}$$

Consequently,

$$\sqrt{\mathbb{E}[(\Delta y)^2]} \approx \varepsilon \|\nabla f(x)\|_2.$$

For a common sufficiently small perturbation scale, the gradient norm is therefore approximately proportional to the root mean squared change in the scalar output. Setting $x = v_i$ and $f(x) = P$ motivates the attribution score $I_i = \|\partial P / \partial v_i\|_2$ in Section 4.1.1, which measures the local sensitivity of the target-word probability to visual token i.

This derivation applies to perturbations and gradients of the same scalar target. It provides a local sensitivity interpretation of our probability-based gradient attribution.

C Hyperparameter Analysis

C.1 Selection of the Gradient Attribution Layer

To assess how the attribution layer affects hallucination mitigation, we compute visual token contribution scores for target object words at different Transformer layers and evaluate the resulting CHAIR scores. Figure 4 shows substantial differences across layers, with both early and late layers underperforming middle layers. Layer 16 yields the best overall performance, achieving CHAIRs of 30.8 and CHAIRi of 9.3 while maintaining recall at 75.8 and leaving output length largely unchanged.

These results corroborate the layerwise visualizations in Appendix A. Early layers primarily retain low-level visual features and lack sufficient semantic alignment. Middle layers offer strong local visual localization, while moderately late layers combine local visual evidence with global semantic context more consistently. Layer 16 lies at a point where visual–semantic alignment is well developed

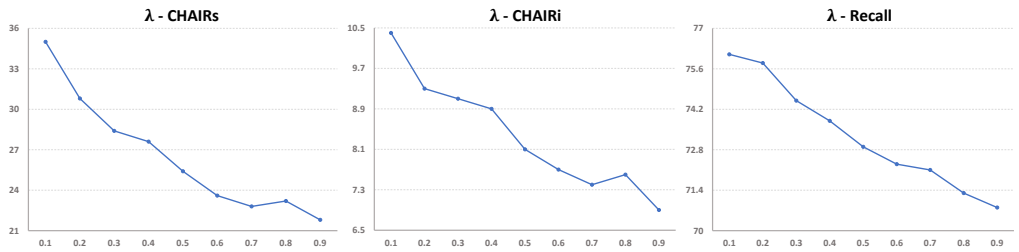


Figure 6: Effect of the hallucination penalty coefficient λ on hallucination mitigation with HT-SCD.

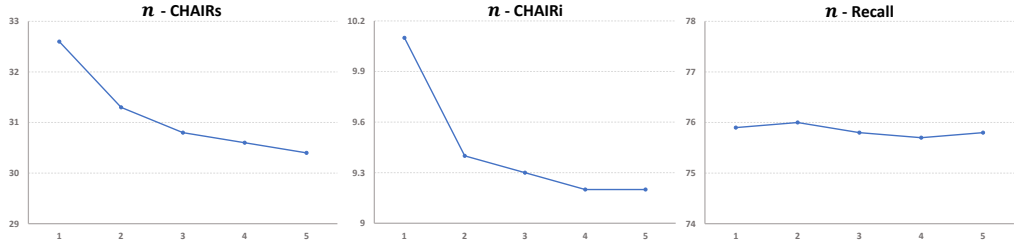


Figure 7: Effect of the number of regeneration attempts $\mathbf{n}$ on hallucination mitigation with HT-SCD.

but deep-layer language generation priors have not yet become dominant. We therefore use it as the default gradient attribution layer.

C.2 Hallucination Detection Threshold

As shown in Figure 5, increasing the detection threshold τ from 0.15 to 0.25 first decreases and then increases CHAIRs and CHAIRi, while recall generally declines. A low threshold is permissive: some unsupported object mentions fail to trigger correction, resulting in higher hallucination rates. Raising the threshold identifies and corrects more potential hallucinations, and both hallucination metrics reach their minima near $\tau = 0.20$, at approximately 29.6 and 8.7, respectively. Further increases in the threshold cause both metrics to rise while recall continues to decrease. This suggests that an excessively high threshold also triggers interventions on some correct object mentions, causing unnecessary rollback and regeneration, reducing information retention, and providing no additional mitigation benefit. We select $\tau = 0.19$ as the default because it offers a favorable balance between hallucination reduction and object coverage.

C.3 Hallucination Penalty Strength

As shown in Figure 6, increasing the penalty coefficient λ from 0.1 to 0.9 generally reduces CHAIRs and CHAIRi, from approximately 35.0 and 10.4 to 21.8 and 6.9, respectively. Stronger penalties therefore suppress object hallucinations effectively. However, recall also decreases from approximately 76.1 to 70.8, indicating that excessive penalties may suppress some visually grounded object descriptions and reduce object coverage. When λ increases from 0.1 to 0.2, CHAIRs and CHAIRi improve substantially while recall decreases only slightly, yielding a favorable trade-off between hallucination mitigation and information retention. Moderate penalties thus reduce hallucinations while preserving high object coverage, whereas larger penalties favor lower hallucination rates at the expense of descriptive completeness. We therefore set the default penalty coefficient to $\lambda = 0.2$.

C.4 Number of Regeneration Attempts

As shown in Figure 7, increasing the number of regeneration attempts n from 1 to 3 reduces CHAIRs from approximately 32.6 to 30.8 and CHAIRi from approximately 10.1 to 9.3. Multiple rounds of rollback and correction can therefore remove hallucinations that persist after a single attempt.

Increasing n further to 4 or 5 produces only small fluctuations in the two metrics, without consistent additional improvement, suggesting diminishing returns from further correction. Recall remains approximately 75.7–76.0 across settings, indicating that additional attempts do not materially reduce object coverage within the tested range. Given the extra regeneration cost of more attempts, we set the maximum number to n = 3.

D Localization Metrics

We quantify localization of the visual region corresponding to a target object using PG-Box, EIB, and Recall@K. These metrics respectively assess the location of the strongest response, the concentration of attribution energy, and target-region coverage. Let an image contain N visual tokens, let B denote the set of tokens covered by the target object's bounding box, and let a_i be the attribution score of visual token i.

(1) PG-Box. This metric checks whether the visual token with the highest attribution score belongs to B and averages the resulting binary hits across examples. A higher PG-Box indicates that the attribution peak falls within the target region more often.

(2) EIB. This metric measures the fraction of nonnegative attribution energy inside the target bounding box relative to the total energy of the attribution map:

$$\text{EIB} = \frac{\sum_{i \in B} a_i^+}{\sum_{i=1}^{N} a_i^+}, \qquad a_i^+ = \max(a_i, 0)$$

A higher EIB indicates that attribution is more concentrated in the target region. Because each method is normalized by its own total attribution energy, this metric is not directly affected by differences in the overall scale of attribution scores across methods.

(3) Recall@K. Let T_K be the set of K visual tokens with the highest attribution scores. Then

$$Recall@K = \frac{|T_K \cap B|}{|B|}$$

We primarily report Recall@5%, selecting 29 tokens when $N = 576$, corresponding to approximately 5% of the visual tokens. This metric measures how well highly attributed tokens cover the target region and differs from the object-category recall used in CHAIR.

PG-Box, EIB, and Recall@5% therefore characterize peak localization accuracy, attribution energy concentration within the target region, and target-region coverage, respectively. Higher values indicate better localization for all three metrics.

E Object Word Identification

After generating each sentence, HT-SCD applies tokenization and lemmatization, matches multiword object expressions first, and then identifies remaining object mentions using category names and synonyms. These mentions are checked by the detector in Section 4.1. The main experiments use the object vocabularies from the official CHAIR and AMBER-G evaluation code.

To assess dependence on benchmark-specific vocabularies, we replace both inference-time vocabularies with a shared LVIS vocabulary on LLaVA-1.5-7B, retaining each benchmark's official evaluation protocol. On MSCOCO-CHAIR, this yields CHAIRs/CHAIRi of 30.3/9.4, recall of 76.1, and output length of 85.6, compared with 29.4/9.2, 75.9, and 85.9 under the original setting. On AMBER-G, CHAIR/Cover/Hal/Cog are 4.31/49.40/20.32/2.78, compared with 4.10/49.20/18.80/2.40.

The shared LVIS vocabulary retains most of the hallucination reduction with comparable object coverage, suggesting that HT-SCD remains effective without exact alignment with benchmark-specific vocabularies. Nevertheless, detection remains vocabulary-dependent; open-vocabulary object identification is left for future work.

F FAILURE CASES IN HALLUCINATION MITIGATION

This appendix examines three representative cases: a hallucinated object misclassified as real, a real object misclassified as hallucinated, and three rounds of rollback and regeneration. We show only sentences or recorded excerpts relevant to detection and rollback, omitting unrelated correct descriptions.

F.1 HALLUCINATED OBJECT MISCLASSIFIED AS REAL

Case 1: A Hallucinated Object Misclassified as Real

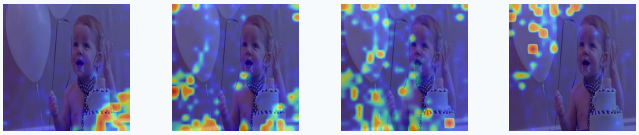


Input image

Green:Detected as Real
Red:Detected as Hallucinated

Sentence Identified as Hallucinated:
The **cake** is placed on a **table**, and there is a **knife**

Regeneration:
The scene also includes a few **balloons**, with one located near the top left corner of the image and another one towards the top right corner.



Cake Table Knife Balloon

Figure 8: Hallucinated object misclassified as real.

Error Analysis As shown in Figure 8, neither a table nor a knife is present in the image. The hallucinated object mention table is incorrectly classified as a real object. To investigate the cause of this error, we visualize the gradient attribution heatmap.The heatmap shows that the model primarily attends to the wooden wall in the background. This misclassification may arise because the wooden wall shares visual characteristics with a table. Moreover, the strong semantic compatibility between table and the overall scene may further cause the model to misclassify the hallucinated object as real.

Case2 : A Real Object Misclassified as Hallucinated

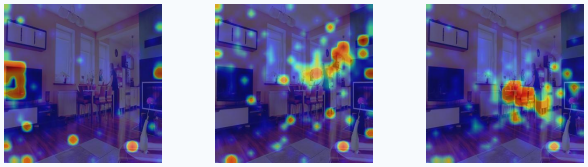


Input image

Green:Detected as Real
Red:Detected as Hallucinated

Sentence Identified as Hallucinated:
A **TV** is mounted on the wall in the background, while a **clock** hangs on the wall above the dining **table**.

Regeneration:
The **table** is accompanied by a **TV**, which is located on the left side of the room.



TV Clock Table

Figure 9: Real object misclassified as hallucinated.

F.2 Real Object Misclassified as Hallucinated

Error Analysis As shown in Figure 9, the image contains a clock, but the model incorrectly classifies it as a hallucinated object. To investigate the cause of this error, we visualize the gradient attribution heatmap. The heatmap shows that the model fails to correctly localize the clock because it occupies only a very small region in the scene, ultimately leading to an incorrect prediction.

F.3 Three Rollbacks

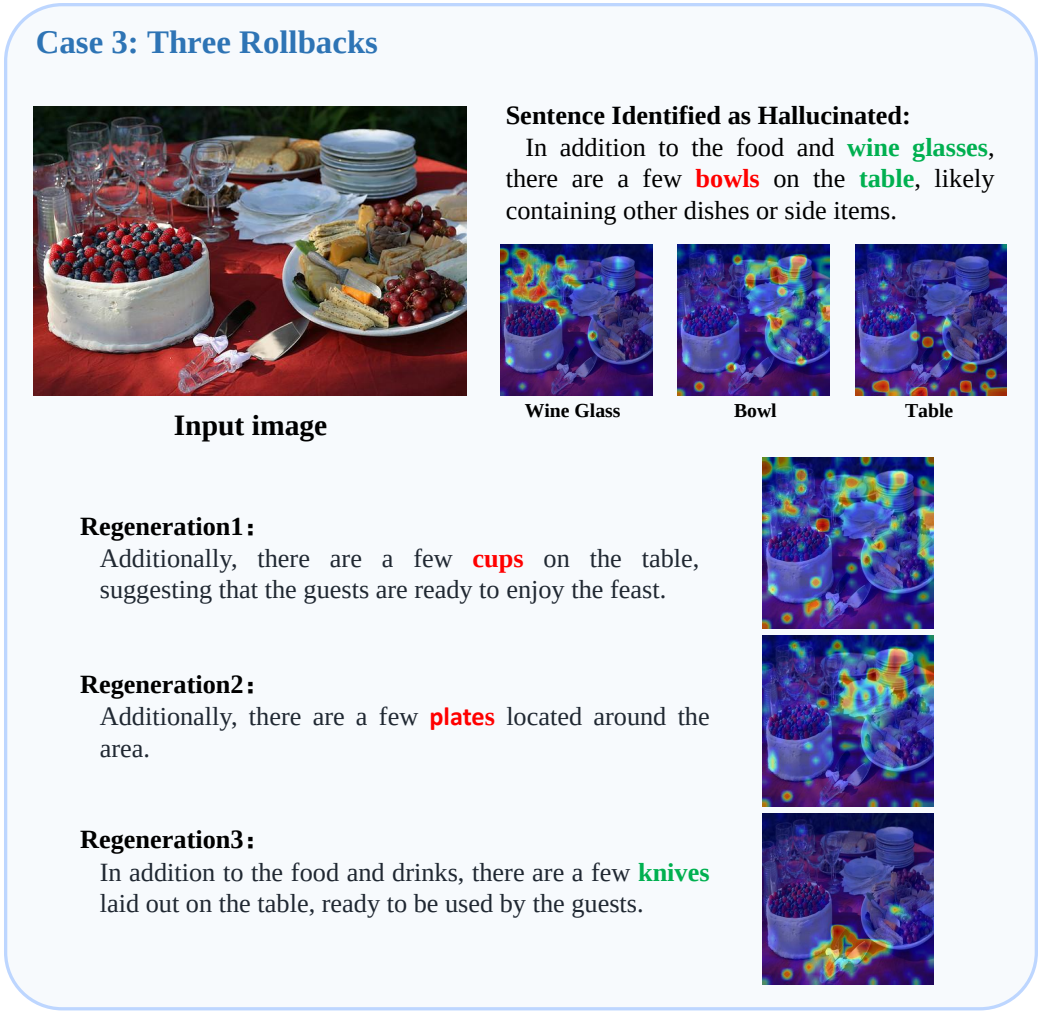


Figure 10: Three rollbacks.

Error Analysis As shown in Figure 10, the image depicts a dining table scene containing cups and plates but no bowls. However, these three object categories are visually similar and also frequently co-occur in such scenes. Across the three rollback and regeneration attempts, the model fails to reliably determine which of the three object categories are actually present. Consequently, after three regeneration attempts, none of the three object categories is selected in the final output.

These cases illustrate three limitations of detection and mitigation. First, similar background textures and scene-level semantic co-occurrence can assign high scores to mentions without actual object evidence, causing hallucinations to be missed. Second, weak visual features and changes in context can lower the scores of real objects, triggering unnecessary rollback and the loss of valid information. Third, for visually similar categories, a limited number of self-contrastive regeneration attempts may still repeatedly produce hallucinations. The online loop in HT-SCD therefore reduces object

hallucinations but does not guarantee that every detection or regeneration decision is correct. Future improvements could better distinguish background cues from object evidence, calibrate contextual effects and detection thresholds, and enable finer-grained category discrimination, while considering both hallucination rates and coverage of real objects.

G Experimental Protocol

Data sampling. For the MSCOCO hallucination mitigation experiments in Table 2, we randomly sample 500 images using seed 42 and use the same images for all three models and all baselines. The ablations in Table 4 use this same image set. The hallucination detection experiment in Table 1 uses 5,000 randomly sampled MSCOCO images.

Hyperparameter selection. The hyperparameter experiments in Appendix C use MSCOCO images sampled with seed 142. We manually select the default configuration based on these experiments and fix it for the main evaluation. All three models use attribution layer $l = 16$, contrastive strength $\lambda = 0.2$, and at most $n = 3$ regeneration attempts. The detection threshold is $\tau = 0.19$ for LLaVA-1.5-7B and InstructBLIP-Vicuna-7B, and $\tau = 0.58$ for Qwen2.5-VL-7B. For each model, these hyperparameters are transferred directly to AMBER-G without retuning. Object extraction follows the benchmark-specific procedures in Appendix E .

Generation and baseline settings. All compared methods use the same task prompt and generate at most 256 new tokens. Baselines use their official default settings and hyperparameters. In Table 4, we replace only the negative-branch construction or correction strategy while keeping the detector, detection threshold, and rollback settings unchanged. The Resample variant samples from the original predictive distribution during correction, constituting the intended exception to greedy decoding.

AI Use Statement

We used generative AI tools for language editing, code development, debugging, brainstorming, and feedback on experimental results. Generative AI was not used to draft the substantive content of the paper or to design the experiments.

All AI-assisted outputs were reviewed and verified by the authors. AI-assisted code was manually inspected and tested, AI-assisted edits and feedback were independently evaluated for technical accuracy and consistency with the experimental evidence. The authors take full responsibility for the final content of this work.