VISUAL OVERINTERPRETATION IN THE PREFILL REPRESENTATIONS OF VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language Models (VLMs) have advanced rapidly in recent years, but they still suffer from hallucinations. Prior studies have identified various causes of hallucination and developed mitigation methods, primarily focusing on the generation process. More recent work moves upstream, revealing that hallucination risk is already encoded in pre-generation representations and can be reduced through prefill-time interventions. Nevertheless, these studies characterize hallucination risk through latent token-level or representation-level signals, lacking semantic interpretability of what kinds of concepts are easily misrepresented and how such biases emerge during prefill. Our work investigates this problem from a new concept-level perspective. Specifically, we formulate a fine-grained plausibility-aware concept system that classifies risk concepts into semantically meaningful categories. This concept system includes plausible absent concepts (PACs) and implausible absent concepts (imPACs), with PACs further distinguished by local visual resemblance, global scene semantics, or both. Based on this, we introduce PAC^3, a manually annotated benchmark for controlled concept-level analysis. We further design a dual-contrast prefill analysis framework that examines PAC-imPAC preference at both behavioral and representational levels and traces its evolution across prefill layers. Experiments across multiple VLMs consistently show that PACs receive stronger support than matched imPACs at both the behavioral and representational levels. Importantly, this preference emerges in intermediate prefill layers. We term this phenomenon *visual overinterpretation*, providing a concept-level semantic interpretation of prefill hallucination risk. Experiments in open-ended generation further show that this phenomenon is associated with subsequent object hallucinations.

1 INTRODUCTION

Vision-language models (VLMs) have recently demonstrated remarkable performance across a wide range of tasks, including image captioning (Liu et al., 2024b) and visual question answering (Li et al., 2023a). However, they still face generating content that is inconsistent with the given images, which is called hallucination (Rohrbach et al., 2018; Li et al., 2023b).

Prior studies explore hallucination in various ways, including language priors (Leng et al., 2024), learned object co-occurrence biases (Dai et al., 2023), insufficient or misleading visual grounding (Geigle et al., 2024; Li et al., 2025), and error propagation during multimodal generation (Zhong et al., 2024; Shoby et al., 2026). These studies identify many causes and mitigation strategies for hallucination, but most learn it in terms of model outputs or the generation process. More recent work proposes that hallucination risk can be identified before generation. Kogilathota et al. (2026) demonstrates that hallucinations in generation can be predicted directly by pre-generation hidden representations. Zhang et al. (2026) further indicates that manipulating representations at this stage can reduce subsequent hallucinations. However, they detect potential risk through latent token-level or representation-level signals, and the semantic behavior of VLMs during prefill is less explored. In particular, it remains unclear whether models misrepresent certain input concepts during prefill and whether such representational biases are related to downstream hallucinations.

Dai et al. (2023) shows that VLMs tend to hallucinate objects that are absent from the image but plausible given learned object co-occurrence patterns. This observation raises a new question of whether

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.

biases toward plausible absent concepts already emerge during prefill. To investigate this question, we distinguish plausible absent concepts (**PACs**) from implausible absent concepts (**imPACs**) and construct matched PAC-imPAC pairs, enabling controlled analysis of how plausibility affects the representation of visually unsupported concepts. Beyond the global contextual associations primarily considered in prior work (Dai et al., 2023; Li et al., 2026), we further characterize PACs according to their source of plausibility, including local visual resemblance (Local-PAC), global scene semantics (Global-PAC), or both (Mixed-PAC). Together, these concepts form our *plausibility-aware concept system*, enabling concept-level interpretation of bias risk during prefill. Specifically, we investigate three research questions around the plausibility-aware concept system: *RQ1: Behavioral susceptibility.* Are VLMs more likely to falsely recognize PACs than matched imPACs? *RQ2: Prefill representation.* Does this preference already emerge during prefill, and how does it evolve across layers and models? *RQ3: Downstream hallucination.* Is the phenomenon observed during prefill associated with subsequent object hallucinations in open-ended generation?

To explore these questions, we construct **PAC**^{3} based on our plausibility-aware concept system, a new manually annotated and verified benchmark with 1,219 questions across 453 controlled instances. We then introduce a *dual-contrast analysis framework* to systematically examine PAC and imPAC preference at both the behavioral and representation levels, tracing this preference across prefill layers. We further extend this contrast analysis to open-ended generation by tracking the same PACs and imPACs examined during prefill, investigating whether the observed preference persists in subsequent generation. Given the limited number of hallucinated samples under this controlled setting, we further complement this analysis with a larger-scale retrospective analysis of naturally generated hallucinations. Together, these analyses provide a semantically interpretable view of how plausible but visually unsupported concepts are represented before generation and whether these representations relate to downstream hallucinations. Experiments show that PACs are significantly more prone to false recognition than matched imPACs. More importantly, the preferential support for PACs already emerges in intermediate layers, providing a semantically interpretable view of prefill hallucination risk at the concept level. We refer to this consistent behavioral and representational tendency as *visual overinterpretation*. This bias is further associated with subsequent object hallucinations in open-ended generation. Our main contributions are summarized as follows:

- To improve the semantic interpretability of latent prefill risk, we introduce a concept-level perspective through a new plausibility-aware concept system for fine-grained concept-level analysis in pre-generation.
- Based on this, we introduce **PAC**^{3}, a human-annotated benchmark for concept-level interpretable analysis of plausibility-induced representational biases.
- We propose a dual-contrast prefill analysis framework, further identify *visual overinterpretation* during prefill, and show that this phenomenon is associated with downstream object hallucinations.

2 Related Work

2.1 Multimodal Hallucination and Interpretation

Hallucination in VLMs has been attributed to language priors (Leng et al., 2024), object co-occurrence biases (Dai et al., 2023), insufficient visual grounding (Geigle et al., 2024; Li et al., 2025), and error propagation during generation (Zhong et al., 2024; Shoby et al., 2026). Recent work further distinguishes hallucination sources arising from visual evidence and contextual priors (Wang et al., 2026; Bukkapatnam, 2026). For example, Phukan et al. (2025) analyzes contextual embeddings across layers to identify hallucination-related features, while Wang et al. (2026) combines visual attention with Logit Lens to reveal different sources of object hallucination. Other work traces hallucinations to specific sources or propagation processes within VLMs (Bukkapatnam, 2026; Shoby et al., 2026). However, most investigate hallucination through model outputs or the generation process. Dai et al. (2023) shows that hallucinated objects can be plausible due to learned object co-occurrence patterns, motivating our consideration of plausible but absent concepts. However, its notion of plausibility primarily captures global contextual associations, whereas our PAC formulation further distinguishes local visual resemblance (Local-PAC), global scene semantics (Global-PAC), and their combination (Mixed-PAC). More recently, Li et al. (2026) further

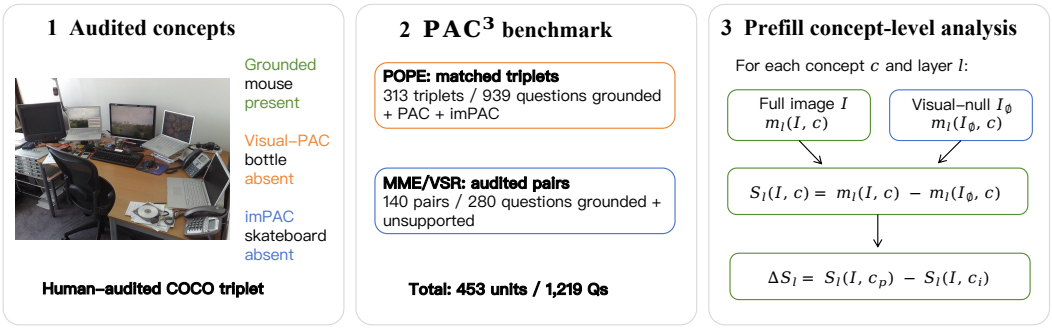

Figure 1: Overview of PAC^3 and our concept-level prefill analysis.

investigates plausible but absent objects through model generation behavior in generation. In contrast, our work uses plausibility as a concept-level semantic variable for interpreting prefill representations, with matched PACs and imPACs enabling us to examine whether equally absent concepts are represented differently depending on their plausibility.

2.2 Hallucination Risk in Prefill

Recent studies have moved beyond decoding analysis to investigate hallucination risk in internal representations during prefill. Phukan et al. (2025) shows that contextual embeddings from intermediate layers contain useful signals for hallucination detection and grounding. Moving further upstream, HALP (Kogilathota et al., 2026) learns probes over visual- and query-token hidden states to predict hallucinations before any output token is generated. PTI (Zhang et al., 2026) further derives object–background directions and modifies the prefill KV cache to mitigate subsequent hallucinations. Together, these studies show that hallucination-related information is already present during prefill and can support both detection and intervention. However, they primarily characterize such risk through latent token- or representation-level signals, leaving its concept-level semantic structure less explored. Our work aims to learn what visually unsupported concepts are preferentially represented during prefill and how such biases evolve across layers.

3 Method

3.1 Overview

Given an image I, we consider a plausible absent concept (PAC) c_p and a matched implausible absent concept (imPAC) c_n, both of which are visually unsupported. We define their relative support difference as

$$\Delta S_l(I) = S_l(I, c_p) - S_l(I, c_n), \tag{1}$$

where $S_l(I, c)$ denote the image-conditioned support for concept c at prefill layer l. Positive ΔS_l indicates preferential support for the PAC over the matched imPAC, while negative values indicate the reverse. Based on this formulation, we investigate three questions. *RQ1* asks whether VLMs systematically exhibit positive ΔS_l, indicating preferential support for PACs over matched imPACs; *RQ2* asks how this preference emerges and evolves across prefill layers, characterized by the trajectory $\Delta S_{l=1}^{L}$. *RQ3* asks whether the PAC–imPAC preference observed during prefill is associated with subsequent object hallucinations in open-ended generation. As shown in Figure 1, our analysis pipeline begins with a fine-grained plausibility-aware concept system that distinguishes PACs from imPACs and further categorizes PACs by local visual resemblance, global scene semantics, or both. Based on this concept system, we construct PAC^3 for controlled analysis. To address *RQ1* and *RQ2*, we introduce a dual-contrast prefill analysis framework that quantifies relative image-conditioned support and traces its evolution across layers using the behavioral and representational dual-contrast for PACs and imPACs. To address *RQ3*, we extend concept-support tracing to open-ended generation and examine its association with subsequent object hallucinations. We additionally evaluate broader concept types beyond object existence. The details of each component are as follows.

3.2 PLAUSIBILITY-AWARE CONCEPTS AND BENCHMARK

Plausibility-Aware Concepts. Prior work has shown that VLMs can hallucinate objects that are absent from the image yet remain plausible (Dai et al., 2023), suggesting plausibility as a meaningful dimension for characterizing hallucinated concepts. Motivated by this observation, we distinguish concepts according to their visual presence and plausibility. Given an image I, a *grounded concept* c_g is visually present in the image, whereas a *plausible absent concept* (PAC) c_p is visually absent but remains plausible given the image. In contrast, an *implausible absent concept* (imPAC) c_n is also visually absent but lacks such plausibility. By contrasting PACs with imPACs under the same absence condition, we examine whether VLMs represent visually unsupported concepts differently depending on their plausibility. To further characterize what makes a visually unsupported concept plausible, we distinguish the source of plausibility rather than treating all PACs as a single category. Specifically, we categorize PACs into three types. Local-PACs (c_p^l) arise from local visual resemblance, where image regions share visual characteristics such as appearance, shape, or color with the absent concept. Global-PACs (c_p^g) arise from global scene semantics, such as scene context or object co-occurrence, that make the absent concept plausible despite the lack of direct visual evidence. Mixed-PACs (c_p^{mix}) combine both local visual resemblance and global scene semantics. This fine-grained categorization enables us to interpret where PACs' plausibility comes from. Together, these concept types and their fine-grained categories constitute our *plausibility-aware concept system*.

PAC^3 Benchmark. To operationalize the above concept system, we construct PAC^3, a manually annotated and verified benchmark for controlled concept-level analysis. PAC^3 instantiates grounded concepts, PACs, and matched imPACs within the same image. Given an annotated PAC c_p, it needs to find the matched imPAC c_n from candidates $\mathcal{C}_n(I)$. We rank the candidates according to concept frequency and token length and retain the top-K candidates for manual review as follows:

$$\mathcal{C}_n^K(I, c_p) = \text{TopK}_{\tilde{c}_n \in \mathcal{C}_n(I)} \left[-D_{\text{match}}(c_p, \tilde{c}_n) \right], \tag{2}$$

$$D_{\text{match}}(c_p, \tilde{c}_n) = d_f(c_p, \tilde{c}_n) + d_\ell(c_p, \tilde{c}_n), \tag{3}$$

$$d_f(c_p, \tilde{c}_n) = \left[\log(1 + f(c_p)) - \log(1 + f(\tilde{c}_n)) \right]^2, \tag{4}$$

$$d_\ell(c_p, \tilde{c}_n) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} \left[\ell_m(c_p) - \ell_m(\tilde{c}_n) \right]^2. \tag{5}$$

Here, $D_{\text{match}}(c_p, \tilde{c}_n)$ denotes the matching distance between the PAC c_p and an imPAC candidate $\tilde{c}_n$, and $\mathcal{C}_n^K(I, c_p)$ denotes the top-K candidates with the smallest D_{match}. d_f and d_ℓ denote the frequency and token-length distances, respectively; $f(c)$ is the image frequency of concept c in COCO (Lin et al., 2014); $\ell_m(c)$ is its token length under model m; and $\mathcal{M}$ is the set of evaluated VLMs. Annotators then select the highest-ranked candidate satisfying the imPAC criteria as the final matched imPAC c_n. This procedure reduces potential confounds arising from concept frequency and model-specific tokenization while ensuring that the selected control satisfies the imPAC criteria.

PAC^3 also preserves the fine-grained PAC categories $c_p^l, c_p^g, c_p^{\text{mix}}$ defined above, enabling comparison of the relative strength of representation biases across different sources of plausibility. Beyond object existence, PAC^3 further includes broader concept types, such as attributes and relations, to examine whether the observed patterns extend to other forms of visually unsupported concepts. Detailed benchmark construction, annotation, and statistics are provided in Appendix A.

3.3 DUAL-CONTRAST PREFILL ANALYSIS FRAMEWORK

To examine whether VLMs preferentially represent plausible but visually unsupported concepts during prefill, we propose a *dual-contrast prefill analysis framework*. The framework consists of two levels of dual-contrast analysis: behavioral and representational.

Behavioral Contrast. At the behavioral level, we compare PACs with their matched imPACs under the same image condition. Since both concepts are visually absent, this contrast allows us to examine whether plausibility leads to additional false recognition beyond general errors on unsupported concepts. Specifically, given an image I, we query the model about the presence of each concept and define the behavioral contrast as

$$\Delta B = P(\hat{y} = 1 \mid I, c_p) - P(\hat{y} = 1 \mid I, c_n), \tag{6}$$

where $\hat{y} = 1$ denotes that the model recognizes the queried concept as present. Since both PAC c_p and imPAC c_n are absent from the image, a positive ΔB indicates a stronger behavioral tendency to falsely recognize the PAC than its matched imPAC.

Representation Contrast. At the representation level, directly comparing PAC and imPAC representations may be confounded by the model's pre-existing preferences for certain concept semantics, making it difficult to attribute the observed differences to visual conditioning. We therefore first quantify how the visual input changes the model's support for each concept. Given an image I and a concept c, let $m_l(I, c)$ denote the concept support measured from the hidden representation at layer l, following the general practice of decoding layer-wise latent predictions (Belrose et al., 2023; Wang et al., 2026). We construct a visual-null counterpart $I_\emptyset$ by removing the visual input while keeping the textual prompt unchanged. The image-conditioned support for concept c at layer l is defined as

$$S_l(I, c) = m_l(I, c) - m_l(I_\emptyset, c). \tag{7}$$

This visual-null contrast reduces the contribution of concept-specific semantic preferences and captures the change in concept support induced by the visual input. We then compare the layer l image-conditioned support ΔS_l of a PAC c_p and its matched imPAC c_n under the same image.

3.4 Concept-Support Propagation Tracing

To examine whether the concept preference observed during prefill is associated with subsequent hallucination, we introduce *concept-support propagation tracing*. Specifically, we track the same predefined PAC–imPAC pairs from their prefill representations to open-ended generation. For each image I_i, we first obtain its layer l relative prefill support $\Delta S_l(I_i)$. We then perform K open-ended generations for each image and define the generation frequency of concept c as

$$R_i(c) = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}\Big[c \in G_i^{(k)}\Big], \tag{8}$$

where $G_i^{(k)}$ denotes the concepts mentioned in the k-th generated response and $\mathbb{I}[\cdot]$ is the indicator function. Accordingly, $R_i^p = R_i(c_{p,i})$ and $R_i^n = R_i(c_{n,i})$ denote the generation frequencies of the PAC and matched imPAC, respectively, and their relative generation preference is defined as $\Delta R_i = R_i^p - R_i^n$. Thus, $\Delta R_i > 0$ indicates a stronger generation preference for the PAC.

To trace the propagation of concept preference, we examine the layer-wise associations between $\Delta S_l(I_i)$ and both the relative generation preference ΔR_i and the PAC generation frequency R_i^p using Spearman correlation. The former evaluates whether the PAC–imPAC preference during prefill is associated with the corresponding preference in subsequent generation, while the latter directly relates the prefill support gap to the generation frequency of the PAC itself. We further obtain a layer-agnostic measure by averaging the support gap across layers:

$$\Delta S_i^{\text{avg}} = \frac{1}{L} \sum_{l=1}^{L} \Delta S_l(I_i), \tag{9}$$

where L denotes the number of analyzed layers. This provides an overall measure of prefill concept preference without relying on a particular layer.

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

PAC^3 contains 453 controlled instances and 1,219 questions. Its object-existence subset comprises 313 POPE-derived triplets (939 questions) (Li et al., 2023b), each pairing a grounded concept and a PAC with a frequency- and token-matched imPAC from the same image. We further include 140 MME/VSR-derived pairs (Fu et al., 2026; Liu et al., 2023) (280 questions) to extend the analysis beyond object existence to attributes and relations. For complementary evaluation of RQ3, we use CHAIR generations on 500 COCO images per model (Rohrbach et al., 2018; Lin et al., 2014) to examine naturally occurring object hallucinations. CHAIR is not part of PAC^3.

Table 1: Behavioral susceptibility on the 313 PAC^3 object-existence triplets. Accuracy, F1, GR, FPR, PI, POR, and confidence intervals are reported in percentage points. Confidence intervals and p-values use parent-image-clustered inference.

Model	Acc. ↑	F1 ↑	GR ↑	$\mathrm{FPR}_{\mathrm{PAC}}$ ↓	$\mathrm{FPR}_{\mathrm{imPAC}}$ ↓	PI ↓	POR ↑	95% CI (PI)	p
LLaVA-1.5-7B	88.8	82.6	79.9	11.8	1.6	10.2	89.8	[6.2, 14.4]	$< .001$
Qwen3-VL-8B	91.8	88.0	90.4	14.4	0.6	13.7	86.3	[9.7, 18.0]	$< .001$
DeepSeek-VL-7B	90.2	85.4	85.6	12.8	2.2	10.5	89.5	[5.7, 15.3]	$< .001$

4.1.2 Implementation Details

We evaluate LLaVA-1.5-7B (Liu et al., 2024a), Qwen3-VL-8B (Bai et al., 2025), and DeepSeek-VL-7B (Lu et al., 2024) using their native chat templates and image preprocessors. Binary recognition uses greedy decoding. At each language layer, we extract the Yes–No logit margin from the final prompt representation and compute the image-conditioned support in Eq. (5) using the model-specific visual-null construction. Based on the discovery split, we fix the primary layers at L20/32 for LLaVA, L22/36 for Qwen, and L18/30 for DeepSeek, and evaluate them on the image-disjoint confirmation split. For open-ended generation, we sample $K = 10$ responses per image with temperature 0.7 and top-$p = 0.9$, using ten random seeds shared across models.

4.1.3 Metrics

We evaluate the models using both standard recognition metrics and overinterpretation-specific metrics. **Accuracy** and **F1** measure overall recognition performance, while Grounded Recall (**GR**) measures the model's ability to correctly recognize visually supported concepts. Plausibility Overinterpretation Resistance (**POR**) measures the model's resistance to plausibility-induced false recognition: $\mathrm{POR} = 1 - \mathrm{PI}$, $\mathrm{PI} = \mathrm{FPR}_{\mathrm{PAC}} - \mathrm{FPR}_{\mathrm{imPAC}}$. **FPR** denotes the false-positive rate, and **PI** measures the excess false-positive rate associated with concept plausibility. We additionally report $\mathrm{FPR}_{\mathrm{PAC}}$, $\mathrm{FPR}_{\mathrm{imPAC}}$, and PI as diagnostic statistics. Prefill Overinterpretation Gap (**POG**) measures representation-level overinterpretation during prefill, corresponding to the paired support difference ΔS_l defined in Eq. 1. A larger positive POG indicates stronger relative prefill support for PACs over matched imPACs. Generation Overinterpretation Gap (**GOG**) measures overinterpretation in open-ended generation, corresponding to the relative generation frequency ΔR defined in Sec. 3.4. A larger positive GOG indicates a stronger tendency to generate PACs than matched imPACs. Overinterpretation Generation Alignment (**OGA**) measures the association between prefill and generation overinterpretation, quantified by Spearman's correlation between POG and GOG. A larger positive OGA indicates that stronger overinterpretation during prefill is more consistently associated with stronger overinterpretation during subsequent generation. For statistical significance, we report 95% confidence intervals (**CIs**) and p**-values** for the relevant differences and associations, with image-level resampling when multiple instances share the same parent image.

4.2 Results

4.2.1 RQ1: Behavioral Susceptibility to PACs

We first examine whether VLMs are more likely to falsely recognize plausible absent concepts than matched implausible absent concepts. As shown in Table 1, all three models exhibit substantially higher false-positive rates for PACs than for imPACs, resulting in positive PI across models. For example, Qwen3-VL-8B yields a PAC false-positive rate of 14.4%, compared with only 0.6% for matched imPACs, corresponding to a PI of 13.7 percentage points ($p < .001$). Similar differences are observed for LLaVA-1.5-7B and DeepSeek-VL-7B. Meanwhile, the models maintain relatively high Grounded Recall, indicating that the observed differences are not simply explained by a general tendency to reject or accept queried concepts. These results show that visual absence alone does not determine false recognition: among concepts that are both absent from the image, plausible concepts are systematically more susceptible to being falsely recognized.

The cross-model comparison further suggests that general multimodal capability may affect where PAC susceptibility emerges. Qwen3-VL-8B achieves the strongest overall recognition performance (91.8% accuracy and 88.0% F1) and later shows the smallest prefill POG, suggesting greater resis-

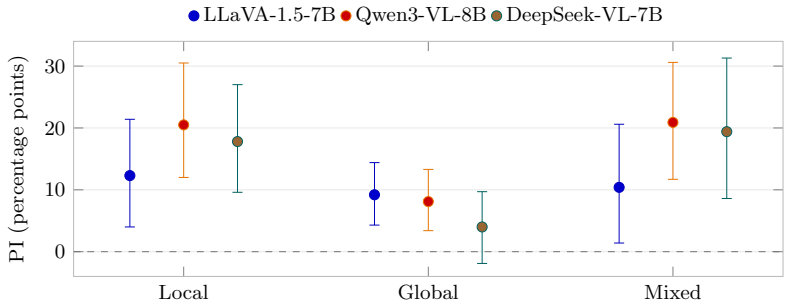

Figure 2: Behavioral PI across local, global, and mixed PACs. Error bars denote parent-image-clustered 95% confidence intervals.

tance to PAC interference in prefill representations. However, its behavioral PI remains substantial, indicating that stronger multimodal ability does not eliminate output-level PAC susceptibility.

We next examine whether this susceptibility varies with the source of plausibility. Figure 2 reports PI separately for Local-, Global-, and Mixed-PACs. Positive PI is observed across PAC categories, although its magnitude varies across both concept types and models. In particular, Qwen3-VL-8B and DeepSeek-VL-7B show pronounced effects for Local- and Mixed-PACs, whereas LLaVA-1.5-7B exhibits smaller differences across categories. For DeepSeek-VL-7B, for example, PI reaches 17.8 and 19.4 percentage points for Local- and Mixed-PACs, respectively, compared with 4.0 points for Global-PACs. The complete subtype statistics are provided in Table 4. Rather than suggesting a universal ordering among PAC types, these results indicate that behavioral susceptibility can arise from different sources of plausibility and that their relative effects are model dependent.

Finally, we examine whether the observed susceptibility extends beyond object existence. Appendix Table 5 reports results on the attribute and relation extensions of PAC^3. Overinterpretation-related errors remain observable in these settings, although their magnitude varies substantially across semantic tasks and models. For instance, LLaVA-1.5-7B shows an unsupported-concept false-positive rate of 11.8% for object existence, increasing to 22.5% for attributes and 57.1% for relations, whereas Qwen3-VL-8B exhibits a substantially lower relation error rate of 9.5%. These results suggest that susceptibility to visually unsupported semantics is not restricted to object existence, while also highlighting the greater model dependence of the broader semantic extensions.

4.2.2 RQ2: Visual Overinterpretation during Prefill

Having established the behavioral difference between PACs and matched imPACs, we next investigate whether this preference is already reflected in prefill representations. Figure 3 shows the layer-wise Prefill Overinterpretation Gap (POG) across the three models. In early layers, PACs and matched imPACs receive similar image-conditioned support, resulting in POG values close to zero. As representations progress through the network, however, POG increases substantially, with the strongest separation emerging in intermediate layers. For LLaVA-1.5-7B, for example, the gap increases rapidly after approximately half of the model depth and reaches its maximum around a relative depth of 0.6. Comparable intermediate-layer emergence is observed in Qwen3-VL-8B and DeepSeek-VL-7B, although the exact layer at which the effect is strongest differs across architectures.

The quantitative results at the discovery-selected primary layers are reported in Table 2. When evaluated on the image-disjoint confirmation split, the selected layers retain positive and statistically significant POG values. For example, LLaVA-1.5-7B obtains a POG of 10.31 at Layer 20, with a 95% confidence interval of $[7.02, 13.39]$ ($p < .001$). The persistence of the gap on held-out images indicates that the layer-wise pattern is not solely a consequence of selecting favorable layers on the discovery data.

Across models, Qwen3-VL-8B combines the highest Accuracy and F1 in Table 1 with the smallest POG, whereas LLaVA-1.5-7B and DeepSeek-VL-7B show larger representation-level gaps. This pattern is consistent with stronger multimodal representations providing greater resistance to PAC interference during prefill, although the comparison does not establish causality.

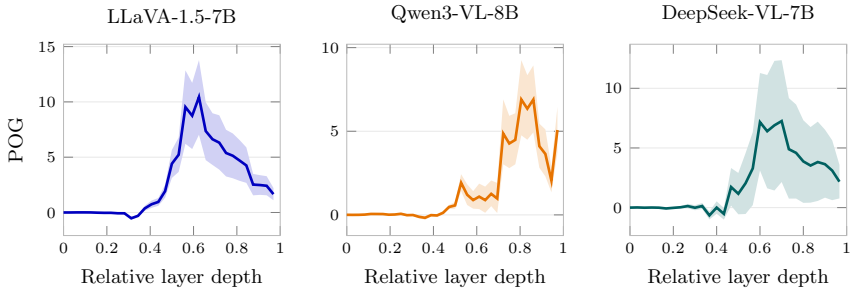

Figure 3: Layer-wise POG on the image-disjoint confirmation split. Shaded regions denote paired-bootstrap 95% confidence intervals.

Table 2: POG at the discovery-selected primary layer on the image-disjoint confirmation split. Confidence intervals are paired bootstrap intervals.

Model	Primary layer	POG ↓	95% CI	p
LLaVA-1.5-7B	20/32	10.310	[7.024, 13.391]	$< .001$
Qwen3-VL-8B	22/36	1.880	[1.146, 2.647]	$< .001$
DeepSeek-VL-7B	18/30	8.267	[4.521, 12.090]	$< .001$

We refer to this systematic representation-level preference for plausible but visually unsupported concepts as *visual overinterpretation*. Importantly, visual overinterpretation is not merely an output-level error: the PAC–imPAC distinction is already reflected in image-conditioned representations before generation begins. Additional subtype results in Table 6 show positive POG for Local-, Global-, and Mixed-PACs across the evaluated models, although the magnitude again varies across categories and architectures. Moreover, the support analysis in Table 7 generally places PACs between grounded and implausible absent concepts, with $S(c_g) > S(c_p) > S(c_n)$. For example, Qwen3-VL-8B yields support values of 10.12, 2.02, and 0.14 for grounded concepts, PACs, and imPACs, respectively. This pattern suggests that PACs acquire partial image-conditioned support despite lacking direct visual evidence.

Figure 4 provides a qualitative example in which PAC support separates from matched imPAC support in intermediate layers.

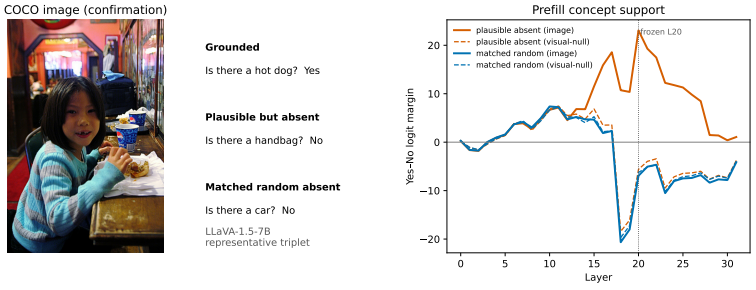

Figure 4: Representative PAC^3 triplet and its layer-wise prefill concept support. The PAC separates from the matched imPAC in intermediate layers.

4.2.3 RQ3: Visual Overinterpretation Association with Downstream Hallucination

We finally investigate whether visual overinterpretation observed during prefill is associated with subsequent hallucination behavior in open-ended generation. Table 3 first compares the generation frequencies of PACs and matched imPACs. Across the evaluated models, PACs are generated more frequently than their matched imPACs, resulting in positive Generation Overinterpretation Gaps (GOGs). For LLaVA-1.5-7B, for example, the GOG reaches 11.0 percentage points. The corresponding grounded-concept generation rates remain substantially higher, providing a reference for the models' ability to mention visually supported concepts.

Table 3: Open-ended generation and prefill–generation alignment. Generation rates, GOG, and its confidence interval are in percentage points. $\mathrm{GR}_{\mathrm{gen}}$ is the grounded-concept mention rate. OGA is the all-layer POG–GOG Spearman correlation.

Model	$\mathrm{GR}_{\mathrm{gen}} \uparrow$	$R^p \downarrow$	$R^n \downarrow$	GOG $\downarrow$	95% CI	OGA $\downarrow$	95% CI	p_{OGA}
LLaVA-1.5-7B	72.4	12.4	0.6	11.0	[8.4, 13.8]	0.52	[0.396, 0.638]	$< .001$
Qwen3-VL-8B	44.4	1.2	0.0	1.5	[0.5, 3.0]	0.17	[0.009, 0.303]	.043
DeepSeek-VL-7B	62.5	3.6	0.4	2.9	[1.6, 4.5]	0.41	[0.262, 0.533]	$< .001$

More importantly, POG and GOG are positively correlated across all three models, although the strength of the association differs substantially. LLaVA-1.5-7B exhibits the strongest Over-interpretation Generation Alignment (OGA), with $\rho = 0.52$ and a 95% confidence interval of $[0.40, 0.62]$ ($p < .001$). DeepSeek-VL-7B also shows a clear positive association ($\rho = 0.41$, 95% CI $[0.26, 0.53]$, $p < .001$), whereas Qwen3-VL-8B exhibits a considerably weaker association ($\rho = 0.17$, 95% CI $[0.01, 0.30]$, $p = 0.043$). Thus, stronger PAC–imPAC preference during prefill tends to coincide with stronger PAC–imPAC preference during subsequent generation, but the strength of this relationship is model dependent. Appendix Figure 5 provides the pair-level visualization.

Overall, the results across RQ1–RQ3 reveal a consistent progression from behavioral susceptibility to representation-level overinterpretation and its association with downstream generation. While this pattern is observed across the evaluated VLMs, its magnitude and downstream association remain model dependent. Importantly, these results establish an association rather than a causal relationship between prefill overinterpretation and subsequent hallucination.

Subtype results in Table 8 further show that generation frequency and prefill–generation alignment capture related but distinct aspects of visual overinterpretation. For example, in LLaVA-1.5-7B, Mixed-PACs exhibit the largest GOG (16.9 percentage points), whereas Local-PACs exhibit the strongest OGA ($\rho = 0.59$). This distinction suggests that a PAC category that is frequently generated is not necessarily the category for which instance-level prefill support is most strongly associated with generation behavior.

Complementary CHAIR results in Appendix Table 9 further show that contextual plausibility is informative for naturally occurring object hallucinations, supporting the downstream relevance of visual overinterpretation while revealing model-dependent association strength.

5 Conclusion

We investigate prefill hallucination risk from a new concept-level interpretability perspective. We formulate a plausibility-aware concept system of plausible absent concepts (PACs) and implausible absent concepts (imPACs), with PACs further categorized by local visual resemblance, global scene semantics, or both. Based on this concept system, we construct PAC^3 and develop a *dual-contrast prefill analysis framework*. Experiments across multiple VLMs reveal *visual overinterpretation*, with preferential support for PACs and emerging bias in intermediate prefill layers. It is further associated with downstream object hallucinations. These findings provide a semantic interpretable view of hallucination risk before generation, complementing existing latent representation-level analyses.

AI Use Statement

Generative AI tools were used to assist with literature discovery and summarization, experimental and statistical design feedback, code drafting and debugging, result interpretation, and manuscript drafting and language editing. The authors inspected the cited literature, executed and tested the analysis code, checked reported values against saved outputs, and reviewed all AI-assisted text. No model output was accepted as scientific evidence without verification. The authors take responsibility for the final content, claims, code, and artifacts in this work.

Ethics Statement

This work reuses public benchmark data and does not redistribute source images; the original dataset licenses and access conditions continue to apply. Human auditing concerns object presence and plausibility in public images and does not intentionally collect sensitive personal attributes. Plausibility labels remain subjective and may reflect annotator or cultural assumptions. We therefore release provenance and audit metadata, report uncertainty, and avoid interpreting the diagnostic as a universal measure of model truthfulness or as evidence about individual images beyond the evaluated concepts.

Reproducibility Statement

Section 4 specifies the models, metrics, frozen endpoints, controls, and clustered inference protocol. The appendix documents data construction, prompts, model-specific visual-null implementations, and additional results. The supplementary material contains machine-readable benchmark files, audit records, reconstruction and validation scripts, analysis code, frozen protocol files, and checksums. We will release these materials anonymously for review and publicly upon acceptance, subject to the source datasets' licenses.

References

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

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Kaustubh S Bukkapatnam. Hallutrace: Causal attribution and source-targeted decoding for hallucination in large vision-language models. In *Proceedings of the 4th Workshop on Advances in Language and Vision Research (ALVR)*, pp. 294–300, 2026.

Wenliang Dai, Zihan Liu, Ziwei Ji, Dan Su, and Pascale Fung. Plausible may not be faithful: Probing object hallucination in vision-language pre-training. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2136–2148, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Does object grounding really reduce hallucination of large vision-language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2728–2742, 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 Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6067–6085, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.287. URL https://aclanthology.org/2026.eacl-long.287/.

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

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, 2025.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Zhaoyang Li, Zhan Ling, Yuchen Zhou, Litian Gong, Erdem Biyik, and Hao Su. Oric: Benchmarking object recognition under contextual incongruity in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23673–23684, 2026.

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 *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: Towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Abin Shoby, Ta Duc Huy, Tuan Dung Nguyen, Minh Khoi Ho, Qi Chen, Anton van den Hengel, Phi Le Nguyen, Johan W Verjans, and Vu Minh Hieu Phan. Overthinking causes hallucination: Tracing confounder propagation in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9185–9194, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25315–25325, 2026.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. *arXiv preprint arXiv:2604.25642*, 2026.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011, 2024.

A Additional Experimental Details

This section provides additional details of PAC^3, model-specific visual-null constructions, prompts, generation settings, discovery–confirmation splitting, and statistical procedures used in the main experiments.

B Additional Results for RQ1

B.1 PAC-Type Behavioral Results

Table 4 provides the complete behavioral results for Local-, Global-, and Mixed-PACs. Across models, the PAC–imPAC false-positive difference is generally positive, while its magnitude and statistical strength vary across PAC types. These results complement Figure 2 and show that the behavioral susceptibility reported in the main text is not restricted to a single source of plausibility.

Table 4: Behavioral PI by PAC source. Values and confidence intervals are in percentage points and use parent-image-clustered inference.

Model	PAC type	n	PI ↓	95% CI	p
LLaVA-1.5-7B	Local	73	12.3	[4.1, 21.4]	.011
LLaVA-1.5-7B	Global	173	9.2	[4.3, 14.4]	$<$ .001
LLaVA-1.5-7B	Mixed	67	10.4	[1.5, 20.6]	.066
Qwen3-VL-8B	Local	73	20.5	[12.0, 30.6]	$<$ .001
Qwen3-VL-8B	Global	173	8.1	[3.4, 13.3]	.003
Qwen3-VL-8B	Mixed	67	20.9	[11.7, 30.6]	$<$ .001
DeepSeek-VL-7B	Local	73	17.8	[9.6, 27.0]	$<$ .001
DeepSeek-VL-7B	Global	173	4.0	[-1.8, 9.7]	.257
DeepSeek-VL-7B	Mixed	67	19.4	[8.6, 31.3]	.003

B.2 Semantic Extensions

Table 5 provides the complete results for the object-existence, attribute, and relation portions of PAC^3. The attribute and relation results extend the analysis to visually unsupported semantics beyond object identity. Because these extensions do not use the same matched PAC–imPAC triplet structure as the object-existence subset, they are interpreted as evidence of broader semantic susceptibility rather than direct estimates of PI.

Table 5: Behavioral extension across semantic units. Attributes pool count, color, and position pairs from MME; relations use spatial-relation pairs from VSR. Rates are percentages.

Model	Semantic unit	Pairs	GR ↑	Plausible/unsupported FPR ↓
LLaVA-1.5-7B	Object	313	79.9	11.8
LLaVA-1.5-7B	Attribute	89	87.6	22.5
LLaVA-1.5-7B	Relation	21	81.0	57.1
Qwen3-VL-8B	Object	313	90.4	14.4
Qwen3-VL-8B	Attribute	89	93.3	14.6
Qwen3-VL-8B	Relation	21	95.2	9.5
DeepSeek-VL-7B	Object	313	85.6	12.8
DeepSeek-VL-7B	Attribute	89	94.4	25.8
DeepSeek-VL-7B	Relation	21	61.9	47.6

C Additional Results for RQ2

C.1 POG across PAC Types

Table 6 reports POG separately for Local-, Global-, and Mixed-PACs across models. Positive gaps are observed across the PAC categories, supporting the conclusion that visual overinterpretation is not confined to a single plausibility source. The relative magnitude of the gaps, however, varies across models.

Table 6: Primary-layer POG by PAC source on the image-disjoint confirmation split.

Model	PAC type	n	POG ↓
LLaVA-1.5-7B	Local	17	11.934
LLaVA-1.5-7B	Global	23	6.212
LLaVA-1.5-7B	Mixed	14	15.068
Qwen3-VL-8B	Local	17	2.598
Qwen3-VL-8B	Global	23	0.949
Qwen3-VL-8B	Mixed	14	2.536
DeepSeek-VL-7B	Local	17	12.739
DeepSeek-VL-7B	Global	23	5.249
DeepSeek-VL-7B	Mixed	14	7.793

C.2 Grounded, PAC, and imPAC Support

Table 7 compares image-conditioned support for grounded concepts, PACs, and matched imPACs. The general ordering $S(c_g) > S(c_p) > S(c_n)$ provides an additional interpretation of POG: PACs remain visually unsupported, but receive greater image-conditioned support than implausible absent concepts while remaining below genuinely grounded concepts.

Table 7: Mean image-conditioned support for grounded concepts, PACs, and matched imPACs at the primary layer on confirmation.

Model	Layer	$S(c_g)$ ↑	$S(c_p)$ ↓	$S(c_n)$ ↓
LLaVA-1.5-7B	20	30.881	13.987	3.677
Qwen3-VL-8B	22	10.121	2.021	0.141
DeepSeek-VL-7B	18	25.126	5.869	-2.398

D Additional Results for RQ3

D.1 Generation and Alignment across PAC Types

Table 8 reports GOG and OGA separately for Local-, Global-, and Mixed-PACs. The results show that the magnitude of generation overinterpretation and the strength of prefill–generation alignment do not necessarily follow the same ordering. This distinction supports treating GOG and OGA as complementary measures of downstream overinterpretation.

Table 8: Generation behavior and POG–GOG alignment by PAC source. GOG is the parent-image-cluster mean in percentage points; OGA is Spearman's ρ.

Model	PAC type	$n_{\text{gen}}/n_{\text{OGA}}$	GR_{gen} ↑	GOG ↓	OGA ↓
LLaVA-1.5-7B	Local	73/58	74.2	11.9	0.59
LLaVA-1.5-7B	Global	173/75	71.7	8.7	0.52
LLaVA-1.5-7B	Mixed	67/44	71.9	16.9	0.38
Qwen3-VL-8B	Local	73/58	54.7	3.5	0.19
Qwen3-VL-8B	Global	173/75	41.3	0.6	0.23
Qwen3-VL-8B	Mixed	67/44	41.3	0.2	0.10
DeepSeek-VL-7B	Local	73/58	67.3	3.9	0.36
DeepSeek-VL-7B	Global	173/75	64.5	2.5	0.44
DeepSeek-VL-7B	Mixed	67/44	52.2	3.1	0.41

Figure 5 visualizes the pair-level association between all-layer POG and GOG. The dense mass at zero reflects PAC–imPAC pairs that are not differentially mentioned in the sampled generations, while the fitted trends summarize the positive, model-dependent association reported in the main text.

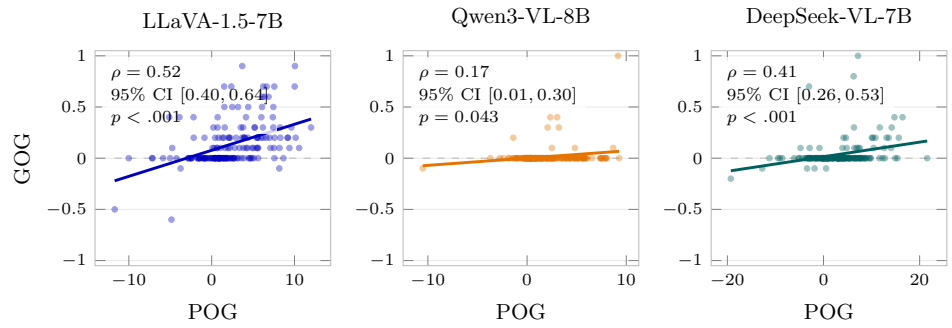

Figure 5: Association between all-layer POG and GOG across PAC–imPAC pairs. Lines show linear trends; annotations report Spearman's ρ with parent-image-clustered 95% confidence intervals and p-values.

D.2 Complementary CHAIR Analysis

Table 9 provides the complete CHAIR results. Unlike the controlled PAC^3 experiment, this analysis starts from naturally occurring open-ended generations and evaluates whether contextual plausibility distinguishes absent concepts that subsequently appear as hallucinations. For LLaVA-1.5-7B, contextual plausibility achieves an AUC of 0.826 on the confirmation data, and the hallucination rate rises from 0.1% in the lowest plausibility quintile to 3.9% in the highest quintile. Similar, though model-dependent, patterns are observed for Qwen3-VL-8B and DeepSeek-VL-7B. Because CHAIR does not provide the same matched PAC–imPAC control as PAC^3, we treat this experiment as complementary evidence rather than a direct test of the full PAC taxonomy. Together, the controlled and natural-generation results indicate that visual overinterpretation during prefill is associated with downstream object hallucination, while the strength of this association varies across models.

Table 9: Complementary CHAIR analysis over all 500 images per model. AUC measures how well contextual plausibility ranks subsequently hallucinated absent concepts; rates compare the lowest and highest plausibility quintiles on confirmation.

Model	Hallucinated ↓	Images	AUC ↑	Q1 rate ↓ (%)	Q5 rate ↓ (%)
LLaVA-1.5-7B	404	236/500	0.826	0.1	3.9
Qwen3-VL-8B	441	267/500	0.689	0.4	2.7
DeepSeek-VL-7B	157	132/500	0.739	0.1	1.4

E Extended Discussion

Interpreting Visual Overinterpretation. Our results suggest that visual absence alone does not fully characterize how VLMs represent unsupported concepts. Although PACs and matched imPACs are both absent from the image, PACs exhibit higher false-recognition rates and receive stronger image-conditioned support during prefill. The observed support ordering among grounded concepts, PACs, and imPACs further suggests that PACs occupy an intermediate representational status: they lack direct visual evidence but nevertheless acquire partial support from image-conditioned representations. We refer to this systematic preferential support for plausible but visually unsupported concepts as *visual overinterpretation*. This perspective complements prior analyses of hallucination risk by providing a concept-level interpretation of what kinds of unsupported semantics are preferentially represented before generation.

Scope and Limitations. Our findings should be interpreted as evidence of association rather than causality. POG captures a representation-level difference that emerges before generation and is associated with subsequent generation behavior, but this does not establish that the measured difference directly causes hallucination. In addition, the matched PAC–imPAC design is most complete for object existence. The attribute and relation extensions broaden the semantic coverage of PAC^3, but do not provide the same controlled triplet structure. Similarly, CHAIR primarily provides complementary evidence for contextual plausibility in naturally occurring hallucinations rather than a direct evaluation of the full PAC taxonomy. Finally, the observed variation across models and PAC

subtypes suggests that visual overinterpretation is a shared phenomenon whose manifestation and downstream relevance can depend on model architecture and semantic context.