Does This Object Exist? Object-Presence Verification for Hallucination Mitigation in LVLMs

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) often hallucinate objects absent from the input image, undermining their reliability in free-form captioning. We ask whether this failure leaves a detectable trace in the model: before an object mention is completed, does the hidden state already encode whether the object is visually grounded? Through layer-wise probing on LLaVA-1.5-7B and 13B, we show that grounded and hallucinated object commitments are linearly separable at intermediate-to-late layers. This signal is stable across model scales, robust under five-fold stratified cross-validation, and significantly above chance under label-permutation tests. Motivated by this finding, we propose Object-Presence Verification (OPV), a fine-tuning-free decoding-time method that verifies completed object mentions with a calibrated linear probe and, when a mention is judged unsupported, rolls back generation while banning the same object identity and its surface variants. On CHAIR and POPE, OPV substantially reduces object hallucination over strong inference-time baselines while improving precision and F1 with modest decoding overhead.

1 Introduction

Large vision-language models (LVLMs) such as LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023a), and InternVL (Chen et al., 2024b) combine visual perception with the generative capacity of large language models, enabling open-ended image captioning and visual question answering. Despite this progress, LVLMs often produce object hallucinations: they mention objects that are not present in the input image (Rohrbach et al., 2018; Li et al., 2023b). This failure is especially concerning because it occurs at the entity level. A single unsupported object mention can change the meaning of a caption and propagate through downstream retrieval, question-answering, or decision-support systems.

This paper asks a representational question: when an LVLM is about to generate an object mention, does its hidden state already contain information about whether that object is grounded in the image? If the answer is yes, then object hallucination is not only an output-level failure, but also a phenomenon that leaves a detectable trace inside the model. Understanding the structure of this trace is important independently of any particular mitigation method: it tells us whether hallucination-related information is accessible, where it appears in the network, and how simple a readout can recover it.

Prior work has made progress on both hallucination mitigation and hidden-state analysis, but the structure of object-grounding signals in LVLMs remains insufficiently characterized. Inference-time methods such as VCD (Leng et al., 2024), M3ID (Favero et al., 2024), OPERA (Huang et al., 2024), Attention Lens (Jiang et al., 2025b), and AGLA (An et al., 2025) reduce hallucination by modifying decoding distributions, attention patterns, or visual grounding signals. Separately, hidden-state methods have shown that internal activations can be predictive of truthfulness or hallucination in language models and LVLMs (Azaria and Mitchell, 2023; Su et al., 2024; Chen et al., 2024a; Duan et al., 2025). However, it remains unclear whether object hallucination in LVLMs is encoded in a simple linearly accessible form, at which layers this signal is strongest, and whether it is stable across model scales.

We address this gap through a layer-wise probing study of LVLM hidden states. We generate free-form captions with LLaVA-1.5-7B and LLaVA-1.5-13B, extract object mentions from the outputs, and label each mention as grounded or hallucinated using the CHAIR synonym mapping (Rohrbach et al., 2018) against image annotations. For each mention, we collect the hidden state at the decoding step

immediately before the object name is completed, i.e., the point at which the model is about to commit to an object identity. We then train a logistic-regression probe independently at each language-model layer and evaluate separability with five-fold stratified cross-validation.

Our main empirical finding is that object hallucination leaves a simple and stable trace in LVLM hidden states. At the moment before an object mention is completed, grounded and hallucinated object commitments become linearly separable across a broad range of intermediate-to-late layers, and this separation is robust under cross-validation and label-permutation tests.

This finding suggests a natural intervention principle. Since the signal is observed at the point where the model completes an object identity, mitigation should operate at the same granularity. We therefore propose Object-Presence Verification (OPV), a decoding-time procedure that uses the calibrated linear probe as an object-commitment verifier. OPV monitors the token stream, queries the probe when an object phrase is completed, and, if the calibrated grounding score falls below a threshold, rolls generation back to the prefix before that object. It then resumes decoding while banning the same object identity, including synonyms and surface variants. OPV is not the starting point of our analysis; rather, it is a direct consequence of the representational structure revealed by the probing study.

We evaluate OPV on CHAIR and POPE with LLaVA-1.5-7B and LLaVA-1.5-13B. Across both model scales, OPV substantially reduces object hallucination compared with strong inference-time baselines, while improving precision and F1. Ablations further show that the AUC-selected layer is a better operational choice than the final layer, and that OPV is stable across nearby threshold and rollback-budget settings.

Our contributions are:

- We provide an empirical characterization of object-grounding signals in LVLM hidden states, showing that grounded and hallucinated object mentions are linearly separable at object-completion boundaries.

- Building on this representational finding, we propose OPV, a fine-tuning-free decoding-time method that verifies completed object commitments and applies identity-level rollback with synonym and surface-variant banning.

- We demonstrate that OPV substantially reduces object hallucination on CHAIR while improving precision and F1 with only modest decoding overhead.

2 Related Work

LVLMs and hallucination benchmarks. Large vision-language models such as LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023a), and InternVL (Chen et al., 2024b) couple a pretrained visual encoder with a large language model via a cross-modal interface, and despite strong performance on captioning and visual question answering they still produce object mentions that the input image does not support. The dominant benchmark for this failure mode is CHAIR (Rohrbach et al., 2018), which measures object hallucination in free-form captioning by comparing generated object mentions against an image's annotations. We adopt CHAIR as our evaluation throughout this paper.

Inference time mitigation of object hallucination. A growing body of work mitigates object hallucination at inference time without retraining the LVLM. *Contrastive-decoding* approaches contrast the model's distribution under the original image against an unconditional or noised distribution and reweight tokens accordingly (VCD (Leng et al., 2024); M3ID (Favero et al., 2024)). *Attention-based* methods reshape self-attention patterns or selectively suppress image-inattentive heads (OPERA (Huang et al., 2024); Attention Lens (Jiang et al., 2025b); SPIN (Sarkar et al., 2025); modular attribution (Yang et al., 2025)). *Visual-grounding guidance* injects evidence from external vision models or strengthens locally grounded attention (MARINE (Zhao et al., 2025); AGLA (An et al., 2025); fine-grained token grounding (Nguyen et al., 2026)). *Post-hoc* approaches edit the response after generation by cross-checking object mentions (LURE (Zhou et al., 2024); LogicCheckGPT (Wu et al., 2024)), and *representation editing* steers latent features away from directions associated with hallucinated objects (ProjectAway (Jiang et al., 2025a)). Across these families, the intervention is typically applied at the token or distribution level, without using the model's internal states to make an explicit per-object grounding decision.

Hidden-state signals. A growing body of work uses model internals to detect truthfulness and grounding (Azaria and Mitchell, 2023; Su et al., 2024; Chen et al., 2024a). In LVLMs, TruthPrInt (Duan et al., 2025) is the most closely related work: it trains a per-token MLP detector at a fixed middle layer and applies a traceback mechanism that replaces the lowest-confidence preceding token. OPV shares the broad idea of hidden-state-based mitigation but reverses the order of investigation. Rather than fixing a layer and engineering a detector, we first characterize the object-grounding signal — its linear accessibility, layer stability, and geometry — and let the analysis dictate the design. The intervention is therefore not an independent engineering choice but follows from of the empirical characterization in Section 3.

3 Hidden States Linearly Encode Object Grounding

The purpose of this section is diagnostic: we use probing analyses to characterize where and how object-grounding information is represented, without yet modifying decoding. We ask whether grounded and hallucinated mentions are separable in hidden-state space, at which layers the separation is strongest, and whether a linear readout already suffices; the answers determine whether a lightweight decoding-time intervention is plausible.

3.1 Representation Collection

We sample randomly 500 images from MS-COCO val2014 (Lin et al., 2014) and generate one caption per image using greedy decoding with LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2023). Object mentions are extracted from the generated captions and labeled as grounded or hallucinated using the CHAIR synonym mapping (Rohrbach et al., 2018) against the image annotations, yielding 4,098 mentions for the 7B model (24.2% hallucinated) and 4,070 mentions for the 13B model (20.2% hallucinated). For each mention we record the hidden state at every language-model layer at the decoding step immediately before the token that completes the object name that is the position at which the model is about to commit to a full object identity.

3.2 Diagnostic Linear Probing

To measure how accessible the grounding signal is in the hidden representation, we fit a logistic-regression probe independently at each language-model layer. Within each cross-validation split, we standardize each feature dimension using statistics computed only on the training folds; the probe at layer ℓ is

$$p_{ij}^{(\ell)} = \sigma\left(\mathbf{w}_\ell^\top \tilde{h}_{ij}^{(\ell)} + b_\ell\right),$$

where $h_{ij}^{(\ell)}$ is the hidden state for object mention j in image i at layer ℓ, and $\mathbf{w}_\ell \in \mathbb{R}^d$, $b_\ell \in \mathbb{R}$ are parameters fit on the training folds via logistic regression against the CHAIR-derived grounded/hallucinated labels. Each layer is evaluated under five-fold stratified cross-validation, which preserves the grounded/hallucinated ratio in each fold.

We use AUC as the primary metric because our goal in this section is to characterize how well each layer's representation encodes the grounding signal, not to evaluate classifier performance at a specific operating point. AUC summarizes separability across the full score range and therefore reflects the layer's representational quality as a whole.

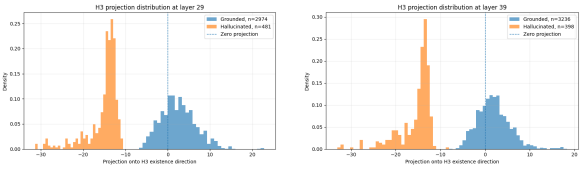
3.3 Layer-wise Separability

The layer-wise probe results show that object-grounding information is not confined to any single layer. On LLaVA-1.5-7B, the highest cross-validated AUC is obtained at layer 17, reaching above 0.97; on LLaVA-1.5-13B, the best layer is layer 35, reaching above 0.96. In both backbones, the high-AUC region forms a broad plateau, with all probed layers exceeding AUC 0.92, indicating that the grounding signal is a stable property of the middle-to-late representation rather than an artifact of a single layer.

The separation is not driven by class imbalance or fold variance: bootstrap confidence intervals at the AUC-peak layers are narrow (< 0.01 wide), and label-permutation tests reject the null at $p < 10^{-3}$. When hidden states are projected onto the learned direction $\mathbf{w}$, grounded and hallucinated mentions occupy clearly different regions of the one-dimensional projection space (Figure 1), which explains why a linear probe suffices: the relevant grounding signal is exposed along a single direction at strong layers.

4 Method

Section 3 shows that hidden states at object-completion boundaries contain a strong linear signal separating grounded from hallucinated object



(a) LLaVA-1.5-7B (b) LLaVA-1.5-13B

Figure 1: Distribution of hidden-state projections along the learned linear direction $\mathbf{w}$ at the AUC-peak layer of each model. Grounded and hallucinated mentions form well-separated clusters in both cases.

mentions. We now turn this diagnostic observation into a decoding-time intervention.

The key design principle is to intervene at the level at which the signal is observed. Since the probe is queried when the model is about to complete an object identity, OPV treats the entire object mention as the unit of verification. If an object commitment is judged unsupported, OPV rolls generation back to the prefix before that object and prevents the model from immediately regenerating the same object identity under a synonymous or surface-variant form.

OPV does not fine-tune the backbone LVLM. The backbone parameters remain fixed throughout. The only learned component is a lightweight linear probe, fitted once per backbone from the representation analysis in Section 3 and reused during inference.

4.1 Overview

Given an image I and textual prompt X, the LVLM generates an output sequence autoregressively:

$$p(y_t \mid I, X, y_{<t}).$$

OPV modifies this decoding process only when the newly generated tokens complete an object phrase. At each such boundary, OPV performs three steps (Figure 2).

First, an object-completion detector checks whether the current token stream ends with a known object phrase. Second, a calibrated linear probe reads the hidden state at the selected layer and estimates whether the current object commitment is visually grounded. Third, if the grounding score falls below a threshold, OPV rolls the sequence back to the position before the object phrase and resumes generation while banning the same object identity, including its synonyms and surface variants.

This design differs from token-level correction methods. OPV does not suppress arbitrary low-confidence tokens, nor does it continuously reshape the output distribution at every decoding step. Instead, it verifies discrete object commitments and intervenes only when the model has just completed an object mention.

4.2 Object-Completion Detection

OPV queries the probe only at object-completion boundaries. This avoids applying the probe at every decoding step and aligns inference-time verification with the probing position analyzed in Section 3.

We maintain a dictionary $\mathcal{D}$ of object identities eligible for verification. For each object category, $\mathcal{D}$ contains a synonym group from the CHAIR object mapping (Rohrbach et al., 2018). Each synonym is expanded into surface variants, including singular and plural forms, capitalization variants, and tokenizer-specific subword segmentations. Let $\mathcal{S}_{\text{obj}}$ denote the resulting set of token sequences.

At decoding step t, let the generated sequence be

$$y_{1:t} = (y_1, \ldots, y_t).$$

We say that an object identity o has just been completed at step t if there exists a token sequence $s(o) \in \mathcal{S}_{\text{obj}}$ of length m such that

$$y_{t-m+1:t} = s(o).$$

The detector is implemented as a suffix match over tokenized object phrases. Multi-token object names are therefore handled naturally: the detector fires only once, at the step where the full object phrase is completed.

4.3 Risk Estimation from Hidden States

When the detector identifies a completed object phrase $s(o)$ at step t, OPV reads the hidden state at the layer selected in Section 3. Let $h_t^{(\ell^\star)}$ be the hidden state at layer $\ell^\star$, where $\ell^\star$ is chosen by cross-validated AUC in the diagnostic probing stage.

The hidden state is standardized using the training statistics of the probe:

$$\tilde{h}_t^{(\ell^\star)} = \frac{h_t^{(\ell^\star)} - \mu}{\sigma}.$$

OPV then computes a scalar projection along the learned linear direction:

$$r_t = \mathbf{w}^\top \tilde{h}_t^{(\ell^\star)}.$$

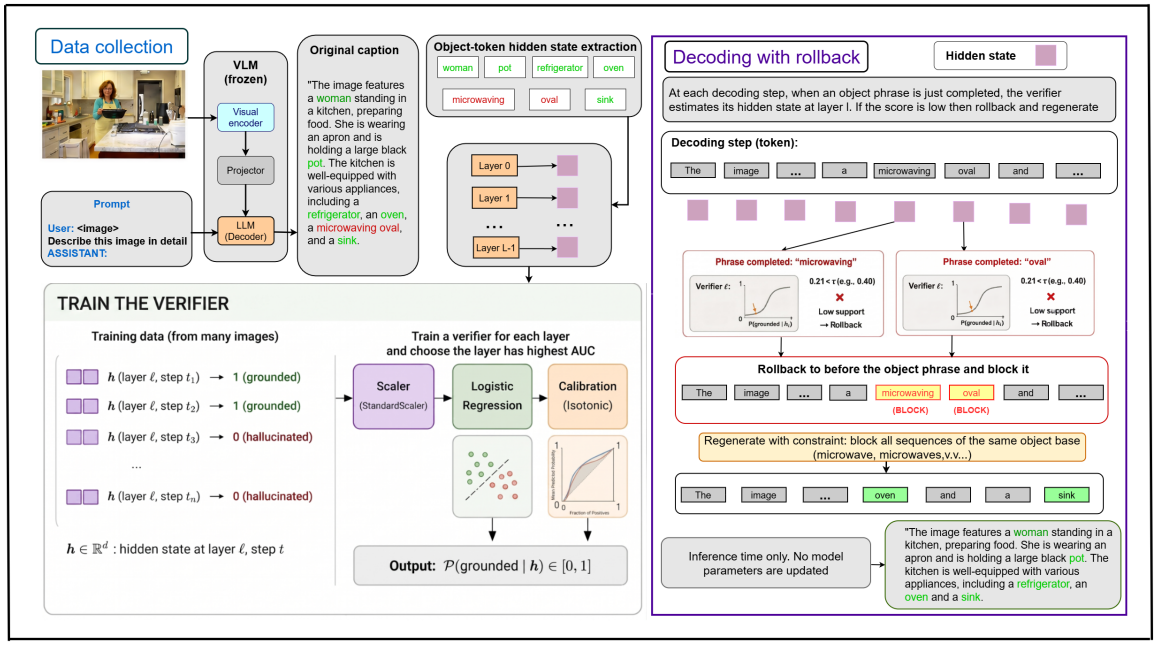


Figure 2: Overview of Object-Presence Verification (OPV)

Because the raw projection score is not necessarily calibrated, we map it to a grounding score using a one-dimensional calibrator:

$$p_t = C(r_t),$$

where $C(\cdot)$ is fitted on a held-out validation split.

The value p_t is used as a risk signal for the current generation state. A higher value indicates that the hidden state is more consistent with grounded object completions observed during probe training. Importantly, p_t is not interpreted as an independent visual detector for the existence of object o in the image. It is a calibrated readout of the LVLM's internal state at the moment the model commits to an object mention.

OPV accepts the object commitment if

$$p_t \geq \tau,$$

where τ is selected on validation. If $p_t < \tau$, the commitment is treated as unsupported and OPV triggers rollback.

4.4 Identity Rollback

When an object commitment is rejected, OPV removes the entire object phrase rather than modifying only the final token. Suppose the detected object phrase $s(o)$ has length m and is completed at step t. OPV rewinds the generated sequence to the prefix before the object phrase:

$$y_{1:t} \leftarrow y_{1:t-m}.$$

The decoding cache and position index are restored consistently with this prefix. Generation then resumes from the same context in which the model originally chose to mention o, but under an additional constraint that prevents immediate regeneration of the same object identity.

A rollback alone is insufficient, because the model may regenerate the same object using a different surface form. For example, after rejecting *bicycle*, the model may generate *bike*, *cycle*, or a plural/capitalized variant. OPV therefore maintains a per-image ban set $\mathcal{B}$ of token sequences. When object identity o is rejected, the ban set is updated as

$$\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{V}(o),$$

where $\mathcal{V}(o)$ contains all token sequences corresponding to the same object identity, including CHAIR synonyms, plural forms, capitalization variants, and alternative tokenizations.

During subsequent decoding, OPV applies a constrained logit update. Let $a_t \in \mathbb{R}^{|\mathcal{V}_{\text{tok}}|}$ be the logits at step t, and let $\mathcal{C}_t$ be the set of next-token candidates that would complete or continue a banned sequence given the current suffix. OPV penalizes these candidates:

$$a_t[k] \leftarrow a_t[k] - \lambda, \qquad k \in \mathcal{C}_t,$$

where λ is a large positive penalty. In practice, λ can be set large enough to approximate hard banning while preserving numerical stability.

This identity-level constraint ensures that OPV does not merely replace one surface form of a hallucinated object with another. The intervention targets the object identity, not the specific tokenization by which it was expressed. To keep inference cost predictable, OPV caps the total number of rollbacks per generation at B; once the cap is reached, decoding continues under the current ban set without further intervention. The complete decoding procedure in pseudocode is given in Appendix C.

5 Experiments

We evaluate whether the hidden-state signal characterized in Section 3 can support effective decoding-time hallucination mitigation when used by OPV. Since OPV verifies and rewrites completed object mentions, we evaluate primarily on CHAIR, which directly measures object hallucination in free-form captioning. We additionally report offline POPE results, to verify that the gains of OPV extend to a complementary evaluation that probes object grounding through binary yes/no decisions derived from generated captions.

5.1 Experimental Setup

Models. We evaluate OPV on LLaVA-1.5-7B and LLaVA-1.5-13B (Liu et al., 2023) with 500 MS-COCO val2014 images, disjoint from the 500 images used for probe training. The backbone LVLMs are kept fixed in all experiments. OPV uses one probing layer per backbone, selected by the layer-wise probing analysis in Section 3, which identifies the representation with the strongest linear separability between grounded and hallucinated object commitments. Specifically, we use layer 17 for LLaVA-1.5-7B and layer 35 for LLaVA-1.5-13B. Section 5.4 further verifies that these AUC-selected layers provide stronger hallucination reduction than the final language-model layer.

OPV configuration. The linear probe is trained using the diagnostic data described in Section 3 and calibrated on a held-out validation split. OPV uses a single verification threshold τ and a maximum rollback budget B. Unless otherwise stated, the main results use the validation-selected threshold and a rollback cap of $B = 4$. OPV uses the same maximum decoding length as greedy decoding, and no backbone parameters are updated.

5.2 Results

OPV substantially reduces object hallucination while preserving both the length and the descriptive coverage of the original captions. Across both LLaVA-1.5 scales, hallucination rates drop by roughly a factor of five to seven on both sentence-level and instance-level CHAIR, yet the proportion of ground-truth objects mentioned in the caption changes by only two to four percentage points (Table 1). Average caption length under OPV remains at the same $L_{\max}$=256, so the reduction is not achieved by truncating output; OPV continues to describe most of what is actually in the image, but rejects the mentions that the linear probe identifies as unsupported.

This selectivity is reflected in precision and F1. Precision rises by more than 20 points on 7B and over 16 points on 13B, indicating that the mentions retained after intervention are predominantly grounded. Coverage drops only modestly, so the F1 trade-off is clearly positive on both backbones. The pattern is consistent across model scales: doubling the parameter count does not change how OPV behaves — hallucinated mentions are removed, while grounded mentions and overall caption length are largely preserved.

5.3 Results on POPE

Beyond free-form captioning, we evaluate OPV on POPE, which derives binary object-existence yes/no questions directly from generated captions rather than asking the LVLM yes/no questions interactively. To elicit object lists for POPE, we prompt the LVLM with the following template, which restricts the response to comma-separated object nouns:

```
USER: <image>
List the clearly visible objects in the image.
Use comma-separated object nouns only.
No full sentences. No colors. No actions.
No details.
ASSISTANT:
```

The yes/no questions are then constructed from the objects listed in the response.

Table 2 shows the results. OPV improves Precision from 92.80 under greedy decoding to 96.2, and $F_{0.1}$ from 91.73 to 94.4. Relative to Truth-PrInt, which uses a hidden-state detector applied per token, OPV further improves Precision by 1.92 points (94.28 $\rightarrow$ 96.2) and $F_{0.1}$ by 1.93 points (92.47 $\rightarrow$ 94.4). These gains are consistent with

Model	Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	Cov. ↑	Prec. ↑	F1 ↑
LLaVA-1.5-7B	Greedy	48.2	13.9	76.9	74.0	75.4
	VCD (Leng et al., 2024)	47.8	14.7	76.5	–	–
	OPERA (Huang et al., 2024)	44.6	13.0	74.8	–	–
	ProjectAway (Jiang et al., 2025a)	42.0	12.2	–	–	–
	AGLA (An et al., 2025)	43.0	14.1	**78.9**	–	–
	Attention Lens (Jiang et al., 2025b)	25.0	6.7	–	–	76.1
	SPIN (Sarkar et al., 2025)	16.6	5.6	–	–	74.6
	TruthPrInt† (Duan et al., 2025)	10.3	3.9	–	–	–
	OPV (ours)	**8.6**	**2.0**	73.3	**95.7**	**83.1**
LLaVA-1.5-13B	Greedy	40.8	10.9	**79.0**	78.4	78.7
	SPIN (Sarkar et al., 2025)	29.2	7.9	–	–	79.1
	OPV (ours)	**9.6**	**2.3**	76.7	**94.9**	**82.4**

Table 1: Object-hallucination results on CHAIR. CHAIR_s is the per-sentence hallucination rate and CHAIR_i is the per-mention hallucination rate. Lower is better for both. "–" indicates the metric is not reported by the corresponding paper. † TruthPrInt numbers are taken directly from Duan et al. (2025) .

Method	**Precision** ↑	$F_{0.1}$ ↑
Greedy	92.80	91.73
TruthPrInt	94.28	92.47
OPV (ours)	**96.2**	**94.4**

Table 2: Overall results on offline POPE for LLaVA-1.5, averaged across the Random, Popular, and Adversarial splits.

Model	Layer	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	Cov. ↑	F1 ↑
7B	0	8.3	22.2	70.0	77.7
	17⋆	**2.0**	**8.6**	73.3	**83.1**
	18	3.0	10.6	73.1	82.5
13B	0	5.6	23.0	**75.8**	81.5
	35⋆	**2.3**	**9.6**	72.8	**82.4**
	36	2.8	11.0	72.9	82.0

Table 3: Ablation on the probe layer used by OPV. The symbol ⋆ marks the AUC-selected layer used in all main experiments. Early layers (layer 0) yield substantially weaker hallucination reduction despite comparable AUC in probe diagnostics, while upper-half layers (17, 18 on 7B; 35, 36 on 13B) achieve strong CHAIR reduction. Bold marks the best value per column within each backbone.

the CHAIR results: by rejecting unsupported object commitments at the granularity at which the hidden-state signal is most reliable, OPV reduces false-positive object mentions both in free-form captioning and in the binary object-existence setting probed by POPE.

5.4 Ablation Study

We ablate three OPV-specific design choices: the probing layer $\ell^\star$, the maximum rollback budget B, and the verification threshold τ. The layer ablation selects a single operational layer per backbone, while the other two ablations test whether OPV depends on aggressive rejection, and whether its performance is stable under nearby threshold choices.

Layer choice. Although Section 3 reports AUC above 0.92 at every probed layer, early layers exhibit substantially higher fold-to-fold variance than upper-half layers (Appendix A.1). Table 3 tests whether this variance difference matters for intervention by comparing three layers on CHAIR. Layer 0 gives the weakest reduction (CHAIR_s 22.2 on 7B, 23.0 on 13B) despite its high diagnostic AUC, consistent with the hypothesis that a noisy probe direction transfers poorly to inference. Upper-half layers, where AUC is both high and stable, achieve the strongest reduction; neighbouring layers within the plateau (17 vs. 18 on 7B, 35 vs. 36 on 13B) give nearly identical results, confirming robustness to small layer perturbations.

Maximum rollback budget. The maximum rollback budget B controls how often OPV is allowed to backtrack and regenerate before falling back to the current sequence. Table 4 shows that hallucination reduction is already strong at $B = 1$ ($\text{CHAIR}_i = 0.0$, $\text{CHAIR}_s = 0.2$), but this comes at the cost of coverage (66.2) and caption length (58.9), suggesting that an overly small budget reduces hallucination partly by suppressing object generation rather than by selectively rejecting unsupported commitments. Raising B to two restores most of the length and coverage (87.2 and 72.4) while keeping CHAIR_s low at 7.6; further increases to $B \in \{3, 4, 5\}$ produce only marginal changes, with F1 saturating around 83.1.

B	CHAIR$_i$	CHAIR$_s$	Cov.	Prec.	F1	Len.
1	0.0	0.2	66.2	99.9	79.7	58.9
2	1.8	7.6	72.4	96.2	82.6	87.2
3	1.9	8.4	73.3	95.9	83.1	90.7
4	2.0	8.6	73.3	95.7	83.1	92.0
5	2.0	8.6	73.3	95.7	83.1	92.1

Table 4: Effect of the maximum rollback budget B on CHAIR performance (LLaVA-1.5-7B). Performance saturates after three to four rollbacks.

Crucially, increasing B does not translate into a proportional increase in inference time (Table 6). Going from $B = 1$ to $B = 5$ raises the slowdown over greedy decoding only from $1.12\times$ to $1.25\times$ — a 13-percentage-point increase despite a five-fold rise in the maximum budget. This sublinear scaling indicates that rollbacks are issued sparingly in practice: the worst-case budget rarely binds, and most captions complete with far fewer than B rollbacks. We therefore use $B = 4$ in the main experiments, which preserves caption length and coverage at an overhead of $1.23\times$ over greedy decoding.

Verification threshold. The robustness of OPV to the threshold τ is a direct consequence of the calibrated probability geometry observed in Section 3. After isotonic calibration, the calibrated grounding probabilities concentrate near 1 for grounded mentions and near 0 for hallucinated mentions, with a narrow ambiguity band in between (Appendix A.2). Because the two modes are well separated, any threshold placed inside the band partitions most mentions identically; the operating point τ is therefore not a fragile tuning parameter but the midpoint of a wide robust interval.

We confirm this empirically in Table 5 by varying τ by small offsets $\delta \in \{-0.2, -0.1, +0.1, +0.2\}$ on LLaVA-1.5-7B. Within $\delta \in [-0.1, +0.2]$, CHAIR$_i$ stays in $[1.7, 2.1]$ and F1 in $[83.0, 83.3]$ — essentially flat, as predicted by the bimodal structure of the calibrated probabilities. Only at $\delta = -0.2$, where the threshold begins to enter the high-density region of the hallucinated cluster, does performance degrade noticeably (CHAIR$_s$ rises to 12.6, precision drops to 93.2). OPV's selection of τ is therefore robust to small calibration errors, and its sensitivity profile reflects the underlying geometry of grounded versus hallucinated representations rather than fine-grained hyperparameter tuning.

δ	CHAIR$_i$ ↓	CHAIR$_s$ ↓	Cov. ↑	Prec. ↑	F1 ↑	Len.
-0.2	3.1	12.6	74.8	93.2	83.0	91.6
-0.1	2.1	8.6	73.8	95.6	83.3	91.6
$+0.1$	2.0	8.8	73.3	95.7	83.0	92.0
$+0.2$	**1.7**	**8.0**	72.9	**96.2**	83.0	91.9

Table 5: Sensitivity of OPV to perturbations of the verification threshold (LLaVA-1.5-7B). δ is the offset applied to the validation-learned threshold τ; the operating point $\delta = 0$ is reported in Table 1. OPV remains stable across small perturbations of τ but degrades when the threshold is shifted aggressively downward.

Method	Time (s) ↓	Slowdown ↓	$L_{\max}$	B
Greedy	6.10	$1.00\times$	256	–
OPV	6.84	$1.12\times$	256	1
OPV	7.14	$1.17\times$	256	2
OPV	7.38	$1.21\times$	256	3
OPV	7.49	$1.23\times$	256	4
OPV	7.61	$1.25\times$	256	5

Table 6: Runtime overhead of OPV under different rollback budgets. OPV adds moderate overhead relative to greedy decoding.

6 Conclusion

We characterized the structure of object-grounding signals in LVLM hidden states, showing that a single linear direction at an intermediate layer separates grounded from hallucinated object mentions across two LLaVA-1.5 scales and is competitive with deeper probe architectures. Building on this characterization, we proposed *Object-Presence Verification* (OPV), a fine-tuning-free decoding-time procedure that performs identity-level rollback with synonym and surface-variant banning when the calibrated linear probe flags an object commitment as unsupported. On CHAIR, OPV substantially reduces object hallucination across both backbones — CHAIR$_s$ drops from 48.2 to 8.6 on the 7B model and from 40.8 to 9.6 on the 13B model — while improving precision by more than 16 points and keeping caption length comparable to greedy decoding. The same calibrated grounding signal also improves Precision and $F_{0.1}$ on offline POPE relative to greedy decoding and prior hidden-state probe methods, confirming that the gains generalize beyond free-form captioning. Natural extensions include scaling the same probing-then-rollback framework to attribute, count, and relation hallucinations, and characterizing the transferability of the probe across backbones and image distributions.

7 Limitations

OPV addresses object hallucination in LVLMs but leaves several axes of generalization open.

Object-only scope. The probe and the rollback mechanism operate on *object identity*. Other hallucination modes that LVLMs are known to exhibit — incorrect attributes (color, shape, count), spatial relations, or actions — are outside the scope of this method. Extending the same probing-then-rollback framework to these modes requires labeling protocols beyond CHAIR and most likely a richer trigger than object-name completion.

Dictionary dependence. The completion detector relies on a dictionary of object names. We instantiate it from the MS-COCO categories with their CHAIR synonym groups, which covers the evaluation benchmarks used here but does not include objects outside this vocabulary. Generalizing OPV to open-vocabulary captioning would require an open-set object detector or a learned object-name segmenter; we leave this to future work.

Distribution of training data. The linear probe and its calibrator are fit on captions generated by LLaVA-1.5 over 500 MS-COCO val2014 images. Transferring the probe to substantially different distributions — e.g., medical, satellite, or screenshot imagery — is likely to require re-fitting on representative data, and we have not characterized how much data is needed.

Architectural breadth. We evaluate OPV on LLaVA-1.5-7B and 13B, two scales of a single architectural family. The probing analysis in Section 3 is consistent with hidden-state-level analyses on other models, but the specific layer choices, thresholds, and ban-penalty magnitudes are not guaranteed to transfer without re-tuning to backbones with different vision–language connectors (e.g., Q-Former-based BLIP-2 or InstructBLIP, or Qwen-VL).

Inference cost in the worst case. The amortized overhead of OPV is small in practice because rollbacks are infrequent. However, on an image whose decoded caption repeatedly drifts toward unsupported objects, the rollback budget can be exhausted before the model produces a grounded continuation, and generation then proceeds under the current ban set without further intervention. The rollback cap B is therefore a hard upper bound on cost, not a guarantee that every hallucination will be removed.

References

Wenbin An and 1 others. 2025. 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)*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024a. INSIDE: LLMs' internal states retain the power of hallucination detection. In *Proceedings of the Twelfth International Conference on Learning Representations*.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, and 23 others. 2024b. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *Preprint*, arXiv:2412.05271.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. TruthPrInt: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7372–7382.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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*.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. 2025a. Interpreting and editing Vision-Language representations to mitigate hallucinations. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

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

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Singapore. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Yutong Xie, Cam-Tu Nguyen, Minh Khoi Nguyen, Dang Huy Pham Nguyen, Anton van den Hengel, Johan W. Verjans, Phi Le Nguyen, and Vu Minh Hieu Phan. 2026. Beyond the global scores: Fine-grained token grounding as a robust detector of LVLM hallucinations. *Preprint*, arXiv:2604.04863.

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

Sreetama Sarkar, Yue Che, Alex Gavin, Peter Anthony Beerel, and Souvik Kundu. 2025. Mitigating hallucinations in vision-language models through image-guided head suppression. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12481–12500, Suzhou, China. Association for Computational Linguistics.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14379–14391, Bangkok, Thailand. Association for Computational Linguistics.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. 2024. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6944–6962, Bangkok, Thailand. Association for Computational Linguistics.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2025. 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*, pages 77461–77486. PMLR.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *Proceedings of the Twelfth International Conference on Learning Representations*.

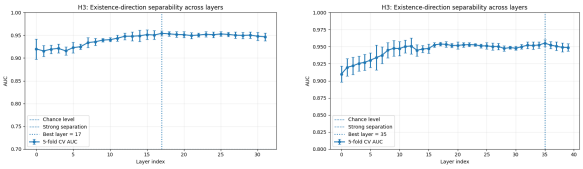
A Additional Probing Analyses

A.1 Layer-wise AUC

Figure 3 reports the linear probe's AUC at every layer of the language component on both backbones. AUC remains high across a broad middle–late range, peaking at layer 17 on LLaVA-1.5-7B and at layer 35 on LLaVA-1.5-13B, with neighboring layers reaching AUC within 0.003 of the maximum. The breadth of the high-AUC region makes the choice of probing layer robust to small implementation differences and allows the same probe to be re-trained for a new model without a fragile per-layer search.

A.2 Calibrated Probability Distributions

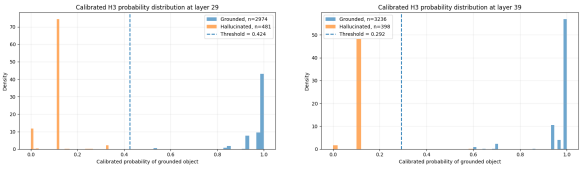
After isotonic calibration, calibrated grounding probabilities concentrate near 1 for grounded mentions and near 0 for hallucinated ones, with a narrow ambiguity band in between (Figure 4). The bimodal shape makes threshold selection robust: small perturbations of the thresholds leave the



(a) LLaVA-1.5-7B (b) LLaVA-1.5-13B

Figure 3: Layer-wise AUC of the linear probe across the language component. AUC remains high across a broad middle–late range, peaking at layer 17 on LLaVA-1.5-7B and at layer 35 on LLaVA-1.5-13B.

accept/reject partitioning of most mentions unchanged.



(a) LLaVA-1.5-7B (b) LLaVA-1.5-13B

Figure 4: Calibrated grounding probability after isotonic calibration at the chosen layer of each model (layer 17 for 7B, layer 35 for 13B). Dashed vertical lines mark the F1-optimal threshold selected on the validation fold.

B Maximum Token Length Ablation

We further examine how the maximum decoded length $L_{\max}$ affects OPV. As widely observed in the CHAIR literature, hallucination is closely tied to caption length: longer captions mention more objects and therefore more opportunities to hallucinate. Table 7 confirms this on LLaVA-1.5-7B: at $L_{\max} = 64$, captions average only 54.5 tokens, CHAIR_s drops to 3.0, but coverage also collapses to 62.4 as the model has no room to describe the image in detail. Raising $L_{\max}$ to 128 and then 256 progressively allows fuller descriptions, at the cost of higher hallucination rates.

We adopt $L_{\max} = 256$ as the default for all main experiments, including all baselines, precisely because it is the most demanding setting: captions are long enough to exercise the failure modes that CHAIR is designed to measure, and methods cannot trivially reduce hallucination by producing short captions. The fact that OPV still achieves $\text{CHAIR}_s = 8.6$ and F1 83.1 at this length is therefore evidence that the reduction is selective rather than a length artifact.

$L_{\max}$	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	Cov. ↑	Prec. ↑	F1 ↑	Len.
64	**0.8**	**3.0**	62.4	**98.4**	76.4	54.5
128	1.9	8.4	73.0	96.0	82.9	89.3
256	2.0	8.6	**73.3**	95.7	**83.1**	**92.0**

Table 7: Effect of the maximum decoded length $L_{\max}$ on OPV decoding (LLaVA-1.5-7B, CHAIR). Reducing $L_{\max}$ lowers hallucination at the cost of coverage and caption length; $L_{\max}$=256 is the default used in the main experiments. Bold marks the best value per column.

C OPV Decoding Algorithm

Algorithm 1 summarizes OPV. The method is applied online during autoregressive decoding. At each step, the LVLM generates a token normally. OPV checks whether this token completes an object phrase. If not, decoding continues unchanged. If an object phrase is completed, OPV computes the calibrated grounding score and either accepts the object or rolls back and updates the ban set.

D Qualitative Examples

Figures 5–7 present qualitative side-by-side comparisons between greedy decoding (baseline) and OPV on three MS-COCO val2014 images. Object mentions shown in green are correctly grounded against the CHAIR-mapped image annotations; mentions in red are hallucinations. Tokens without color highlighting are not tracked by the CHAIR mapping.

Example 1 (Street corner)

Baseline (greedy) caption

The image features a street corner with a traffic light and a street sign. The traffic light is positioned on the left side of the scene, while the street sign is located on the right side. The street sign is green and white, providing directions for drivers and pedestrians. There are two traffic lights in the scene, one being closer to the left side and the other being closer to the right side. The traffic lights are red, indicating that vehicles should stop at the intersection.

OPV caption

The image features a street corner with a traffic light and a street sign. The traffic light is positioned on the left side of the scene, while the street sign is located on the right side. The street sign is green and white, providing directions for the area. There are two traffic signs in the scene, one on the left side and the other on the right side. The left sign is positioned closer to the traffic sign, while the right sign is further away. The scene also includes a building in the background, adding to the urban atmosphere.

Figure 5: Qualitative comparison on a street-corner image. The baseline hallucinates *pedestrians* that are not present in the image; OPV avoids this hallucinated object and steers the description toward the actual contents of the scene.

Example 2 (Model airplane)

Baseline (greedy) caption

The image features a small, blue and white airplane with a star on its tail, flying through the air. The airplane appears to be a model or toy, as it is not a full-sized aircraft. The plane is positioned in the middle of the scene, with its wings spread out. There are two people visible in the image, one on the left side and the other on the right side. They seem to be observing the airplane as it flies by. The scene also includes a car parked in the background, slightly to the left of the center of the image.

OPV caption

The image features a small, blue and white airplane with a star on its tail, flying through the air. The airplane appears to be a model or toy, as it is not a full-sized aircraft. The model is positioned in the middle of the scene, with its wings spread out. The background of the image is a grassy field, providing a natural setting for the airborne model. The scene captures the essence of a small, colorful airborne object soaring through the sky.

Figure 6: Qualitative comparison on a model-airplane image. The baseline hallucinates both *people* and a *car* that are not present; OPV avoids these hallucinations and instead grounds the description in the actual background of the scene.

Example 3 (Office setting)

Baseline (greedy) caption

The image depicts a group of people working in an office setting. There are four people in the scene, with one man sitting at a desk using a laptop, and the other three people standing around him. The man using the laptop is wearing a blue shirt. The office is equipped with multiple computers, including a desktop computer with a monitor and keyboard, and a laptop on the desk. There are also two cell phones visible in the scene, one near the center and the other closer to the right side. In addition to the computers, there are two chairs in the room, one near the center and the other towards the right side. A handbag can be seen placed on the floor, close to the center of the room.

OPV caption

The image depicts a group of people working in an office setting. There are four people in the scene, with one man sitting at a desk using a laptop, and the other three people standing around him. The man using the laptop is wearing a blue shirt. The office is equipped with multiple desktops, keypads, and a mouse. There are two keypads on the desktops, and a mouse is placed on one of the desktops. Additionally, there are two laptops in the scene, one being used by the man and the other one placed on a desktop. A cellular telephone can be seen on one of the desktops, and a hand bag is placed on the floor near the standing people. The office environment appears to be a mix of traditional and modern technology.

Figure 7: Qualitative comparison on an office-setting image. OPV removes the hallucinated *cell phones* and *handbag* produced by the baseline while preserving the grounded mentions of *people*, *man*, *desk*, *laptop*, and *mouse*.

Algorithm 1 Object-Presence Verification

Require: Image I, prompt X, LVLM M, object dictionary $\mathcal{D}$, probe $(\mathbf{w}, C)$, layer $\ell^\star$, threshold τ, rollback budget B

Initialize generated sequence $y \leftarrow \emptyset$
Initialize ban set $\mathcal{B} \leftarrow \emptyset$
Initialize rollback counter $k \leftarrow 0$
while not stopped **do**
    Compute next-token logits from $M(I, X, y)$
    Penalize logits that would extend or complete sequences in $\mathcal{B}$
    Sample or select next token y_t
    Append y_t to y
    if y completes an object phrase $s(o)$ **and** $k < B$ **then**
        Read hidden state $h_t^{(\ell^\star)}$
        Compute calibrated score $p_t = C(\mathbf{w}^\top \tilde{h}_t^{(\ell^\star)})$
        if $p_t < \tau$ **then**
            Roll back y to the prefix before $s(o)$
            Update $\mathcal{B} \leftarrow \mathcal{B} \cup \mathcal{V}(o)$
            Restore decoding cache to the rolled-back prefix
            $k \leftarrow k + 1$
        end if
    end if
end while
return y