STATE-DEPENDENT VISUAL SUPPORT IN VISION-LANGUAGE HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prior studies have associated vision-language model hallucination with apparently conflicting internal visual signals, including reduced visual dependence and highly attended image tokens. We observe a more specific difference between settings. Hallucinated objects receive less visual support in free-form generation but more when the model makes queried object existence decisions. We show that this difference mainly reflects comparisons of internal representations serving different computational roles. To compare these roles directly, we present STATEHALL, which applies the same measure of visual support to states corresponding to an object introduced during free-form generation, a queried object, and the queried decision. Hallucinated object representations receive less support around intermediate decoder layers in both free-form generation and explicit querying, while relative over-support appears only at the late decision state. A same-image, same-object decomposition shows that the state-role component constitutes most of the final difference between the generated-object and queried-decision gaps, while the component associated with concept origin is substantially smaller. These patterns are also useful for hallucination detection. For free-form hallucination, the resulting midpoint concept-visual support (MCVS) score matches or improves upon recent training-free baselines and transfers across models and datasets. For queried object existence, compact detectors based on the identified states demonstrate strong transfer across benchmarks and complement output confidence.

1 INTRODUCTION

Vision-language models (VLMs) can describe complex images and answer visual questions while still asserting objects that are not present. Object hallucination is therefore a major reliability concern, motivating training-free and learned detectors Jiang et al. (2025); Che et al. (2025); Hoang et al. (2026); Wang et al. (2026); Kogilathota et al. (2026); Zhang et al. (2025; 2026), as well as dedicated benchmarks for object-existence probing Li et al. (2023); Neuhaus & Hein (2025); Xu et al. (2026) and free-form generation Wang et al. (2024); Kaul et al. (2024); Rohrbach et al. (2018).

A common explanation for object hallucination is weak visual grounding: The language model asserts the presence of an object that is plausible from the linguistic context but insufficiently supported by the image. This view is intuitive and often empirically useful. Yet recent analyses do not yield a universal internal signature. Hallucinated outputs can show reduced visual dependence or attention Jiang et al. (2025); Hoang et al. (2026), while other studies identify highly attended image tokens that can drive hallucination Che et al. (2025), and attention magnitude itself need not separate true from hallucinated outputs Wang et al. (2026). In our initial analysis, we observed an even sharper contradiction: Free-form hallucinated objects were visually *under-supported*, while false-positive object-existence decisions were visually *over-supported* late in the decoder.

The question is whether these observations reflect genuinely opposite hallucination mechanisms, or whether they arise because the two analyses examine different kinds of internal states. We argue for the latter. In free-form generation, visual support is measured at a state representing an object mentioned in the free-form response, while in object-existence probing, it is measured at a state used to form the final decision. These states serve different computational roles and therefore are not directly comparable. This motivates our analysis: We compare states serving the same computational

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.

role using the same visual support measure, and distinguish differences associated with whether the object concept is generated or supplied in the query from those associated with computational role.

We present STATEHALL, a state-aware framework for comparing visual support at different stages of VLM computation. We consider three text-side states: an object introduced during free-form generation, the same object when it appears in an explicit query, and the answer decision produced for that query. For each state, visual support measures how strongly it aligns with relevant visual representations, using exactly the same measure across all three states. This allows us to distinguish differences associated with concept origin from those associated with computational role. At the *object-representation stage*, hallucinated concepts are visually under-supported around intermediate layers in both free-form generation and explicit querying. At the *decision stage*, the support pattern reverses: Hallucinated cases become relatively more visually supported than true cases late in the decoder. For queried object existence, both midpoint object under-support and late decision over-support distinguish hallucinated from true cases.

A same-image, same-object comparison makes this distinction quantitative. We decompose the difference between the generated-object and queried-decision gaps into a component associated with concept origin, obtained by comparing generated-object and queried-object states, and a component associated with computational role, obtained by comparing queried-object and queried-decision states. At the final layer, the state-role component constitutes approximately 77% of this difference. Thus, differences associated with concept origin remain, but they constitute a substantially smaller part of this difference. An exact support-change decomposition further separates changes in support for visual evidence already emphasized at the queried-object state from changes in which visual evidence is emphasized at the decision state. The latter contribution becomes larger for hallucinated examples late in the decoder. We use this decomposition to characterize the measured state change rather than as evidence of a causal pathway.

The primary contribution of this work is the state-aware characterization of hallucination signatures across computational states. We further evaluate whether the identified patterns support simple hallucination detection. For free-form object hallucination, the resulting midpoint concept-visual support (MCVS) score uses low midpoint object support and matches or improves upon recent training-free detectors. For queried object existence, STATEHALL-U combines midpoint object support with final decision support and reaches 83.6% AUROC on Causal-HalBench under direct transfer. A learned detector using the same two state features improves performance further, and combining these state features with output confidence reaches 88.4% AUROC. Although output confidence is strong for binary object-existence decisions, the internal dynamics identified by STATEHALL remain useful for detection in both free-form generation and explicit querying. They also clarify how hallucination signatures differ across computational states.

2 RELATED WORK

Object-hallucination evaluation. POPE Li et al. (2023) popularized explicit object-existence probing, followed by the more challenging RePOPE Neuhaus & Hein (2025) and the causally motivated Causal-HalBench Xu et al. (2026). Free-form object hallucination is commonly evaluated with CHAIR Rohrbach et al. (2018), AMBER Wang et al. (2024), and THRONE Kaul et al. (2024). We study object hallucination in both settings, which differ in how the object concept enters the computation: In object-existence probing, the concept is supplied in the query before the model forms its answer, while in free-form generation, the model itself introduces the concept.

Internal visual signals and decoding-time mitigation. Decoding-time methods intervene directly in generation. Visual Contrastive Decoding (VCD) compares output distributions from original and distorted visual inputs Leng et al. (2024), while OPERA adjusts beam-search decoding using an over-trust penalty and retrospection-allocation strategy Huang et al. (2024). These are complementary to our setting as they modify the generated sequence, while we analyze and score fixed model outputs. Training-free detection and analysis methods have used attention and measures of visual dependence Phukan et al. (2025); Jiang et al. (2025); Che et al. (2025); Hoang et al. (2026); Wang et al. (2026). Reduced visual dependence can indicate hallucination Jiang et al. (2025); Hoang et al. (2026), while strongly attended image tokens can also contribute to hallucinated generation Che et al. (2025), and attention magnitude alone need not distinguish true from hallucinated outputs Wang et al. (2026). Instead of treating any single internal quantity as a universal signature, we apply the same

visual-support measure across computational states and decoder depths, asking whether apparently conflicting signatures reflect the measured quantity or the computational role of the state.

Learned hallucination detectors. Recent methods also learn hallucination signals from hidden representations or attention patterns. HALP Kogilathota et al. (2026) probes query-side hidden representations before generation, DHCP Zhang et al. (2025) learns from cross-modal attention patterns, and VIB-Probe Zhang et al. (2026) applies a variational bottleneck to attention-head outputs. Such learned signatures can perform well on the data used to fit them, while robustness under benchmark shift is important for practical detection. We therefore evaluate queried hallucination detection both in-domain and under direct benchmark transfer without refitting or recalibration.

3 State-aware visual support framework

We present a state-aware framework for measuring visual support across text states with different computational roles. Applying the same support measure throughout allows us to compare object representation and decision formation without state-specific scoring rules.

Object and decision states. Consider an image I, an object o, and a decoder with layers $\ell \in \{0, \dots, L-1\}$. Let $\boldsymbol{v}_{\ell j} \in \mathbb{R}^d$ be the hidden state of visual token j after decoder block ℓ. For the same object, we consider the generated-object state $\boldsymbol{h}_\ell^{\mathrm{gen,obj}}(o)$ when o is generated in a caption, the queried-object state $\boldsymbol{h}_\ell^{\mathrm{qry,obj}}(o)$ when o is supplied in an existence query, and the queried-decision state $\boldsymbol{h}_\ell^{\mathrm{qry,dec}}(o)$ at the last prompt position immediately before generation of the first answer token. For an object represented by multiple tokens, we average the corresponding token states. The first two states represent the object concept, while $\boldsymbol{h}_\ell^{\mathrm{qry,dec}}(o)$ corresponds to the model's decision about that object.

3.1 A common visual-support measure

For any text state $\boldsymbol{h}$ at decoder depth ℓ, define its cosine affinity to visual state $\boldsymbol{v}_{\ell j}$ as

$$a_{\ell j}(\boldsymbol{h}) = \frac{\boldsymbol{h}^\top \boldsymbol{v}_{\ell j}}{(\|\boldsymbol{h}\|_2 \|\boldsymbol{v}_{\ell j}\|_2)}. \tag{1}$$

We define the *visual support of the text state* as

$$S_\ell(\boldsymbol{h}) = \frac{1}{K} \sum_{j \in \mathbb{T}_K(\boldsymbol{h}, \ell)} a_{\ell j}(\boldsymbol{h}), \tag{2}$$

where $\mathbb{T}_K(\boldsymbol{h}, \ell)$ is the set of indices of the K visual states with the largest affinities. The same visual-support measure is applied to every predefined computational state, so comparisons across states do not depend on state-specific scoring rules.

To compare models with different decoder depths, we use normalized depth $\rho = \ell/(L-1)$, with midpoint $\ell_{\mathrm{mid}} = \lceil 0.5(L-1) \rceil$ and final layer $\ell_{\mathrm{final}} = L-1$. In our experiments, these correspond to blocks 18 and 35 for Qwen2.5-VL-3B-Instruct and blocks 16 and 31 for LLaVA-1.5-7B. We use S_{mid} and S_{final} to denote support at ℓ_{mid} and ℓ_{final}, respectively.

3.2 Hallucination support gaps and decomposition

Let H and T be the hallucinated and true object populations, respectively. We define a hallucination support gap for each of the generated-object, queried-object, and queried-decision states:

$$\Delta_\ell^s = \mathbb{E}[S_\ell^s \mid H] - \mathbb{E}[S_\ell^s \mid T], \tag{3}$$

where $S_\ell^s = S_\ell(\boldsymbol{h}_\ell^s)$ is the visual support evaluated at state s. Thus, $\Delta_\ell^s < 0$ indicates that hallucinated objects receive less visual support than true objects, while $\Delta_\ell^s > 0$ indicates that they receive more.

To decompose the difference between the two modes, we evaluate all three state gaps on the same set of image-object pairs (as described in Sec. 3.3). We define the cross-mode difference between the generated-object and queried-decision gaps,

$$C_\ell = \Delta_\ell^{\mathrm{qry,dec}} - \Delta_\ell^{\mathrm{gen,obj}}. \tag{4}$$

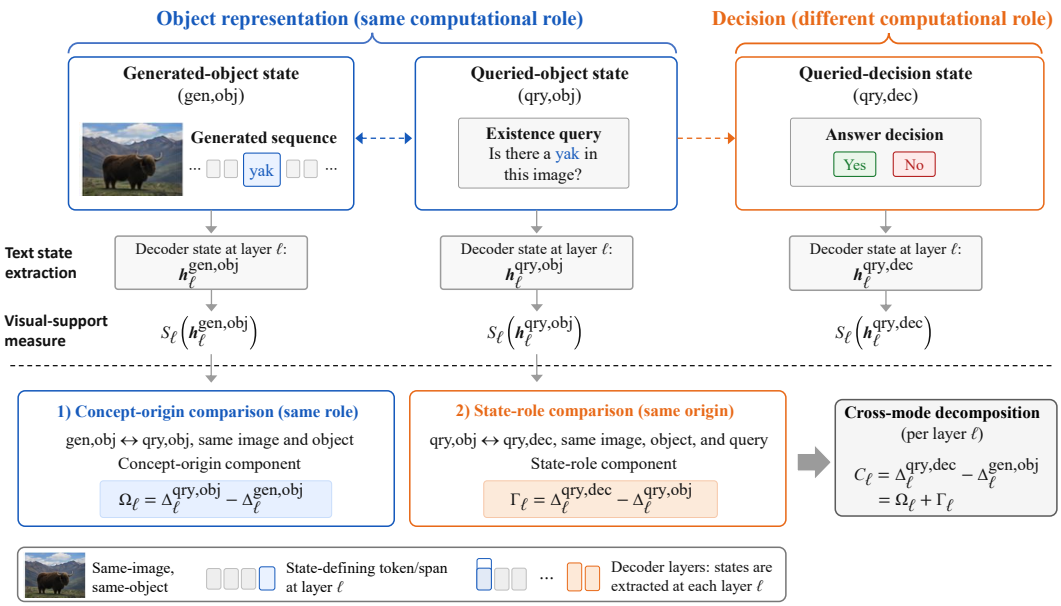

Figure 1: State-aware visual-support analysis. StateHall applies the same visual-support measure to generated-object, queried-object, and queried-decision states. The first two represent the object under different concept origins, while the third represents the resulting decision. The comparison between the two object states captures the concept-origin component, while that between the queried-object and queried-decision states captures the state-role component.

This reflects two factors simultaneously: the origin of the object concept (generated or supplied in the query) and the computational role of the state (object representation or decision). Using the queried-object state, we separate these two factors as

$$\Omega_\ell = \Delta_\ell^{\mathrm{qry,obj}} - \Delta_\ell^{\mathrm{gen,obj}}, \quad \Gamma_\ell = \Delta_\ell^{\mathrm{qry,dec}} - \Delta_\ell^{\mathrm{qry,obj}}, \quad C_\ell = \Omega_\ell + \Gamma_\ell. \tag{5}$$

Here, Ω_ℓ measures the difference associated with concept origin by comparing two object-representation states, while Γ_ℓ measures the difference associated with computational role by comparing the queried-object and queried-decision states. Equation (5) therefore separates the cross-mode difference into a concept-origin component and a state-role component.

3.3 State-aware analysis

Figure 1 summarizes the state definitions and the two comparisons used in the decomposition. For the same-image, same-object analysis, we take an object from a generated sequence and form an existence query for that object on the same image. The generated-object and queried-object states differ in concept origin while sharing the same image, object, and computational role. The queried-object and queried-decision states share the same image, object, and query but differ in computational role. We retain only the first occurrence of each unique (image,object) pair so that repeated occurrences of an object do not dominate the estimates.

The generated-object and queried-decision states differ in both concept origin and computational role. The queried-object state provides the intermediate representation needed to separate these two factors. Generated and queried token positions do not have to coincide, and a prompt token does not have to correspond directly to a generated answer token. Each evaluated state is defined by its computational role.[1]

In Eq. (5), Ω_ℓ and Γ_ℓ are differences between population support gaps and therefore characterize the cross-mode population difference rather than individual examples. Given model outputs and object labels, no learned component is required. The complete procedure is presented in Algorithm 1.

[1]We use *concept-origin component* for the generated-object and queried-object difference. Since the two states may also differ in linguistic context and token position, Ω_ℓ should be interpreted as associated with concept origin rather than as its causal effect.

3.4 DECOMPOSITION OF THE OBJECT-TO-DECISION SUPPORT CHANGE

The decomposition in Eq. (5) separates the concept-origin and state-role components. We further characterize the state-role component without assuming that the same visual states remain in the Top-K set. At a fixed layer ℓ, let

$$\mathbb{A}_\ell^{\mathrm{qry,obj}} = \mathbb{T}_K(\boldsymbol{h}_\ell^{\mathrm{qry,obj}}, \ell) \tag{6}$$

be the visual-state index set selected at the queried-object state, and let $S_{\ell,\mathbb{A}}(\boldsymbol{h})$ denote support computed using a fixed size-K index set $\mathbb{A}$. Then the object-to-decision support change decomposes as

$$\begin{aligned} &S_\ell(\boldsymbol{h}_\ell^{\mathrm{qry,dec}}) - S_\ell(\boldsymbol{h}_\ell^{\mathrm{qry,obj}}) \\ &\quad = \Big[S_{\ell,\mathbb{A}_\ell^{\mathrm{qry,obj}}}(\boldsymbol{h}_\ell^{\mathrm{qry,dec}}) - S_{\ell,\mathbb{A}_\ell^{\mathrm{qry,obj}}}(\boldsymbol{h}_\ell^{\mathrm{qry,obj}})\Big] + \Big[S_\ell(\boldsymbol{h}_\ell^{\mathrm{qry,dec}}) - S_{\ell,\mathbb{A}_\ell^{\mathrm{qry,obj}}}(\boldsymbol{h}_\ell^{\mathrm{qry,dec}})\Big]. \end{aligned} \tag{7}$$

The first bracket measures the support change on the visual states selected at the queried-object state. The second is the nonnegative gain from changing the Top-K set at the queried-decision state, since that set maximizes the mean affinity over size-K visual-state sets. Taking hallucinated-minus-true population differences gives

$$\Gamma_\ell = \Gamma_\ell^{\mathrm{fix}} + \Gamma_\ell^{\mathrm{switch}}. \tag{8}$$

Thus, the state-role component can reflect both a change in affinity to the visual states selected at the object state and a change in which visual states enter the Top-K set at the decision state. The final-layer switching component is positive and substantially larger for hallucinated examples, while the detailed analysis and bootstrap intervals are reported in Sec. B.3.

Independently of this decomposition, Eq. (5) gives

$$\Delta_\ell^{\mathrm{qry,dec}} = \Delta_\ell^{\mathrm{qry,obj}} + \Gamma_\ell. \tag{9}$$

Therefore, an under-supported object state ($\Delta_\ell^{\mathrm{qry,obj}} < 0$) and an over-supported decision state ($\Delta_\ell^{\mathrm{qry,dec}} > 0$) can coexist when

$$\Gamma_\ell > -\Delta_\ell^{\mathrm{qry,obj}}. \tag{10}$$

The decomposition in Eqs. (7) and (8) is descriptive and does not establish a causal pathway from the object state to the decision.

4 OBJECT REPRESENTATION AND DECISION FORMATION

We first test whether the apparent difference between free-form and queried hallucination persists when visual support is measured at comparable object-representation states. We then examine how the support pattern changes when visual support is measured at the corresponding decision state.

4.1 EXPERIMENTAL PROTOCOL

Qwen2.5-VL-3B-Instruct Bai et al. (2025) is used for the main analysis, while LLaVA-1.5-7B Liu et al. (2023) is used to evaluate transfer of the free-form object-state result. Free-form hallucination is evaluated on AMBER-G Wang et al. (2024) and on COCO using CHAIR-style object labels Lin et al. (2014); Rohrbach et al. (2018). Queried object existence is evaluated on RePOPE Neuhaus & Hein (2025) and Causal-HalBench Xu et al. (2026).

For the analyses on RePOPE and Causal-HalBench, we restrict attention to responses predicted as Yes. False positives are labeled hallucinated, while true positives are labeled true. The same-image, same-object decomposition instead follows the free-form labeled image-object pairs through the corresponding existence query and is not restricted to Yes predictions. For free-form generation, we map extracted objects to the corresponding COCO or AMBER object vocabulary and retain the first occurrence of each unique (image,object) pair. Complete preprocessing details are presented in Sec. B.1.

Throughout the experiments, we use $K = 8$. For the main comparisons, we use bootstrap resampling at the image level to quantify uncertainty in the observed differences. The midpoint layer is selected on Qwen with AMBER-G and then evaluated on LLaVA with COCO, Qwen with COCO, and LLaVA with AMBER-G. Queried state comparisons use prespecified midpoint and final depths.

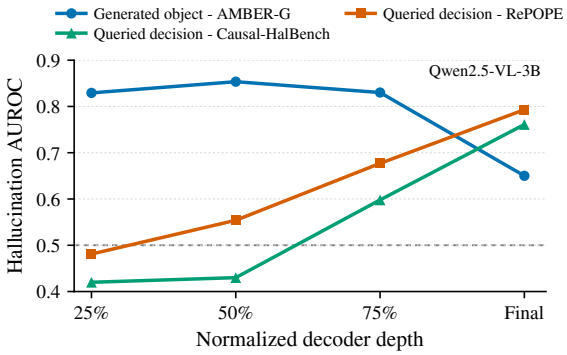

Figure 2: Depth-dependent visual support across computational states. Generated-object hallucinations are most separable around the midpoint using *low* support, while queried false-positive *decisions* are increasingly separable toward the final layer using *high* support. This apparent difference is resolved by distinguishing object-representation and decision states (Fig. 1 and Tab. 1).

Across architectures, decoder depth is aligned by normalized layer index, and object states follow each model's native tokenization. Visual support is computed using the visual states produced by each model. As hidden-state geometries can differ across architectures, we compare within-model differences rather than absolute cosine magnitudes across models. Details of state extraction and layer selection are presented in Secs. A.2 and B.1.

4.2 Visual support of object representations

In free-form generation, hallucinated generated-object states receive substantially less visual support than true objects around the decoder midpoint (Fig. 2). On Qwen with AMBER-G, low midpoint support reaches AUROC approximately 0.854, while low final support is weaker. We then evaluate the selected midpoint on three additional model-dataset settings without further layer selection. Across these settings, low midpoint support reaches AUROC 0.827-0.897 (Tab. 2). For example, on LLaVA with AMBER-G, mean midpoint support is 0.499 for true objects and 0.345 for hallucinated objects.

The same midpoint pattern also appears within individual images. Among images containing both true and hallucinated objects, hallucinated midpoint support is lower in 89.0% of LLaVA/COCO images, 91.2% of Qwen/COCO images, and 89.9% of LLaVA/AMBER-G images, indicating that this overall difference between hallucinated and true objects cannot be explained only by differences across images.

Queried objects show the same midpoint pattern (Fig. 2). When visual support is measured at the queried object state, false positives are under-supported around the midpoint, with low-support AUROC approximately 0.720 on RePOPE and 0.801 on Causal-HalBench (Tab. 3). Thus, object-state under-support appears in both free-form generation and explicit querying despite the different origin of the object concept.

4.3 Visual support at the decision state

At the state corresponding to the queried decision, support at the midpoint has limited discriminative power, while high support at the final layer distinguishes false positives from true positives on both RePOPE and Causal-HalBench (Fig. 2 and Tab. 3). For the same query, the object state is under-supported around the midpoint, while the decision state shows relative visual over-support late in the decoder.

4.4 Same-image, same-object decomposition

We next follow the same image-object pairs from Qwen on AMBER-G across free-form generation and explicit querying. Using the queried object state instead of the queried decision state preserves the object-representation role and reveals a different view of the cross-mode difference. As shown in Tab. 1, generated and queried object states are both under-supported at the midpoint and remain mildly under-supported at the final layer, while the queried decision gap becomes positive.

At the final layer, the state-role component constitutes about 77% of the cross-mode difference, while the concept-origin component is much smaller (Tab. 1). The concept-origin component is positive, with a 95% confidence interval (CI) of approximately $[+0.0014, +0.0269]$. The final-layer values also meet the condition in Eq. (10): The state-role component is large enough for the negative queried-object gap to become a positive queried-decision gap. At the midpoint, the concept-origin

Table 1: Same-image, same-object state decomposition. Negative Δ indicates lower visual support for hallucinated than true objects. The cross-mode difference C is decomposed by Eq. (5) into the concept-origin component Ω and state-role component Γ.

MEASURE	MIDPOINT	FINAL
Generated object $\Delta^{\mathrm{gen,obj}}$	-0.1013	-0.0431
Queried object $\Delta^{\mathrm{qry,obj}}$	-0.1263	-0.0289
Queried decision $\Delta^{\mathrm{qry,dec}}$	$+0.0009$	$+0.0195$
Concept origin Ω	-0.0250	$+0.0142$
State role Γ	$+0.1272$	$+0.0484$
Cross-mode $C=\Omega+\Gamma$	$+0.1022$	$+0.0626$

component is smaller in magnitude than the state-role component and has the opposite sign, with a 95% CI of approximately $[-0.0318,-0.0189]$. Differences associated with concept origin therefore remain, but constitute a substantially smaller part of the cross-mode difference.

Restricting the same-image population to queried responses predicted as Yes preserves the same final-layer sign pattern. The generated-object and queried-object gaps remain negative (-0.0202 and -0.0198), while the queried-decision gap is positive ($+0.0225$). The resulting concept-origin component is near zero, while the state-role component remains positive at $+0.0423$. Complete bootstrap intervals are reported in Sec. B.2. The exact support-change decomposition also shows that the final-layer state-role component contains a positive Top-K switching component, with the detailed depth analysis reported in Sec. B.3. These analyses identify where the statistical signature changes without establishing a causal pathway for hallucination. The late over-support behavior is *relative*, in the sense that it captures the difference in support between hallucinated and true objects. It does not imply that absolute support increases monotonically for every hallucinated example.

5 Detection from state dynamics

The state-dependent support patterns identified in the preceding section also motivate task-specific hallucination scores. We evaluate these scores to determine whether the identified state-dependent signatures remain informative under model and benchmark transfer.

5.1 Evaluation protocol

Free-form evaluation follows the populations and midpoint selection in Sec. 4.1. For queried detection, all methods operate on the same population of Yes predictions. Detection performance is measured by area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). RePOPE is divided based on image identity into a detector-training set and a test set. Learned parameters are fit only on the detector-training set, and the test set contains 2,428 responses predicted as Yes, including 160 hallucinations. To evaluate benchmark transfer, we use Causal-HalBench, which contains 3,580 responses predicted as Yes, including 257 hallucinations. This benchmark is not used for fitting, standardization, layer selection, or threshold selection. The corresponding random-ranking AUPRCs are 0.066 on RePOPE test and 0.072 on Causal-HalBench. Details of the evaluation setting and implementation are presented in Secs. B.1 and B.8.

5.2 Free-form hallucination detection

For a generated object, the training-free MCVS score is

$$R_{\mathrm{MCVS}}(o) = -S^{\mathrm{gen,obj}}_{\ell_{\mathrm{mid}}}(o). \tag{11}$$

For comparison, the final-layer baseline uses the analogous low-support score $-S^{\mathrm{gen,obj}}_{\ell_{\mathrm{final}}}(o)$. No detector weights, target-set normalization, layer search, or threshold calibration are required. Table 2 reports the transfer results obtained using the selected midpoint across the three additional model-dataset settings without further layer selection.

Table 2: Free-form transfer of the midpoint object-state score.

MODEL	DATASET	MCVS		FINAL SUPPORT	
		AUROC	AUPRC	AUROC	AUPRC
LLaVA	COCO	**0.827**	**0.421**	0.727	0.277
Qwen	COCO	**0.897**	**0.481**	0.693	0.188
LLaVA	AMBER-G	**0.894**	**0.547**	0.777	0.198

Table 3: Training-free queried object-existence detection. Bold entries among the state-based support scores indicate the best result in each column.

TRAINING-FREE SCORE	REPOPE TEST		CAUSAL TRANSFER	
	AUROC	AUPRC	AUROC	AUPRC
Confidence-based	0.883	0.400	0.877	0.313
Entropy-based	0.885	0.388	0.877	0.305
Object-mid	0.721	0.164	0.801	0.218
Decision-final	**0.817**	**0.254**	0.760	0.165
Final transition	0.750	0.223	0.715	0.137
StateHall-U	0.807	0.240	**0.836**	**0.262**

Using identical LLaVA outputs on COCO, MCVS obtains higher AUROC and AUPRC than PAS and SVAR. In comparison to PAS, the AUPRC improvement is statistically significant, while the AUROC difference is not. Similarly, on Qwen with COCO, MCVS achieves higher AUROC and AUPRC than PAS and SVAR. On both Qwen/COCO and LLaVA/AMBER-G, the improvements over PAS and SVAR are statistically significant in both metrics under paired bootstrap resampling at the image level. Details are reported in Sec. B.5. All comparisons use identical generated outputs and object pairs. Runtime and memory usage are presented in Sec. B.10.

5.3 Queried object-existence detection

As shown in Fig. 2, false positives are under-supported at the midpoint object state and over-supported at the final decision state. We therefore define the training-free StateHall-U score

$$R_{\mathrm{U}}(o) = S^{\mathrm{qry,dec}}_{\mathrm{final}}(o) - S^{\mathrm{qry,obj}}_{\mathrm{mid}}(o). \tag{12}$$

This uses the same prespecified midpoint and final depths as the preceding analysis and requires no fitted weights or target-set calibration. We compare it with support-based alternatives and two uncertainty references (Tab. 3). The confidence-based score uses the negative first-step Yes-minus-No logit margin, while the entropy-based score uses the entropy of the first answer-token distribution. The support-based alternatives are Object-mid, $-S^{\mathrm{qry,obj}}_{\mathrm{mid}}$, Decision-final, $S^{\mathrm{qry,dec}}_{\mathrm{final}}$, and Final transition, $S^{\mathrm{qry,dec}}_{\mathrm{final}} - S^{\mathrm{qry,obj}}_{\mathrm{final}}$.

On Causal-HalBench, StateHall-U reaches AUROC 0.836 and AUPRC 0.262, exceeding the object-mid, decision-final, and final-transition scores. Paired bootstrap resampling shows that all three AUROC improvements are statistically significant. Thus, the cross-depth state score transfers better than either state-based component alone. Its advantage over the final-transition score further indicates that retaining the midpoint object state is useful, while the confidence-based score remains stronger for object-existence queries.

5.4 Learned detection from state features

We next test whether the two state variables identified above support effective learned detection. Based on these variables, StateHall-L represents each object using

$$\boldsymbol{x}_{\mathrm{L}}(o) = \left[-S^{\mathrm{qry,obj}}_{\mathrm{mid}}(o), S^{\mathrm{qry,dec}}_{\mathrm{final}}(o) \right]. \tag{13}$$

We standardize these features and fit a logistic regression model using only the RePOPE detector-training set. Both fitted coefficients are positive, matching the support directions identified in the

Table 4: Queried false-positive detection among responses predicted as Yes. Learned detectors are fit only on the RePOPE detector-training set and applied to Causal-HalBench without refitting or recalibration. Bold entries indicate the best result in each column.

METHOD	TRAINING	REPOPE TEST		CAUSAL TRANSFER	
		AUROC	AUPRC	AUROC	AUPRC
Confidence-based	No	0.883	0.400	0.877	0.313
Entropy-based	No	0.885	0.388	0.877	0.305
STATEHALL-U	No	0.807	0.240	0.836	0.262
RepProbing	Yes	0.852	0.411	0.588	0.092
HALP-QT	Yes	**0.886**	**0.470**	0.769	0.152
DHCP-adapted	Yes	0.726	0.152	0.581	0.102
VIB-Probe	Yes	0.877	0.448	0.725	0.144
STATEHALL-L	Yes	0.835	0.274	0.845	0.270
STATEHALL-HYBRID	Yes	**0.886**	0.389	**0.884**	**0.333**

preceding section: Greater midpoint object under-support and greater final decision support are associated with a higher hallucination score. STATEHALL-HYBRID additionally includes output confidence as a third feature.

We compare these detectors with RepProbing, HALP-QT Kogilathota et al. (2026), DHCP-adapted Zhang et al. (2025), and VIB-Probe Zhang et al. (2026) on the same population of Yes predictions (Tab. 4). The comparison evaluates the complete detector pipelines under the same data partition and transfer protocol. Details of the architectures and fitting procedures are presented in Sec. B.8.

On RePOPE, STATEHALL-L has lower AUROC than HALP-QT and VIB-Probe and shows no statistically significant difference from RepProbing. Under direct transfer to Causal-HalBench, STATEHALL-L reaches AUROC 0.845 and is higher than the four learned baseline implementations in Tab. 4. All four paired AUROC differences are statistically significant. We interpret this as a comparison of the evaluated detector pipelines rather than evidence that the state representation is intrinsically more informative than the baseline representations.

STATEHALL-HYBRID combines the two state features with output confidence. On RePOPE, its AUROC is not statistically different from that of the confidence-based score. On Causal-HalBench, it improves AUROC from 0.877 to 0.884 and reaches AUPRC 0.333. This result indicates that the state features provide complementary information to output confidence under this transfer setting. Complete comparisons and fitting details are presented in Sec. B.8.

6 CONCLUSION

We presented a state-aware framework for characterizing visual support in VLM object hallucination. Our main finding is that the interpretation of an internal hallucination signature depends on the computational role of the measured state. When generated and queried objects are compared at states serving the same object-representation role, hallucinated objects are visually under-supported around intermediate layers in both settings. Relative over-support appears only late in the decoder at the queried decision state. A same-image, same-object decomposition shows that the state-role component accounts for most of the cross-mode difference, while the concept-origin component is substantially smaller.

These state-dependent patterns are also useful for hallucination detection. MCVS uses midpoint object under-support for free-form generation, while STATEHALL-U combines midpoint object under-support with final decision support for queried object existence. MCVS transfers across models and datasets, while the queried state-based detectors transfer across benchmarks and provide information complementary to output confidence. Output confidence remains effective for object-existence queries. Internal hallucination signatures should therefore be interpreted with respect to the computational state being measured. Additional discussion and limitations are presented in Sec. C. Code will be made available upon acceptance.

AI USE STATEMENT

In this work, we used generative AI tools to assist with draft code generation, the revision and improvement of the text, and the verification of mathematical derivations. AI-generated code was verified and revised by the authors, and all mathematical derivations were independently checked by the authors. AI-assisted text was also reviewed and revised by the authors to ensure its accuracy and consistency with the work. We take full responsibility for the final content of this work, including all text, claims, mathematical derivations, and code.

REFERENCES

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. *arXiv:2502.13923*, 2025.

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 *ICCV*, 2025.

Nhat Hoang, Minh Vu, My T. Thai, and Manish Bhattarai. PAS: Prelim attention score for detecting object hallucinations in large vision-language models. In *CVPR*, 2026.

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 *CVPR*, 2024.

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 *CVPR*, 2025.

Prannay Kaul, Zhizhong Li, Hao Yang, Yonatan Dukler, Ashwin Swaminathan, C. J. Taylor, and Stefano Soatto. THRONE: An object-based hallucination benchmark for the free-form generations of large vision-language models. In *CVPR*, 2024.

Sai Akhil Kogilathota, Sripadha Vallabha E G, Luzhe Sun, and Jiawei Zhou. HALP: Detecting hallucinations in vision-language models without generating a single token. In *EACL*, 2026.

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 *CVPR*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, 2023.

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 *ECCV*, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.

Yannic Neuhaus and Matthias Hein. RePOPE: Impact of annotation errors on the pope benchmark. *arXiv:2504.15707*, 2025.

OpenBMB. Minicpm-v 4.6. OpenBMB MiniCPM-V model repository, 2026.

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 *NAACL*, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, 2018.

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:2311.07397*, 2024.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate LVLM object hallucination. In *CVPR*, 2026.

Zhe Xu, Zhicai Wang, Junkang Wu, Jinda Lu, and Xiang Wang. Causal-HalBench: Uncovering LVLMs object hallucinations through causal intervention. In *AAAI*, 2026.

Feiran Zhang, Yixin Wu, Zhenghua Wang, Xiaohua Wang, Changze Lv, Xuanjing Huang, and Xiaoqing Zheng. VIB-probe: Detecting and mitigating hallucinations in vision-language models via variational information bottleneck. In *ACL*, 2026.

Yudong Zhang, Ruobing Xie, Xingwu Sun, Yiqing Huang, Jiansheng Chen, Zhanhui Kang, Di Wang, and Yu Wang. DHCP: Detecting hallucinations by cross-modal attention pattern in large vision-language models. In *ACM MM*, 2025.

A Support-change decomposition and state extraction

A.1 Exact object-to-decision support-change decomposition

Fix a layer ℓ and suppress the layer index. Normalize each visual state as $\boldsymbol{u}_j = \boldsymbol{v}_j / \|\boldsymbol{v}_j\|_2$. For a fixed index set $\mathbb{A}$ of size K, define

$$S_{\mathbb{A}}(\boldsymbol{h}) = \frac{1}{K} \sum_{j \in \mathbb{A}} \frac{\boldsymbol{h}^\top \boldsymbol{u}_j}{\|\boldsymbol{h}\|_2}. \tag{14}$$

The support in Eq. (2) can equivalently be written as

$$S(\boldsymbol{h}) = \max_{|\mathbb{A}|=K} S_{\mathbb{A}}(\boldsymbol{h}), \tag{15}$$

because the Top-K affinities maximize the sum over size-K index sets.

Proposition 1. *Let $\boldsymbol{h}^{\mathrm{qry,obj}} \neq 0$ and $\boldsymbol{h}^{\mathrm{qry,dec}} \neq 0$, and let $\mathbb{A}^{\mathrm{qry,obj}} = \mathbb{T}_K(\boldsymbol{h}^{\mathrm{qry,obj}})$. Then*

$$\begin{aligned} &S(\boldsymbol{h}^{\mathrm{qry,dec}}) - S(\boldsymbol{h}^{\mathrm{qry,obj}}) \\ &\quad = \big[S_{\mathbb{A}^{\mathrm{qry,obj}}}(\boldsymbol{h}^{\mathrm{qry,dec}}) - S_{\mathbb{A}^{\mathrm{qry,obj}}}(\boldsymbol{h}^{\mathrm{qry,obj}})\big] + \big[S(\boldsymbol{h}^{\mathrm{qry,dec}}) - S_{\mathbb{A}^{\mathrm{qry,obj}}}(\boldsymbol{h}^{\mathrm{qry,dec}})\big], \end{aligned} \tag{16}$$

and the second term is nonnegative.

Proof. Since $\mathbb{A}^{\mathrm{qry,obj}}$ is the Top-K set at the object state, $S(\boldsymbol{h}^{\mathrm{qry,obj}}) = S_{\mathbb{A}^{\mathrm{qry,obj}}}(\boldsymbol{h}^{\mathrm{qry,obj}})$. Adding and subtracting $S_{\mathbb{A}^{\mathrm{qry,obj}}}(\boldsymbol{h}^{\mathrm{qry,dec}})$ gives Eq. (16). By Eq. (15), $S(\boldsymbol{h}^{\mathrm{qry,dec}}) \geq S_{\mathbb{A}^{\mathrm{qry,obj}}}(\boldsymbol{h}^{\mathrm{qry,dec}})$, which proves nonnegativity. □

For each example x, let $\delta_\ell^{\mathrm{fix}}(x)$ and $g_\ell^{\mathrm{switch}}(x)$ denote the first and second terms in Eq. (7). Taking hallucinated-minus-true differences gives

$$\Gamma_\ell^{\mathrm{fix}} = \mathbb{E}[\delta_\ell^{\mathrm{fix}} \mid H] - \mathbb{E}[\delta_\ell^{\mathrm{fix}} \mid T], \tag{17}$$

$$\Gamma_\ell^{\mathrm{switch}} = \mathbb{E}[g_\ell^{\mathrm{switch}} \mid H] - \mathbb{E}[g_\ell^{\mathrm{switch}} \mid T], \tag{18}$$

with $\Gamma_\ell = \Gamma_\ell^{\mathrm{fix}} + \Gamma_\ell^{\mathrm{switch}}$ exactly. Although $g_\ell^{\mathrm{switch}}(x) \geq 0$ for every example, its hallucinated-minus-true population difference can have either sign.

A.2 State and visual-support extraction

Block outputs. We use post-block hidden states on both the visual and text sides. State definitions and multi-token aggregation follow Sec. 3. Visual support is computed directly from these states using Eq. (2), with $K = 8$ and no learned projection. The resulting statistic is a hidden-state alignment measure, not attention mass.

Table 5: Zero-based decoder block indices. Free-form detection uses only the midpoint. Queried-state comparisons use midpoint object support and final decision support. The 25% and 75% Qwen depths are descriptive only.

MODEL	L	25%	50%	75%	FINAL
Qwen2.5-VL-3B-Instruct	36	9	18	26	35
LLaVA-1.5-7B	32	–	16	–	31

Depth and layer indices. We use the normalized depth defined in Sec. 3. Table 5 gives the exact zero-based indices used for the midpoint and final states. The 25%, 50%, 75%, and final depth profile used for descriptive queried-depth characterization is also shown for Qwen.

Free-form sequence reuse and tokenization. For free-form analysis, saved captions are reused rather than regenerated for support extraction. We teacher-force the saved sequence and map the original object character or token span to the corresponding subtokens under the model's native tokenizer.

For LLaVA/AMBER-G, exact generated token IDs are retained and reused by the PAS and SVAR evaluations. For Qwen/COCO and LLaVA/COCO, saved caption strings are canonically retokenized with the corresponding tokenizer, with original character spans remapped to subtokens. The same sequence handling is used for MCVS and the PAS and SVAR comparisons on identical outputs and object pairs. We verified that canonical retokenization produces nearly identical support scores and AUROC.

Queried-object and decision extraction. For queried object existence, the benchmark question is followed by the instruction `Answer only Yes or No.`. Qwen decoding is greedy with at most four new tokens. Analyses on RePOPE and Causal-HalBench are restricted to examples whose recorded answer is `Yes`. False-positive Yes responses are labeled hallucinated, while true-positive Yes responses are labeled true. The same-image, same-object decomposition instead follows the free-form labeled image-object pairs through the corresponding existence query and is not restricted to `Yes`responses.

The queried-object state is obtained from the query subtokens corresponding to the object concept. The queried-decision state is the last prompt-position hidden state from the prompt-prefill computation, immediately before generation of the first answer token.

State gaps and decomposition. The hallucination support gaps and the exact decomposition into concept-origin and state-role components are defined in Eqs. (3) and (5). All terms are measured with the same support measure. In the same-image, same-object comparison, image and object identity are fixed across the generated and queried modes.

The exact support-change decomposition in Sec. A.1 characterizes the same-layer role comparison $\boldsymbol{h}_\ell^{\mathrm{qry,obj}} \to \boldsymbol{h}_\ell^{\mathrm{qry,dec}}$, while the cross-depth construction of StateHall-U follows the measured depth profile.

The state-aware analysis procedure is summarized in Algorithm 1.

B Experimental details and additional results

B.1 Datasets, populations, and evaluation protocol

Free-form object populations. COCO generations are evaluated with CHAIR-style canonical object labels Lin et al. (2014); Rohrbach et al. (2018). The primary unit is the first occurrence of each unique $(\mathrm{image,canonical\ object})$ pair, preventing repeated mentions of the same object from dominating the result. The Qwen/COCO evaluation population contains 500 generated captions, 672 image-object pairs after first-occurrence filtering, 43 hallucinated pairs, and 437 contributing images. The LLaVA/COCO evaluation contains 1,098 pairs after first-occurrence filtering, with 116 hallucinations. The generation prompt is `Please describe this image in detail.`

AMBER-G Wang et al. (2024) uses each released generative query verbatim. Object labeling follows the fixed AMBER object-level pipeline, with noun extraction, restriction to the AMBER object

Algorithm 1 State-aware visual support analysis

Require: Labeled collection $\mathcal{D}$ of image-object examples with saved generated sequences and/or existence queries, layer set $\mathbb{L}$, and K

Ensure: State-specific support trajectories, population gaps, and, when available, decomposition terms

for each example (I,o) in $\mathcal{D}$ **do**
  if a saved generated sequence is available **then**
    Teacher-force the sequence and cache visual states $\boldsymbol{v}_{\ell j}$ for $\ell \in \mathbb{L}$
    Extract $\boldsymbol{h}_{\ell}^{\mathrm{gen,obj}}$ for each $\ell \in \mathbb{L}$
  end if
  if an existence query is available **then**
    Run the query and cache visual states $\boldsymbol{v}_{\ell j}$ for $\ell \in \mathbb{L}$
    Extract $\boldsymbol{h}_{\ell}^{\mathrm{qry,obj}}$ from the object span for each $\ell \in \mathbb{L}$
    Extract $\boldsymbol{h}_{\ell}^{\mathrm{qry,dec}}$ at the last prompt position before first answer-token generation for each $\ell \in \mathbb{L}$
  end if
  Apply Eq. (2) to every available state-layer pair
end for
Compute Δ_{ℓ}^{s} over $\mathcal{D}$ for each available state role $s \in \{\mathrm{gen,obj,qry,obj,qry,dec}\}$
if $\mathcal{D}$ contains same-image, same-object generated and queried states **then**
  Compute the concept-origin component Ω_{ℓ}
  Compute the state-role component Γ_{ℓ}
end if
return State-specific support trajectories, $\{\Delta_{\ell}^{s}\}$, and, when available, $\{\Omega_{\ell}, \Gamma_{\ell}\}$

vocabulary, reference and relation checks, and semantic matching to the image annotations. Unmatched vocabulary objects are labeled hallucinated. The LLaVA/AMBER run contains 1,004 queries, 4,854 scored object mentions with 252 hallucinated instances, and 2,830 pairs after first-occurrence filtering, with 214 hallucinations from 1,001 images. The Qwen/AMBER-G run also contains 1,004 queries. After the same filtering, the population used for midpoint selection and same-image, same-object comparisons contains 2,793 image-object pairs from 990 images, with 178 hallucinated and 2,615 true pairs. Among the hallucinated pairs, 97 are also falsely accepted by the queried existence decision.

Queried detector population and data partition. All queried detectors use the same fixed Qwen2.5-VL-3B-Instruct population of Yes predictions. RePOPE Neuhaus & Hein (2025) contains 3,289 responses predicted as Yes, of which 197 are hallucinated. The assignment is deterministic and based on image identity. An image is assigned to the detector-training set when

$$\mathrm{SHA256}(\texttt{image_key}) \bmod 4 = 0, \tag{19}$$

and to the test set otherwise. This gives 861 rows for detector training, with 37 hallucinations, and 2,428 test rows with 160 hallucinations. All fitting, standardization, and learned depth selection are restricted to the detector-training set.

Causal-HalBench Xu et al. (2026) is used only for direct transfer and contains 3,580 responses predicted as Yes with 257 hallucinations from 2,025 images. No Causal-HalBench label is used for standardization, feature selection, layer choice, hyperparameter selection, or threshold selection.

Selection protocol. Following the midpoint-selection protocol in Sec. 4.1, Qwen with AMBER-G shows a broad peak near the normalized midpoint, with AUROC approximately 0.829 at 25% depth, 0.854 at 50%, 0.830 at 75%, and 0.650 at the final layer. We use the midpoint rule $\ell_{\mathrm{mid}} = \lceil 0.5(L-1) \rceil$ and $K = 8$ throughout the subsequent transfer evaluations. No fitted fusion, target-set calibration, alternative layer selection, or K search is performed on those settings.

Queried state comparisons use the prespecified midpoint and final layers from the normalized-depth convention. The same-image, same-object comparisons change only which computational state is extracted, not the support measure or layer-selection rule.

Table 6: Final-layer same-image, same-object decomposition after restricting queried responses to `Yes`. Intervals are 95% image-level bootstrap intervals.

MEASURE	VALUE	95% CI
Generated object $\Delta^{\mathrm{gen,obj}}$	-0.0202	$[-0.0346,-0.0063]$
Queried object $\Delta^{\mathrm{qry,obj}}$	-0.0198	$[-0.0291,-0.0106]$
Queried decision $\Delta^{\mathrm{qry,dec}}$	$+0.0225$	$[+0.0158,+0.0292]$
Concept origin Ω	$+0.0004$	$[-0.0123,+0.0136]$
State role Γ	$+0.0423$	$[+0.0310,+0.0535]$
Cross-mode C	$+0.0427$	$[+0.0256,+0.0597]$

Table 7: Exact decomposition of the state-role component into fixed-object-set and Top-K switching terms. Brackets show 95% image-level bootstrap intervals.

POPULATION	DEPTH	Γ_{fix}	Γ_{switch}
All	Midpoint	$+0.1504$ $[+0.1288,+0.1705]$	-0.0232 $[-0.0337,-0.0125]$
All	Final	$+0.0140$ $[+0.0004,+0.0278]$	$+0.0344$ $[+0.0234,+0.0453]$
Queried Yes	Midpoint	$+0.0928$ $[+0.0683,+0.1165]$	-0.0005 $[-0.0122,+0.0119]$
Queried Yes	Final	$+0.0013$ $[-0.0138,+0.0157]$	$+0.0410$ $[+0.0271,+0.0545]$

B.2 Robustness of the state decomposition

Queried-Yes sensitivity. The primary same-image, same-object decomposition follows all 2,793 Qwen/AMBER-G image-object pairs regardless of the queried answer. To test whether correctly rejected `No`responses drive the result, we repeat the decomposition after restricting the queried responses to `Yes`. This retains 2,648 pairs, including 97 hallucinated and 2,551 true pairs. At the final layer, both object-state gaps remain negative, while the queried-decision gap remains positive (Tab. 6). The concept-origin component is near zero, while the state-role component remains clearly positive. Thus, the final state-role pattern does not depend on including hallucinated free-form objects that are subsequently rejected by the queried decision.

B.3 Analysis of Top-K switching

The exact decomposition in Eq. (7) does not require the Top-K set to remain unchanged between the queried-object and queried-decision states. We nevertheless measured endpoint set stability to determine whether set changes are common. At both the midpoint and final layer, none of the 2,793 examples has identical queried-object and queried-decision Top-8 sets. Mean endpoint Jaccard similarity is 0.0083 at the midpoint and 0.0062 at the final layer, with median zero at both depths. The selected visual states therefore change substantially between the two computational states.

Table 7 reports the corresponding population decomposition. At the final layer, $\Gamma_{\mathrm{switch}} = +0.0344$ with 95% CI $[+0.0234,+0.0453]$ on the full population. On the queried-Yes sensitivity population, $\Gamma_{\mathrm{switch}} = +0.0410$ with CI $[+0.0271,+0.0545]$, while Γ_{fix} is near zero. The switching component also increases from midpoint to final depth by $+0.0576$ $[+0.0414,+0.0736]$ on the full population and $+0.0415$ $[+0.0203,+0.0617]$ after restricting to queried `Yes`responses. These results characterize a depth-dependent change in how the Top-K visual states contribute to the state-role component. They do not imply that Top-K switching causes hallucination.

B.4 Multi-token queried-object aggregation

The main analysis averages the hidden states over the queried object subtokens before computing support. For multi-token objects, we additionally evaluate support at the first object subtoken and the mean of the support scores computed separately for each object subtoken. Single-token objects are identical under all three definitions. In the full same-image population, 129 of 2,793 objects contain multiple subtokens, including two hallucinated objects. We therefore evaluate robustness on the full population rather than drawing a separate hallucinated-versus-true conclusion from the small multi-token subset.

Table 8: Final-layer sensitivity to queried-object aggregation. Mean state is the primary definition. Mean token support averages support values computed separately for the object subtokens.

POPULATION	AGGREGATION	$\Delta^{\mathrm{qry,obj}}$	Γ
All	Mean state	-0.0289	$+0.0484$
All	First token	-0.0304	$+0.0499$
All	Mean token support	-0.0281	$+0.0476$
Queried Yes	Mean state	-0.0198	$+0.0423$
Queried Yes	First token	-0.0214	$+0.0439$
Queried Yes	Mean token support	-0.0190	$+0.0415$

At the final layer, the queried-object gap and state-role component remain stable across the three definitions (Tab. 8). All image-level bootstrap intervals preserve a negative queried-object gap and a positive state-role component. The same result holds after restricting queried responses to `Yes`.

B.5 Additional free-form results

Paired transfer comparisons. For LLaVA/COCO, the paired image-level bootstrap improvement of MCVS over final support is $+0.101$ AUROC with 95% CI $[0.066,0.138]$, and $+0.138$ AUPRC with CI $[0.061,0.213]$. For Qwen/COCO, the corresponding differences are $+0.204$ $[0.136,0.277]$ and $+0.293$ $[0.175,0.413]$. For LLaVA/AMBER-G, they are $+0.117$ $[0.091,0.143]$ and $+0.346$ $[0.282,0.410]$.

Same-image support comparison. The same-image direction rates are reported in Sec. 4. The corresponding mixed-class populations contain 91, 34, and 189 images for LLaVA/COCO, Qwen/COCO, and LLaVA/AMBER-G, with mean paired midpoint-support differences of -0.121, -0.113, and -0.157, respectively.

Baselines. PAS Hoang et al. (2026) and SVAR Jiang et al. (2025) are evaluated on the same generated sequences and the same object pairs as MCVS. PAS uses zero-based decoder layer 0 and the first object subtoken as the query position, and treats attention mass to preceding generated or preliminary text as hallucination risk, with higher values indicating greater risk. SVAR uses zero-based layers 5-18 and the first object subtoken as the query position, sums head-averaged visual-token attention over the selected layers, and treats lower visual attention as higher risk. On Qwen/COCO, PAS and SVAR are our reproductions. We retain the prespecified PAS query position and SVAR layer range without Qwen-specific layer selection, sign selection, calibration, or regeneration.

Additional baselines: EAZY and LLCC. EAZY Che et al. (2025) uses layer 15 in one-based paper notation (block 14 zero-based). We identify the Top-5 attended projected image tokens, zero those projected tokens, regenerate, and mark the object hallucinated if the target disappears. Since this produces a hard decision rather than a continuous score, its AUROC/AUPRC is not directly comparable to that of a continuous ranking score. On Qwen/COCO, our reproduction obtains $0.693/0.153$.

For LLCC Wang et al. (2026), we reproduce its semantic-consistency score on LLaVA-1.5-7B and Qwen2.5-VL-3B-Instruct. On LLaVA, the Top-3 attended image tokens are selected at decoder layers 20-27, their hidden states are decoded through the Logit Lens, and the resulting tokens are compared semantically with the generated object. For Qwen, we use the same procedure with zero-based layers 21 and 23-29. For threshold-free evaluation, hallucination risk is one minus the maximum semantic similarity. We additionally report object-token NLL, computed as the mean negative log-probability over the generated subtokens of the target object, as a separate likelihood baseline. Neither EAZY nor LLCC is used to tune MCVS. The LLCC results are reported in Tab. 9, while EAZY is reported separately above as it produces a hard decision rather than a continuous ranking score.

Additional comparison on Qwen/COCO. The Qwen/COCO population contains 672 first-occurrence image-object pairs from 437 images, including 43 hallucinated pairs. Table 9 shows that MCVS reaches 0.897 AUROC and 0.481 AUPRC, compared with $0.747/0.117$ for PAS and $0.769/0.231$ for SVAR. Under paired image-level bootstrap resampling with 5,000 replicates, the MCVS improvement over PAS is $+0.150$ AUROC with 95% CI $[+0.090, +0.209]$ and $+0.368$

Table 9: Free-form baseline comparisons on the same generated outputs and object pairs. Difference columns report mean paired image-level bootstrap differences for MCVS minus the comparator. Brackets show 95% bootstrap intervals. On LLaVA/COCO, the MCVS-PAS AUROC interval includes zero, so those two methods are not statistically separated in AUROC.

DATASET	METHOD	AUROC	AUPRC	AUROC DIFF.	AUPRC DIFF.
LLaVA/COCO	MCVS	0.827	0.421	–	–
	PAS	0.805	0.302	$+0.022$ [-0.016,0.058]	$+0.115$ [0.026,0.204]
	SVAR	0.752	0.231	$+0.075$ [0.037,0.112]	$+0.184$ [0.119,0.254]
	LLCC semantic	0.659	0.187	$+0.169$ [0.115,0.224]	$+0.235$ [0.158,0.312]
	Object NLL	0.650	0.166	$+0.178$ [0.118,0.238]	$+0.253$ [0.163,0.347]
Qwen/COCO	MCVS	0.897	0.481	–	–
	PAS	0.747	0.117	$+0.150$ [0.090,0.209]	$+0.368$ [0.232,0.501]
	SVAR	0.769	0.231	$+0.130$ [0.058,0.207]	$+0.251$ [0.083,0.410]
	LLCC semantic	0.479	0.060	$+0.417$ [0.316,0.514]	$+0.428$ [0.284,0.564]
	Object NLL	0.657	0.192	$+0.241$ [0.151,0.342]	$+0.281$ [0.103,0.457]
LLaVA/AMBER-G	MCVS	0.894	0.547	–	–
	Final Support	0.777	0.198	$+0.117$ [0.091,0.143]	$+0.346$ [0.282,0.410]
	PAS	0.749	0.222	$+0.144$ [0.114,0.174]	$+0.321$ [0.254,0.388]
	SVAR	0.732	0.177	$+0.162$ [0.134,0.189]	$+0.367$ [0.310,0.426]
	Object NLL	0.628	0.150	$+0.265$ [0.219,0.311]	$+0.394$ [0.311,0.475]

Table 10: Cross-architecture evaluation of the late queried decision-state support pattern on MiniCPM-V 4.6. Direction rate is the fraction of same-image comparisons where false-positive decisions have higher support than true-positive decisions.

BENCHMARK	FINAL-SUPPORT AUROC	DIRECTION RATE
RePOPE	0.623	79.3%
Causal-HalBench	0.676	83.8%

AUPRC with CI $[+0.232, +0.501]$. In comparison to SVAR, the corresponding improvements are $+0.130$ $[+0.058, +0.207]$ and $+0.251$ $[+0.083, +0.410]$. All four intervals exclude zero.

B.6 Cross-architecture results for the queried decision-state pattern

We evaluate the late queried-decision support pattern on MiniCPM-V 4.6 OpenBMB (2026) using the same decision-state definition and support direction. This experiment evaluates whether the qualitative pattern observed with Qwen2.5-VL transfers across architectures. The pattern is observed on both queried benchmarks, although discrimination is weaker than for Qwen (Tab. 10).

Although the AUROC values are lower than for Qwen, the same-image direction agrees with the state-role prediction in a majority of cases on both benchmarks. These results provide cross-architecture evidence for the late decision-state pattern.

B.7 Training-free queried detector

Training-free queried performance is summarized in Tab. 3. On RePOPE test, the paired AUROC differences for StateHall-U are -0.076 versus the confidence-based reference with 95% CI $[-0.126, -0.032]$, $+0.086$ versus object-mid with CI $[0.061, 0.113]$, -0.010 versus decision-final with CI $[-0.063, 0.041]$, $+0.057$ versus final-transition with CI $[0.002, 0.113]$, and -0.078 versus the entropy-based reference with CI $[-0.127, -0.034]$.

On Causal-HalBench, the corresponding differences are -0.041 versus the confidence-based reference with CI $[-0.059, -0.024]$, $+0.035$ versus object-mid with CI $[0.026, 0.044]$, $+0.076$ versus decision-final with CI $[0.050, 0.101]$, $+0.121$ versus final-transition with CI $[0.091, 0.151]$, and -0.041 versus the entropy-based reference with CI $[-0.058, -0.024]$.

B.8 LEARNED DETECTOR PROTOCOLS AND TRANSFER

STATEHALL-L and STATEHALL-HYBRID. STATEHALL-L uses the two-dimensional feature defined in Eq. (13). Each feature is standardized using the mean and standard deviation estimated on the RePOPE detector-training set, and logistic regression with inverse regularization strength set to 1 and no class weighting is fit on the same set.

STATEHALL-HYBRID additionally uses the negative first-step Yes-minus-No logit margin as a third feature,

$$\boldsymbol{x}_{\mathrm{hyb}}(o) = \left[-S_{\mathrm{mid}}^{\mathrm{qry,obj}}(o), S_{\mathrm{final}}^{\mathrm{qry,dec}}(o), -m_{\mathrm{Y-N}}(o) \right], \tag{20}$$

where $m_{\mathrm{Y-N}}(o)$ denotes the Yes-minus-No logit margin for the query about object o. The same standardization and fitting convention is used, with no Causal-HalBench data used to estimate the standardization statistics or fit the classifier.

RepProbing. A linear probe is fit on the final hidden representation associated with the generated `Yes`token. This generated-token representation is distinct from the STATEHALL queried-decision state, which is taken from the last prompt position before answer generation. The raw feature dimension is 2,048. Training uses only the RePOPE detector-training set.

HALP-QT. We implement a HALP-style query-token MLP Kogilathota et al. (2026). Its normalized extraction depth is selected by grouped cross-validation restricted to the RePOPE detector-training set. The raw queried hidden state is 2,048-dimensional. No Causal-HalBench information is used for depth or model selection.

DHCP-adapted. We use the released DHCP one-hidden-layer 128-unit detector structure Zhang et al. (2025). Qwen2.5-VL uses dynamic image resolution, while the released detector expects a fixed spatial attention representation. Each attention map is therefore spatially resampled to 12×12 before flattening. Across 36 layers and 16 heads, the raw input dimension is $36 \times 16 \times 12 \times 12 = 82{,}944$. DHCP requires eager attention extraction, which changes the first-token argmax for fewer than 0.1% of examples. To preserve an identical evaluation population across detectors, we retain the recorded Yes decision and teacher-force the first answer token during feature extraction. The unforced first-step logits are retained for the entropy-based score and the Yes-minus-No logit margin.

VIB-Probe. We reproduce the VIB-Probe design Zhang et al. (2026) using pre-projection outputs from all layers and heads, followed by a Gaussian information bottleneck. The raw representation has dimension $36 \times 16 \times 128 = 73{,}728$. The reported VIB-Probe results are from our implementation.

Paired detector comparisons. Overall detector performance is summarized in Tab. 4. The paired comparisons below use the corresponding unrounded score vectors.

On RePOPE test, the paired AUROC differences for STATEHALL-L are -0.017 versus RepProbing with 95% CI $[-0.067, +0.032]$, -0.052 versus HALP-QT with CI $[-0.096, -0.004]$, $+0.108$ versus DHCP with CI $[+0.032, +0.187]$, and -0.042 versus VIB-Probe with CI $[-0.0864, -0.0002]$.

Under direct transfer to Causal-HalBench, STATEHALL-L exceeds RepProbing by $+0.256$ $[+0.218, +0.294]$, HALP-QT by $+0.075$ $[+0.049, +0.100]$, DHCP by $+0.264$ $[+0.220, +0.312]$, and VIB-Probe by $+0.120$ $[+0.088, +0.152]$. STATEHALL-HYBRID improves over the confidence-based reference by $+0.007$ AUROC $[+0.004, +0.010]$ on Causal-HalBench, while the RePOPE gain $+0.003$ has CI $[-0.003, +0.009]$ and is not statistically significant.

B.9 STATISTICAL EVALUATION

Image-level bootstrap resampling. Object pairs from the same image are not independent, so primary confidence intervals resample images rather than individual object observations. Each replicate samples image IDs with replacement, includes all rows belonging to the sampled images, recomputes both methods on the same resample, and stores their paired AUROC or AUPRC difference. Reported difference values are the means of these paired bootstrap differences. Replicates containing only one class are skipped, and 95% intervals use the empirical 2.5/97.5 percentile quantiles. We use

Table 11: Post-generation runtime and memory measurements. Peak memory reports the increase in allocated GPU memory above the allocation immediately before each timed pass.

METHOD	MEAN TIME (SEC./IMAGE)	RELATIVE MEAN	PEAK MIB
Plain pass	0.0897	1.000	67.6
MCVS	0.1065	1.188	67.6
PAS	0.0902	1.006	67.6
SVAR	0.0927	1.034	67.6

5,000 replicates for free-form comparisons and 2,000 for learned-detector comparisons. Training-free queried intervals use the same paired image-level procedure.

Same-image support comparison. The free-form same-image statistic is different from the bootstrap. We first retain images containing at least one true and at least one hallucinated object. For each such image we compute the mean support over hallucinated objects minus the mean support over true objects, then report the mean image-level difference and the fraction of images for which the difference is negative. This statistic asks whether the population-level under-support finding is also visible after conditioning on image identity.

B.10 Runtime and memory overhead

Per-image runtime is measured on 50 saved LLaVA/AMBER-G sequences using a single NVIDIA RTX 6000 Ada GPU, excluding autoregressive caption generation. Results are summarized in Tab. 11.

The average runtime overhead relative to the plain teacher-forced pass is approximately $+16.88$ ms/image for MCVS, $+0.54$ ms/image for PAS, and $+3.08$ ms/image for SVAR. The corresponding 95% paired bootstrap intervals are $[+12.62,+21.44]$, $[-0.60,+1.67]$, and $[+2.23,+3.96]$ ms/image, respectively. With N_v visual tokens, the support reduction itself requires $O(N_v d)$ operations per text state and layer after the hidden states are available and requires only $O(N_v)$ temporary similarity scores.

C Limitations and future directions

Output confidence remains an informative reference for object-existence queries, while the queried state features provide complementary information under benchmark shift, as reflected by StateHall-Hybrid on Causal-HalBench. Future work should investigate whether this complementarity persists across a wider range of model families, datasets, and forms of distribution shift, and how state-based signals can be combined with output confidence more generally.

The empirical scope of this work is limited to object hallucination in free-form generation and queried object-existence decisions. We do not assume that attributes, counting, spatial relations, or general VQA exhibit the same state dynamics. Extending the framework to these settings requires identifying semantically corresponding representation and decision states and verifying that the visual-support measure remains meaningful for the target variable. The exact switching decomposition characterizes the measured support change between states but does not establish that changing the selected visual states causes a hallucinated decision. While the free-form result transfers across Qwen and LLaVA and the queried decision-state pattern is also observed on MiniCPM-V, the complete same-image, same-object decomposition is established only on Qwen2.5-VL.

Our primary focus is on identifying and characterizing state-dependent hallucination signatures rather than optimizing a fully supervised detector. The learned variants use simple models fitted to the identified state features. A natural direction for future work is to incorporate these state-aware signals into richer detectors trained on broader and more diverse hallucination data.