DON'T GRADE YOUR OWN EXAM: ITERATIVE EXTERNAL-DETECTOR VERIFICATION AND REWIND DECODING FOR LVLM HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (LVLMs) often generate fluent responses that are not fully grounded in visual evidence. Most existing hallucination-mitigation methods rely on the generator's own internal representations to revise its predictions, leaving the correction process exposed to the same biases that caused the original error. External vision models offer a complementary source of evidence that can verify generated claims independently of the generator. Building on this idea, we introduce **SIEVE**, a training-free decoding framework that performs selective, iterative verification during generation. Whenever the model produces a verifiable claim about the image, an external verifier evaluates that claim against the visual evidence. If the claim is rejected, SIEVE rolls generation back to the rejected word, suppresses it at the same decoding position, and resumes from the model's next-best continuation. The regenerated claim is verified again, allowing correction to continue until the output is supported by the evidence rather than simply deleting unsupported content. We instantiate SIEVE for object hallucination, where established benchmarks and reliable external verifiers are readily available, using an object detector as the verifier. Across LLaVA-1.5, InstructBLIP, and Qwen2.5-VL, SIEVE reduces object hallucinations by 42-80% while preserving caption length and fluency.

1 INTRODUCTION

Large vision-language models (LVLMs) describe scenes and answer questions with remarkable fluency, yet the text they produce is not always grounded in the image. Models assert objects that are absent, attributes the image contradicts, and relations that never hold (Rohrbach et al., 2018; Liu et al., 2024a). These hallucinations are hard to catch from the outside because neither fluency nor model confidence separates a hallucinated claim from a grounded one (Li et al., 2023).

The dominant approach has been to have the model improve itself using its own internal information. For example, contrastive decoding perturbs the input or the layers and reads off the difference (Leng et al., 2024; Wang et al., 2025). Attention- and representation-level methods suppress patterns the generator associates with hallucination (Huang et al., 2024; Liu et al., 2024c; 2025). Retraining-based revision learns its corrections from the very network that made the errors (Wu et al., 2025). Structurally, this asks the student who took the exam to grade their own answer sheet: the LVLM and its internal critic share visual representations and linguistic biases, so an error of perception is graded with the same faulty perception.

The natural alternative to self-reference is to verify the model's claims with an external tool. A hallucinated claim is, at bottom, a factual statement about the image, and a specialized vision model can check it. Because that checker is trained separately, it does not share the generator's visual representations or linguistic priors, so it does not inherit the bias that produced the error. And unlike the LVLM, which must solve open-ended generation, the checker faces a single, well-defined recognition task, a narrow scope that is comparatively easy to optimize to high reliability.

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.

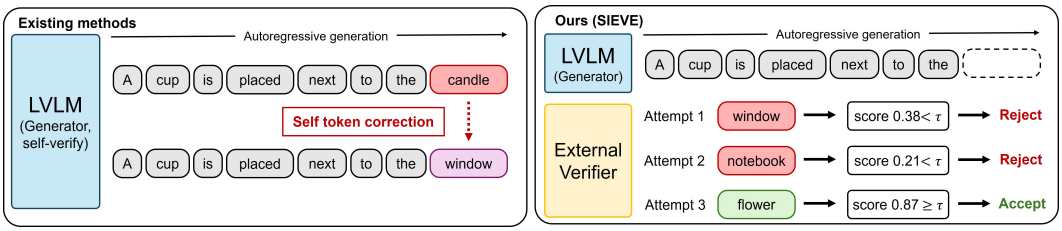


Figure 1: Existing approaches versus SIEVE. (a) A single correction driven by the model's own internal signals; the generator commits to an unsupported object and cannot take it back. (b) Iterative correction against an external reference model: each object claim is verified at completion and, on rejection, generation rewinds to the word boundary with a position-local ban and regenerates until the claim is supported.

A few methods have incorporated external vision models into the decoding process (Zhao et al., 2025; Lee & Song, 2025; Chen et al., 2026), yet their outputs still contain many hallucinated claims. These methods rely on a static, one-shot integration of the external signal: they merge it with the generator's own prediction into a single adjusted distribution and decode only once. As a result, the linguistic bias that produced the hallucination can still dominate the resulting token selection. What is needed instead is dynamic verification that engages only at tokens that assert visual content and repeatedly revises a rejected claim until it is supported by visual evidence.

To this end, we propose **SIEVE** (**S**elective **I**terative **E**xternal **V**erification and **E**diting), a decoding framework that actively suppresses hallucinated visual claims during generation. Whenever the model produces a visually grounded claim that can be verified against the image, an external verifier evaluates its consistency with the visual evidence. If the claim is deemed unsupported, SIEVE (i) rewinds generation to the beginning of the rejected word, (ii) suppresses that word at the same decoding position, and (iii) resumes generation from the model's next-best alternative. Every regenerated claim is then verified again, and this verification-and-revision cycle continues until the generated claim satisfies the visual evidence criterion. In contrast, existing revision methods perform only a single correction pass, as illustrated in Fig. 1.

Hallucination takes many forms, including objects, attributes, relations, and actions, and a verified decoder is only as trustworthy as the verifier behind it. We therefore validate our approach on the form that can be held to the highest standard today: object-existence hallucination, in which absent objects are mentioned as present. It is the most frequent and most intensively studied form, and the standard generative and discriminative benchmarks are built around it (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2023; Liu et al., 2024a). The framework itself is agnostic to the type of claim, and stronger verifiers for attributes, relations, and actions extend it directly.

SIEVE is training-free, and the external evidence is what makes it effective. Across three representative LVLMs, SIEVE removes 42–80% of object hallucinations without shortening captions or degrading fluency and outperforms every baseline reproduced under our unified protocol, while incurring only a modest increase in decoding cost.

2 Related Work

LVLMs perform strongly on captioning, VQA, and multimodal reasoning (Liu et al., 2024b; Dai et al., 2023; Bai et al., 2025), yet remain prone to fluent claims about objects, attributes, actions, or relations the image does not support, errors attributed to linguistic priors, under-use of visual evidence, and biased training data (Rohrbach et al., 2018; Li et al., 2023; Liu et al., 2024a); reducing object hallucination does not automatically mitigate action- or state-level hallucination (Wang et al., 2026). As in the introduction, we restrict attention to object-existence hallucination.

Existing mitigations divide into training-based, post-hoc, and decoding-control approaches. Training-based methods improve faithfulness through fine-tuning or hallucination-targeted preference optimization (Fu et al., 2025) at the cost of additional data and parameter updates. Post-hoc methods revise a completed response with external tools and therefore require a separate analysis-

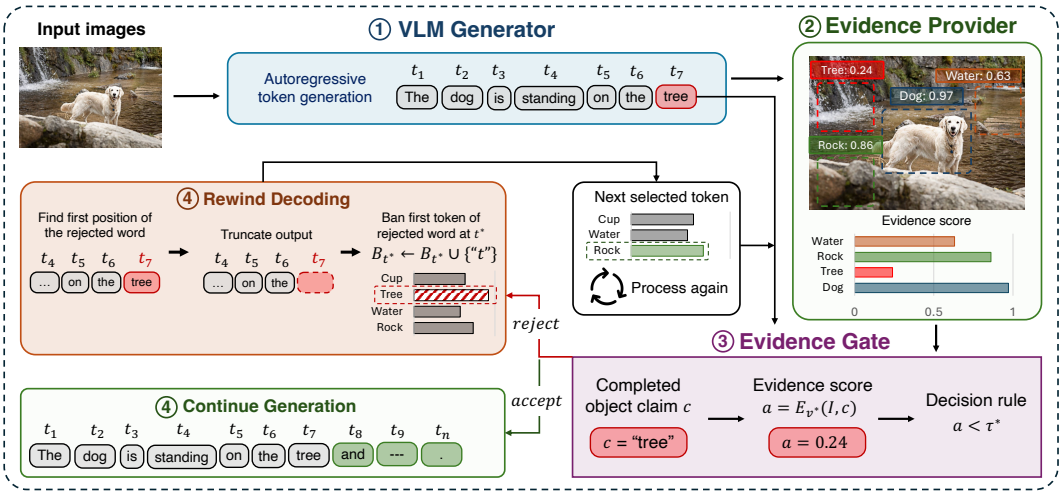


Figure 2: SIEVE overview. A completed object claim is scored by the evidence provider. A claim with a score below the threshold is rejected, after which decoding rewinds to the first token of the rejected word, bans that token at that position, and regenerates from the accepted prefix. Verification repeats until a supported claim is produced.

and-revision stage: Woodpecker verifies key concepts and rewrites unsupported content (Yin et al., 2024), while VBackChecker and InEx check consistency through grounding or iterative cross-modal verification (Guo et al., 2026; Yang et al., 2026). Decoding-control methods intervene during generation without retraining: VCD, OPERA, HALC, and SID reshape token selection with contrastive distributions, attention patterns, or internally selected visual information (Leng et al., 2024; Huang et al., 2024; Chen et al., 2024; Huo et al., 2025), and more recent methods manipulate attention or hidden representations directly (Ouyang et al., 2026; Tang et al., 2026; Yu et al., 2026; Li et al., 2026a; Cheng et al., 2026; Li et al., 2026b). Their corrective signals, however, are still derived primarily from the generator's own internals.

2.1 Decoding with External Visual Evidence

A distinct line brings in visual information from outside the generator. MARINE runs an external object-grounding model on the image and injects the extracted object features as additional guidance for the generator's token predictions (Zhao et al., 2025). RVCD prepares reference images that do and do not contain the candidate objects and contrasts the token distributions the two induce (Lee & Song, 2025). VCGD first lets an auxiliary captioner describe the image and then uses the resulting clues to adjust the token distribution (Chen et al., 2026). In all three, the evidence from the external vision model is blended with the LVLM's own distribution and the output is decoded once from the mixture. SIEVE keeps the two sources separate and instead re-decodes: a token is accepted only once it satisfies the evidence criterion, and generation is rewound and decoded again until it does. A single mixed pass is still decoded under the generator's linguistic bias and can reintroduce the very hallucination the external evidence was meant to remove.

3 Method

3.1 Problem Setup and Overview

Given an image I and a prompt p, a frozen vision-language model G autoregressively generates a response that may contain object mentions unsupported by the image. SIEVE reduces such mentions without updating the parameters of G or of any model it consults.

SIEVE augments autoregressive decoding with an evidence-gated verification loop (Figure 2). Decoding proceeds as usual until the tokens realizing an object mention are complete. The mention is then scored by an evidence provider and either accepted, leaving the emitted text and the retained

prefix untouched, or rejected, in which case generation rewinds to the beginning of the mention, the rejected word is blocked at that position, and decoding resumes from the retained prefix. The replacement is verified in the same way once it completes, so the loop repeats until a mention is accepted.

3.2 Evidence-Gated Rewind Decoding

An evidence provider E maps an image and a candidate object category to a scalar score,

$$a = E(I, c), \tag{1}$$

where c is the object category and higher a indicates stronger visual support for c in I. All parameters remain fixed throughout provider selection, calibration, and evaluation.

A model generator G claims that a category c is present in the image. During decoding, the model's claim is represented as a tuple (c, m, t^*), where m is the object mention and t^* is the first subword position of that mention. Once a claim is completed, the provider returns an evidence score $a = E(I, c)$, and the deployed policy accepts or rejects the claim by thresholding a.

An accepted claim leaves the output and key-value cache unchanged. A rejected claim triggers the minimal reversible edit: the first token of m is added to a position-specific ban set B_{t^*}, the output and cache are truncated to the prefix preceding t^*, and decoding resumes from the retained prefix.

3.3 Evidence-Provider Selection

Standalone recognition accuracy is not sufficient for selecting an evidence provider E. SIEVE benefits only when the evidence provider E rejects unsupported claims produced by the target generator, whereas rejecting a supported claim may remove valid content.

For an evidence provider E and threshold τ, *Rescue* is the fraction of hallucinated claims that the provider correctly rejects, whereas *Damage* is the fraction of supported claims that it incorrectly rejects. Thus, Rescue measures how many hallucinations are caught, while Damage measures how much correct content is unnecessarily blocked. Because regeneration may recover a rejected object using another supported expression, Damage conservatively estimates the potential content loss.

For each evidence provider E, we evaluate a range of evidence thresholds τ. We first retain only settings whose Damage does not exceed the allowed tolerance ϵ:

$$\mathrm{Damage}(E, \tau) \leq \epsilon. \tag{2}$$

Among these retained settings, we select the provider–threshold pair that achieves the highest Rescue:

$$(E^*, \tau^*) = \arg\max_{E,\tau} \mathrm{Rescue}(E, \tau), \tag{3}$$

where the maximization considers only settings satisfying Eq. 2. The threshold τ^* is used only to compare providers.

3.4 Risk-Controlled Policy Selection

The evidence threshold τ controls how aggressively SIEVE rejects object claims. A claim is accepted when $E^*(I, c) \geq \tau$ and rejected otherwise. Rather than selecting τ through manual hyperparameter tuning, we calibrate it on a held-out subset of the benchmark's training split following Learn-then-Test (LTT) (Angelopoulos et al., 2025). No model parameters are updated during calibration.

We define a fixed set of candidate thresholds and run SIEVE with each threshold on the entire calibration set. A caption is counted as a failure if it contains at least one object claim unsupported by the image annotation.

Let $R(\tau)$ denote the true probability that a final caption generated with threshold τ contains at least one unsupported object claim. Following LTT, we select a threshold $\widehat{\tau}$ satisfying

$$\Pr\left(R(\widehat{\tau}) \leq \alpha\right) \geq 1 - \delta, \tag{4}$$

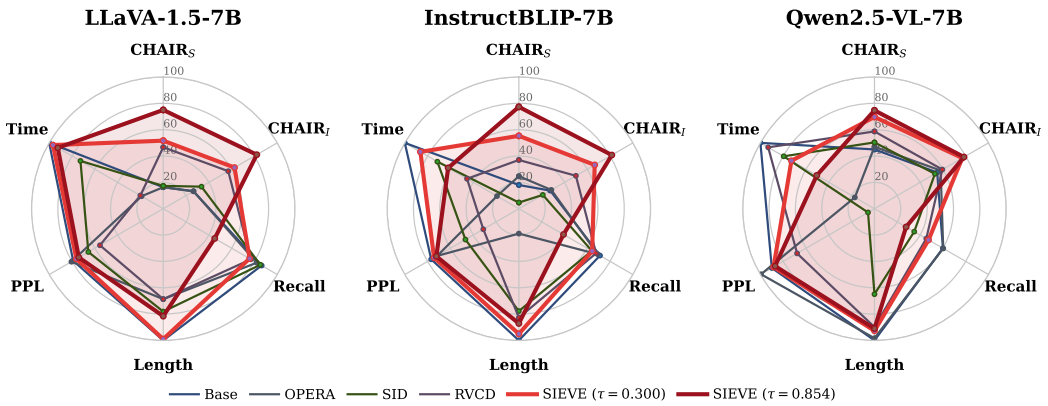


Figure 3: Comparison on COCO-500 for LLaVA-1.5-7B, InstructBLIP-7B and QWEN2.5-VL-7B. Larger radii indicate better performance, with CHAIR$_S$, CHAIR$_I$, PPL, and decoding time plotted on reversed scales.

where α is the maximum acceptable caption-level hallucination risk and δ is the allowed probability that the guarantee fails. The outer probability is taken over the random calibration sample. Among the certified thresholds, SIEVE selects the one that preserves the most supported object mentions.

4 Experiments

4.1 Experimental Setup

Models. We evaluate SIEVE on three representative LVLMs: LLaVA-1.5-7B (Liu et al., 2024b), InstructBLIP-7B (Dai et al., 2023), and Qwen2.5-VL-7B (Bai et al., 2025).

Datasets. We evaluate SIEVE on both generative hallucination and open-ended response quality. For generative hallucination, *COCO-500* contains 500 held-out COCO images following the CHAIR evaluation protocol (Rohrbach et al., 2018). *COCO-Dense* contains 500 object-dense COCO images drawn from the POPE-adversarial image pool (Li et al., 2023); it contains 3.88 annotated object categories per image on average, compared with 2.74 in COCO-500. We further evaluate open-vocabulary hallucination on *AMBER* (Wang et al., 2023) extends the evaluation to object, attribute, and relation hallucinations.

Baselines. We compare SIEVE with ten hallucination-mitigation baselines. The comparison includes contrastive-decoding methods VCD (Leng et al., 2024), DoLa (Chuang et al., 2024), SID (Huo et al., 2025), and RVCD (Lee & Song, 2025); attention- and logit-intervention methods OPERA (Huang et al., 2024), PAI (Liu et al., 2024c), OWL (Yu et al., 2026), AGLA (An et al., 2025), and AIR (Sun et al., 2026); and the self-verification method REVERSE (Wu et al., 2025).

Evaluation metrics. CHAIR$_S$ and CHAIR$_I$ measure hallucinations at the caption and object-mention levels, respectively (Rohrbach et al., 2018). Recall measures the preservation of ground-truth objects; lower CHAIR and higher Recall indicate better performance. We additionally report output length, perplexity, and relative decoding time.

Implementation details. The evidence provider selected by the audit is Faster R-CNN, chosen under a damage cap of ϵ=0.18. All experiments use the prompt "Please describe this image in detail." and are conducted on an NVIDIA RTX A5000 GPU with 24 GB of memory.

4.2 Main Results

Table 1 reports the complete COCO-500 results across three generators, and Figure 3 summarizes the overall trade-offs among hallucination, content preservation, fluency, and decoding efficiency on

Table 1: Free-form object hallucination results on COCO-500 across three LVLMs. Bold and underlined values indicate the best and second-best.

	Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Length	PPL ↓	Time ↓
LLaVA-1.5-7B	Base	50.2	14.7	80.2	100.5	4.26	1.00×
	VCD	59.0	17.1	**83.1**	101.6	5.40	1.97×
	DoLa	58.8	18.0	73.6	79.0	10.73	1.20×
	AIR	54.6	15.5	<u>80.6</u>	104.5	<u>4.23</u>	1.20×
	OPERA	50.2	14.7	78.7	94.2	**4.17**	6.46×
	SID	49.6	13.3	79.8	96.1	5.06	1.88×
	OWL	38.6	12.3	74.7	83.5	7.53	9.70×
	REVERSE	36.4	10.8	70.0	90.7	5.13	10.00×
	RVCD	32.0	8.6	76.9	94.2	5.67	6.30×
	PAI	25.0	7.1	65.9	78.0	5.37	1.98×
	AGLA	27.2	9.4	76.2	82.7	4.98	2.58×
	SIEVE (τ=0.300)	29.0	7.5	76.5	100.7	4.43	**1.08**×
	SIEVE (τ=0.677)	19.4	4.7	71.6	102.2	4.48	<u>1.10×</u>
	SIEVE (τ=0.854)	<u>15.0</u>	<u>3.6</u>	65.8	104.2	4.56	1.19×
	SIEVE (τ=0.920)	**12.6**	**2.9**	61.0	105.3	4.65	1.32×
InstructBLIP-7B	Base	49.2	14.5	74.9	101.1	4.39	1.00×
	SID	57.2	15.8	72.4	96.6	6.19	1.93×
	OPERA	45.2	14.3	**74.2**	84.7	4.70	6.43×
	RVCD	37.8	10.0	<u>73.7</u>	97.6	7.13	3.51×
	SIEVE (τ=0.300)	26.8	6.7	72.5	102.1	**4.54**	**1.37**×
	SIEVE (τ=0.677)	17.0	4.3	68.7	103.0	<u>4.59</u>	<u>1.55×</u>
	SIEVE (τ=0.854)	<u>13.6</u>	<u>3.7</u>	63.7	103.7	4.66	2.40×
	SIEVE (τ=0.920)	**9.8**	**2.6**	59.0	102.8	4.76	3.34×
Qwen2.5-VL-7B	Base	33.0	9.0	71.0	178.0	3.60	1.00×
	OPERA	31.8	8.4	**71.1**	178.6	**3.03**	6.73×
	SID	29.8	9.4	62.2	165.5	8.66	<u>1.60×</u>
	RVCD	24.8	8.1	65.9	174.4	4.93	**1.17**×
	SIEVE (τ=0.300)	18.2	4.6	<u>66.6</u>	180.6	<u>3.69</u>	1.87×
	SIEVE (τ=0.677)	15.4	**4.1**	64.1	181.3	3.71	2.12×
	SIEVE (τ=0.854)	<u>15.2</u>	<u>4.3</u>	59.7	181.2	3.77	3.12×
	SIEVE (τ=0.920)	**15.0**	**4.1**	56.4	181.3	3.81	3.44×

COCO-500. Together, the results show that SIEVE does not correspond to a single aggressively filtered decoding point, but instead provides a family of operating points that can be selected according to the desired hallucination–content trade-off.

(1) Best trade-off against all ten baselines. For every baseline, SIEVE provides an operating point with lower hallucination at comparable or higher Recall. On LLaVA-1.5, where all ten baselines are evaluated, SIEVE provides multiple calibrated operating points for different hallucination-risk budgets. Increasing τ reduces $CHAIR_S$ from 29.0 to 12.6 and $CHAIR_I$ from 7.5 to 2.9.

(2) Large gains at matched Recall. SIEVE improves upon the baselines at comparable levels of Recall. At a Recall of approximately 77, it reduces $CHAIR_S$ from 32.0 to 29.0 relative to RVCD. REVERSE's Recall level, it reduces $CHAIR_S$ from 36.4 to 19.4 and $CHAIR_I$ from 10.8 to 4.7.

(3) Consistent reductions across generators. Every SIEVE operating point reduces both $CHAIR_S$ and $CHAIR_I$ relative to Base on LLaVA-1.5, InstructBLIP, and Qwen2.5-VL. In contrast, several internal-signal baselines increase one or both CHAIR metrics.

(4) The gains do not come from shorter outputs. Across all three generators, SIEVE produces captions that are as long as, or slightly longer than, those of Base, while perplexity remains close to that of the original model.

Table 2: Object hallucination on COCO-Dense. Bold and underlined values indicate the best and second-best CHAIR results.

Model	Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Length
LLaVA-1.5-7B	Base	60.6	14.4	76.4	108.5
	PAI	25.6	7.0	59.7	80.0
	SIEVE (τ=0.677)	<u>23.2</u>	<u>4.1</u>	69.0	111.0
	SIEVE (τ=0.920)	**12.0**	**2.3**	55.3	113.6
InstructBLIP-7B	Base	51.8	12.8	68.2	103.0
	PAI	65.0	14.4	69.0	130.4
	SIEVE (τ=0.677)	<u>16.8</u>	<u>4.0</u>	60.6	104.7
	SIEVE (τ=0.920)	**8.6**	**1.8**	50.2	108.9
Qwen2.5-VL-7B	Base	41.6	9.3	64.8	195.0
	PAI	35.4	8.1	59.2	165.9
	SIEVE (τ=0.677)	<u>21.6</u>	**4.8**	56.8	198.4
	SIEVE (τ=0.920)	**20.2**	<u>5.0</u>	49.4	199.2

Table 3: Open-ended hallucination and response results. Hal., Acc., and Help. denote hallucination rate, accuracy, and helpfulness, respectively.

	MMHal		LLaVA	
Method	Rating ↑	Hal. ↓	Acc. ↑	Help. ↑
Base	3.39	0.375	7.48	6.98
OPERA	<u>3.41</u>	0.366	7.51	6.99
SID	3.40	<u>0.351</u>	<u>7.58</u>	<u>7.08</u>
RVCD	3.35	0.365	7.22	6.88
SIEVE	**3.49**	**0.344**	**7.62**	**7.13**

Table 4: Evidence-provider comparison on COCO. Res., Dam., and C_S denote rescue rate, damage rate, and $CHAIR_S$, respectively.

Provider	AUC ↑	Res. ↑	Dam. ↓	$C_S \downarrow$
CLIP sim.	0.683	0.41	0.31	62.0
FRCNN	0.817	<u>0.50</u>	**0.17**	**40.8**
YOLOv8n	**0.828**	0.38	0.21	54.0
Y-World	<u>0.821</u>	0.43	0.21	<u>44.4</u>
G-DINO	0.664	0.32	0.31	51.2
OWLv2	0.802	**0.51**	<u>0.20</u>	50.0

4.3 Dense Scenes and Open Vocabulary

Generalization to object-dense scenes. We evaluate SIEVE on COCO-Dense to test whether its gains persist in scenes containing more annotated object. As shown in Table 2, SIEVE reduces both CHAIR metrics across all three generators while maintaining comparable caption lengths.

4.4 Preservation of Response Quality

As shown in Table 3, SIEVE slightly improves all judge scores on MMHal-Bench and LLaVA-Bench, indicating that it reduces hallucination without degrading the accuracy, helpfulness, or fluency of open-ended responses.

4.5 Evidence-Provider Audit

A strong standalone verifier is not necessarily an effective evidence provider for SIEVE. Table 4 compares candidate providers using AUC, Rescue, and Damage. YOLOv8n achieves the highest AUC but provides relatively low Rescue and yields a higher $CHAIR_S$. OWLv2 achieves the highest Rescue, whereas Faster R-CNN attains the lowest Damage among external providers while maintaining the second-highest Rescue and yielding the lowest $CHAIR_S$. SIEVE therefore selects Faster R-CNN as its evidence provider.

4.6 Case Study

Figure 4 compares captions generated from the same image. In this example, the base model and all baselines mention objects that are absent from the image, such as a mouse, cell phone, remote control, or chair. SIEVE rejects these unsupported mentions and regenerates a coherent caption using visually supported content. This shows SIEVE corrects hallucinations through replacement rather than simple deletion.

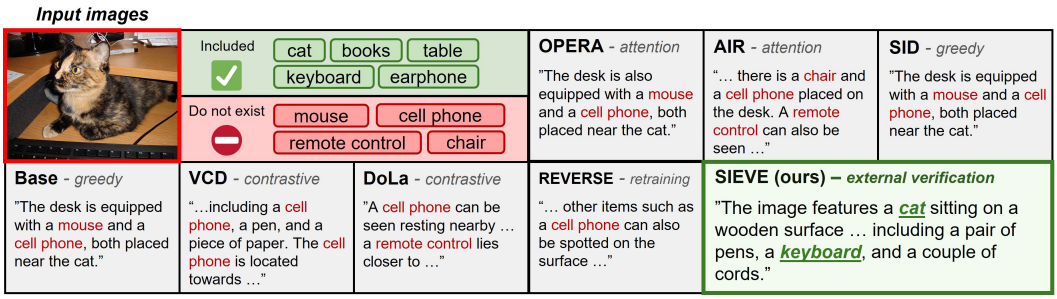


Figure 4: Qualitative comparison on a COCO image using LLaVA-1.5. Red and green indicate unsupported and detector-verified object mentions, respectively.

Table 5: Claim-rejection criteria.

	CHAIR ↓	
Revision criterion	S	I
No intervention	50.2	14.7
Random selection	49.0	15.0
Low model confidence	48.2	14.6
External evidence (Ours)	**29.0**	**7.5**

Table 6: Post-rejection actions.

	CHAIR ↓		
Action	S	I	Rew./img. ↓
Rewind only	50.2	14.7	6.4
Resample	39.4	11.3	5.2
Block word (Ours)	29.0	7.5	5.0
Block synonyms	**27.2**	**7.0**	32.0

5 Ablation

Which claims should be rejected? Randomly selecting claims does not reduce hallucination, while selecting claims with low model confidence provides a small improvement. Table 5 shows that rejecting claims unsupported by external evidence substantially reduces both CHAIR metrics while maintaining similar recall. This shows that model confidence alone does not reliably identify hallucinated claims.

How should decoding continue? Table 6 compares post-rejection actions. Rewinding alone reproduces the rejected word under greedy decoding, while resampling provides only a limited reduction. Position-local blocking forces a different continuation and substantially reduces both CHAIR metrics. Blocking all synonyms yields only a marginal additional reduction, from 29.0 to 27.2 in CHAIR_S and from 7.5 to 7.0 in CHAIR_I, while increasing rewinds from 5.0 to 32.0 per image. SIEVE therefore blocks only the rejected word and verifies each replacement when it completes.

5.1 Rewind Frequency and Decoding Cost

Figure 5 shows the cumulative distribution of rewinds per image across evidence thresholds. Increasing τ shifts the distribution to the right, with the median number of rewinds increasing from 1 at τ=0.300 to 7 at τ=0.920. Most images are completed with relatively few rewinds, while substantially more rewinds are required for only a small fraction. This pattern illustrates the trade-off between hallucination reduction and decoding cost.

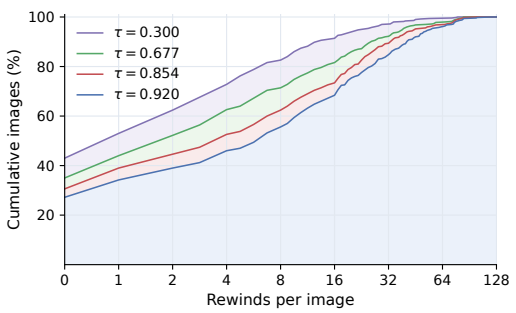


Figure 5: Cumulative distribution of rewinds per image across evidence thresholds.

Table 7: Selected thresholds and observed caption-level risk.

				Observed risk ↓	
Model	Domain	Target	τ	Cal.	Eval.
LLaVA	COCO	0.10	0.920	0.072	0.126
LLaVA	COCO	0.15	0.854	0.099	0.150
LLaVA	COCO	0.20	0.854	0.099	0.150
LLaVA	Dense	0.30	0.677	0.232	0.289
Qwen	COCO	0.15	0.677	0.093	0.062
Qwen	COCO	0.20	0.401	0.148	0.106
LLaVA	AMBER	0.25	0.200	0.207	0.233

5.2 Certified Operating Points

SIEVE selects its threshold from a user-specified caption-level risk target. For each generator and domain, Learn-then-Test evaluates a fixed set of candidate thresholds on a separate calibration set and returns a threshold that satisfies the target. Table 7 reports the requested risk target, the selected threshold, and the resulting risks on the calibration and evaluation sets. Stricter targets generally lead to higher thresholds.

6 Discussion

Scope and verifier limitations. SIEVE reduces object existence hallucination, but it does not eliminate it. Its performance is limited by the external verifier. In our audit, Faster R-CNN rejects 50% of hallucinated claims while incorrectly rejecting 17% of supported claims. Hallucinations missed by the verifier can remain, and incorrect rejections can remove valid content. Extending SIEVE to attributes, relations, or actions therefore requires reliable claim extractors and verifiers for those claim types.

7 Conclusion

In this paper, we introduced SIEVE, a training-free decoding framework for object hallucination. An evidence provider checks each object mention as it is generated. When a mention is unsupported, SIEVE rewinds to the word boundary, blocks the rejected word at that position, and regenerates text.

Our experiments show three findings. First, evidence providers should be selected by Rescue under a Damage constraint, rather than by standalone AUC. Second, position-local blocking prevents rejected claims from reappearing, while word-boundary rewind preserves already accepted text. Third, Learn-then-Test selects a decoding threshold from a user-specified hallucination target.

Across LLaVA-1.5, InstructBLIP, and Qwen2.5-VL, SIEVE reduces object hallucination by 42–80% while preserving caption length and response quality. These results show that iterative external verification is an effective training-free approach to object hallucination. Future work will broaden SIEVE beyond object existence to verify attributes, relations, and actions.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926, June 2025. URL https://arxiv.org/abs/2406.12718.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Guoqing Chen, Fu Zhang, Bingqian Liu, Chenglong Lu, and Jingwei Cheng. VCGD: Visual clue guided decoding with caption model for mitigating hallucination in multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(24):20041–20049, 2026. doi: 10.1609/aaai.v40i24.39089. URL https://ojs.aaai.org/index.php/AAAI/article/view/39089.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 2024. URL https://arxiv.org/abs/2403.00425.

Junyu Cheng, Zhibiao Liang, Yidong Chen, and Shuangyin Li. RFI: Rectified flow intervention for mitigating object hallucination in large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(5):3255–3263, 2026. doi: 10.1609/aaai.v40i5.37320. URL https://ojs.aaai.org/index.php/AAAI/article/view/37320.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.06500.

Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16563–16577, 2025. doi: 10.18653/v1/2025.findings-acl.850. URL https://aclanthology.org/2025.findings-acl.850/.

Pinxue Guo, Chongruo Wu, Xinyu Zhou, Lingyi Hong, Zhaoyu Chen, Jinglun Li, Kaixun Jiang, Sen-Ching Samson Cheung, Wei Zhang, and Wenqiang Zhang. Seeing is believing: Rich-context hallucination detection for MLLMs via backward visual grounding. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37):30871–30879, 2026. doi: 10.1609/aaai.v40i37.40345. URL https://ojs.aaai.org/index.php/AAAI/article/view/40345.

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*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

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*, 2025. URL https://arxiv.org/abs/2408.02032.

Jihoon Lee and Min Song. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8200–8219, 2025. doi: 10.18653/v1/2025.findings-acl.430. URL https://aclanthology.org/2025.findings-acl.430/.

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*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(37): 31645–31653, 2026a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhaoxu Li, Chenqi Kong, Yi Yu, Qiangqiang Wu, Xinghao Jiang, Ngai-Man Cheung, Bihan Wen, Alex Kot, and Xudong Jiang. SAVER: Mitigating hallucinations in large vision-language models via style-aware visual early revision. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):35617–35625, 2026b. doi: 10.1609/aaai.v40i42.40873. URL https://ojs.aaai.org/index.php/AAAI/article/view/40873.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models, 2024a. URL https://arxiv.org/abs/2402.00253.

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*, pp. 26296–26306, 2024b. doi: 10.1109/CVPR52733.2024.02484. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Sheng Liu, Haotian Ye, and James Y. Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=LBl7Hez0fF.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to images: A training-free method for alleviating hallucination in LVLMs. In *Computer Vision – ECCV 2024*, pp. 125–140. Springer Nature Switzerland, 2024c. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/10933_ECCV_2024_paper.php.

Shuyi Ouyang, Hongyi Wang, Gongfan Fang, Xinyin Ma, Lanfen Lin, and Xinchao Wang. Taming the phantom: Token-asymmetric filtering for hallucination mitigation in large vision-language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(10):8206–8214, 2026. doi: 10.1609/aaai.v40i10.37768. URL https://ojs.aaai.org/index.php/AAAI/article/view/37768.

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, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in LVLMs via attention imbalance rectification, 2026. URL https://arxiv.org/abs/2603.24058.

Lexiang Tang, Xianwei Zhuang, Bang Yang, Zhiyuan Hu, Hongxiang Li, Lu Ma, Jinghan Ru, and Yuexian Zou. Not all tokens and heads are equally important: Dual-level attention intervention for hallucination mitigation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40 (11):9439–9447, 2026.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Zehao Wang, Xinpeng Liu, Yudonglin Zhang, Xiaoqian Wu, Zhou Fang, Yifan Fang, Junfu Pu, Cewu Lu, and Yong-Lu Li. Verb Mirage: Unveiling and assessing verb concept hallucinations in multimodal large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12):10349–10357, 2026. doi: 10.1609/aaai.v40i12.38005. URL https://ojs.aaai.org/index.php/AAAI/article/view/38005.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5eff08bd064f0cdd92182cdf6fd06b99-Abstract-Conference.html.

Zhongyu Yang, Yingfang Yuan, Xuanming Jiang, Baoyi An, and Wei Pang. InEx: Hallucination mitigation via introspection and cross-modal multi-agent collaboration. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(35):29829–29837, 2026. doi: 10.1609/aaai.v40i35.40229. URL https://ojs.aaai.org/index.php/AAAI/article/view/40229.

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. doi: 10.1007/s11432-024-4251-x. URL https://doi.org/10.1007/s11432-024-4251-x.

Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in LVLMs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(42):36021–36029, 2026.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77461–77486, 2025. URL https://proceedings.mlr.press/v267/zhao25j.html.