Detecting Hallucinations in Vision-Language Models without Generating a Single Token

Anonymous ACL submission

Abstract

Hallucinations remain a persistent challenge for vision–language models (VLMs), which often describe nonexistent objects or fabricate facts. Existing detection methods typically operate *after* text generation, making intervention both costly and untimely. We investigate whether hallucination risk can instead be predicted *before* any token is generated by probing a model's internal representations in a single forward pass. Across a diverse set of vision–language tasks and eight modern VLMs including Llama-3.2, Gemma-3, Phi-4, and Qwen2.5-VL we examine three families of internal representations: (i) visual-only features without multimodal fusion, (ii) vision-token representations within the text decoder, and (iii) query-token representations that integrate visual and textual information before generation. Probes trained on these representations achieve strong hallucination-detection performance without decoding, reaching up to 0.93 AUROC on Gemma-3-12B, Phi-4-VL, and Molmo. Late query-token states are the most predictive for most models, while visual or mid-layer features dominate in a few architectures (e.g., ∼0.79 for Qwen2.5-VL using visual-only features). These results demonstrate that (1) hallucination risk is detectable *pre-generation*, (2) the most informative layer and modality vary across architectures, and (3) lightweight probes can enable early abstention, selective routing, and adaptive decoding to improve both safety and efficiency.

1 Introduction

Recent vision–language models (VLMs) such as Gemma-3 (Steiner et al., 2024), LLaVA-Next (Liu et al., 2024), Llama-3.2 (AI, 2024), Phi-4 (Li et al., 2022), Molmo (Molmo, 2023), Qwen2.5-VL (Bai et al., 2024), SmolVLM (Team, 2024b), and FastVLM (Team, 2024a) have achieved remarkable progress in generating coherent, human-like text grounded in visual inputs, redefining the frontier of multimodal AI. Yet despite these advances, VLMs remain prone to hallucination producing descriptions of nonexistent objects, invented attributes, or confident but unsupported claims that are factually inconsistent with the image. Such failures undermine the reliability and trustworthiness of these systems, especially in high-stakes applications such as autonomous navigation, medical imaging, and assistive technologies (Rohrbach et al., 2018; Li et al., 2023b).

Current approaches to mitigating hallucinations in vision–language models are largely **reactive**. Post hoc evaluation metrics such as CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023a), and FaithScore (Jing et al., 2024) require complete caption generation to identify hallucinated content, making them computationally expensive and unsuitable for real-time assessment. Other methods aim to reduce hallucinations *during* generation through decoding-time interventions such as HALC (Chen et al., 2024), Uncertainty-Guided Dropout Decoding (Fang et al., 2024), and perception-driven grounding augmentation (Ghosh et al., 2025). However, these techniques still depend on autoregressive decoding and cannot estimate hallucination risk before generation begins. This reveals a critical gap: few existing methods exploit internal model representations to anticipate hallucinations preemptively and enable real-time risk detection.

To address this gap, we introduce **HALP** (*HALlucination Prediction via Probing*), a lightweight framework that predicts hallucination risk from *pre-generative* internal states of vision–language models. HALP performs a **single forward pass** over an image–query pair and extracts three types of representations: (1) pooled visual features from the vision encoder, (2) decoder states at the last vision tokens across multiple layers, and (3) decoder states at the final query token at corresponding depths. Probes trained on these representations

Figure 1: Through out the paper, we consider 8 different domains, with distribution showing in Fig. 3(a). The probes scores reflect how possible the probe think the model is hallucinating. And we describe the probe in Section 3.

estimate whether a model is likely to hallucinate based on the visual context *before* any token is generated. Once trained, these probes require neither hallucination labels nor decoding, enabling efficient, generation-free risk prediction.

To assess generality, we construct a 10k-example benchmark covering diverse hallucination stressors including object presence (AMBER (An et al., 2023), POPE (Li et al., 2023a)), diagram and math reasoning (MathVista (Lu et al., 2023)), compositional and open-ended knowledge (MME (Fu et al., 2023)), illusion-style traps (HaloQuest (Cheng et al., 2023)), and targeted hallucination cases (HallusionBench (Guan et al., 2023)). Experiments on eight state-of-the-art open-source VLMs show that hallucination signals are accessible *before generation*: probes trained on internal states reach high AUROC without decoding, with query-token representations near the top of the decoder most predictive ($\sim$0.93 for Gemma-3, Phi-4, and Molmo). The most informative locus varies by architecture Qwen-VL and Llama-3.2 achieve strong detection from visual features ($\sim$0.79 and $\sim$0.77) or mid-decoder vision tokens indicating distinct failure pathways across models. Overall, lightweight probes generalize across architectures and representation types, offering a practical, low-cost solution for early hallucination risk assessment.

2 Background and Related Work

Hallucination in Vision-Language Models Hallucination in Vision-Language Models (VLMs) refers to the generation of content that does not align with the visual input, manifesting in multiple forms including object hallucinations, attribute hallucinations, and relational errors (Rohrbach et al., 2018). These errors reduce the reliability of VLMs, especially in critical applications like medical imaging, autonomous navigation, and assistive technologies. Hallucinations can arise from mismatches between the model's linguistic priors and visual grounding, or from biases present in the training data. For example, models might describe non-existent objects (object hallucinations), misattribute features to the wrong objects (attribute hallucinations), or incorrectly infer relationships between objects (relational hallucinations). These failures expose a disconnect between the model's linguistic fluency and its grounding in the visual world, making mitigation essential for ensuring trust in VLM outputs, particularly in safety-critical domains.

Representation Probing Representation probing is a diagnostic approach where lightweight classifiers or regressors are trained on internal model activations to assess the encoding of specific properties. In NLP, probes have been used to capture syntactic dependencies, part-of-speech tags, and coreference relations (Hewitt and Liang, 2019; Marvin and Linzen, 2018). In vision, linear probes on convolutional features have been employed to reveal emergent object detectors in scene-classification networks (Zhou et al., 2015). Recent work has extended probing to Vision Transformers, where intermediate layers encode rich class-specific and scene-level semantics (Chen et al., 2022; Sheta et al., 2025). These studies demonstrate the effectiveness of probing as a lightweight tool for detecting task-relevant features, motivating its use to identify hallucination signals early in VLMs.

Pre-Generation Hallucination Detection Recent advances have demonstrated that hallucination-related signals are present in the internal states of language models before generation occurs. Orgad et al. (2025) showed that LLMs' internal representations encode substantial information about truthfulness, with critical signals concentrated around specific tokens. They trained classifiers on hidden states to predict factual correctness, achieving high accuracy across multiple layers. Similarly, Wang et al. (2024) designed probes trained on trans-

former internal representations to predict hallucinatory behavior across grounded generation tasks, finding that probes can reliably detect hallucinations at many transformer layers, achieving 95% of peak performance as early as layer 4. Kadavath et al. (2022) demonstrated that language models possess significant self-knowledge about their outputs through calibration experiments, supporting the hypothesis that error signals are accessible before generation completes. Building upon these insights, HALP extends probing-based approaches to Vision-Language Models, extracting internal representations from various stages of the VLM pipeline to predict hallucination risk before any tokens are generated.

Post-Generation Hallucination Detection and Mitigation Complementary to pre-generation approaches, substantial research has focused on detecting and mitigating hallucinations after content generation. Post-hoc detection methods include uncertainty-guided dropout decoding (Fang et al., 2024), adaptive focal-contrast decoding (HALC) (Chen et al., 2024), and perception-driven grounding augmentation (Ghosh et al., 2025). Mishra et al. (2024) introduced fine-grained hallucination detection with detailed error taxonomies, enabling precise identification of entity, relation, and contradictory errors in generated text. While these methods achieve high detection accuracy, they require complete sequence generation, making them computationally expensive and unsuitable for real-time applications where early intervention is crucial. Evaluation typically relies on metrics like CHAIR_i (Rohrbach et al., 2018), which necessitates full caption generation for assessment.

3 HALP: Hallucination Prediction via Probing

3.1 Vision-Language Model Preliminaries

While there are variations, modern vision-language models (VLMs) generally adopt a three-component architecture (Liu et al., 2023). Given an input image I and a text query Q, the processing pipeline operates as follows:

Vision Encoding: A vision encoder $\text{Enc}(\cdot)$, such as the one from CLIP (Radford et al., 2021), processes the input image I, producing a sequence of vectors of *visual features* $U = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_M\}, \mathbf{u}_i \in \mathbb{R}^{d_I}$, where M is the number of visual patches and d_I is the dimension of the feature.

Multimodal Projection: A vision-to-text projection layer $\mathcal{M}$ maps the visual features into the language model's input embedding space of dimension d, producing a sequence of *vision tokens* $V = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M\}$, with $\mathcal{M}(\mathbf{u}_i) \mapsto \mathbf{v}_i, \mathbf{v}_i \in \mathbb{R}^d$, enabling cross-modal fusion.

Language Decoding: The input text instruction or query Q is tokenized and embedded into $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of N tokens. Togetherwith the projected vision tokens, the concatenation of (V, X) forms the input sequence to a Transformer-based decoder (Vaswani et al., 2017) with L layers. Each layer $\ell \in \{1, 2, \ldots, L\}$ produces hidden states $\mathbf{h}_i^{(\ell)}$ for each token position i. The model then autoregressively generates the output text sequence $Y = \{y_1, y_2, \ldots, y_T\}$ that is supposed to ground the vision information.

Overview Our goal is to explore whether VLMs already encode information about their knowledge certainty or propensity to hallucinate within their internal representations, before any text generation occurs. HALP intercepts and analyzes the internal representations at three critical stages of this pipeline, extracting features that encode both visual understanding and multimodal reasoning. We hypothesize that these pre-generation features contain sufficient information to predict whether the model will produce hallucinated content, enabling real-time risk assessment without requiring complete sequence generation.

Next we describe how HALP explores simple probes on various VLM features for hallucination detection. An overview of the pipeline is shown in Figure 2.

3.2 Hallucination Evaluation

Hallucination detection in VLMs typically requires generating *complete* responses and comparing them against the ground truth or visual evidence to identify factual inconsistencies. Prior research has adopted different VLM hallucination evaluation protocols, such as CHAIR (Rohrbach et al., 2018) and POPE (Li et al., 2023a) that rely on the ground truth, and FaithScore (Jing et al., 2024) that relies on the visual evidence.

To explore how pre-generation internal features encode information for detecting whether VLMs hallucinate on their visual inputs, we first systematically evaluate model hallucinations through a unified assessment process.

Given a collection of (I^j, Q^j) testing examples

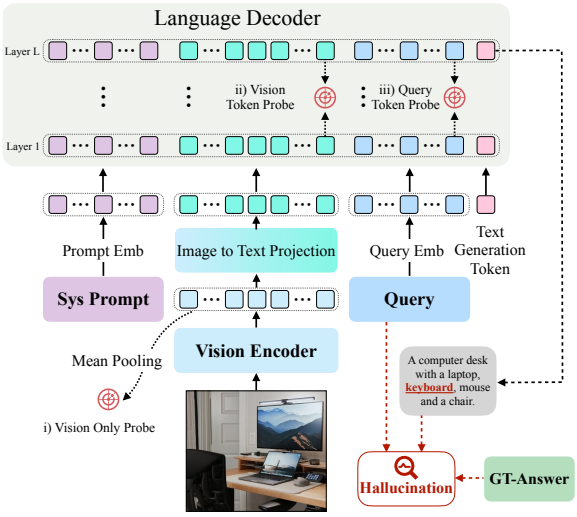

Figure 2: HALP detection pipeline. We extract three types of representations: (1) visual feature vectors from the encoder, (2) vision token states from decoder layers at the final vision patch position, and (3) query token states from the same layers at the final query position. These representations are used independentlyprobed for hallucination prediction.

along with their ground truth answers or reference generations Y^j, we run each VLM using the model's standard inference procedure to complete its generation $\hat{Y}^j$ given inputs (I^j, Q^j). We then employ an automated evaluation framework to determine the hallucination occurrence, denoted as $b^j \in \{0, 1\}$ which is a binary label indicating whether any hallucination occurred in the generated response $\hat{Y}^j$.

In particular, we use LLM-based judging (Zheng et al., 2023) to ensure reliable and standardized hallucination detection across diverse queries and visual domains. Formally, $\text{LLM}_{\text{judge}}(\hat{Y}^j, Y^j, Q^j) \mapsto b^j$, where we purpose an advanced large language model $\text{LLM}_{\text{judge}}$ to compare the generated response $\hat{Y}^j$ against the ground truth Y^j while considering the query Q^j, to identify whether there exists hallucinations such as factual inconsistencies, object/relation/attribute misinterpretations, etc. The obtained hallucination labels b^j will serve as targets for our probing. Experimental discussions on the reliability of LLM-as-a-judge in our context can be found in Section 5.5.

3.3 Pre-Generation Feature Extraction

HALP extracts three distinct types of internal representations (see Figure 2) from a *single* forward pass through the VLM before response token generation, each capturing different aspects of visual understanding and multimodal reasoning:

Visual Features (VF) The globally mean pooled output from the vision encoder after processing the image I but before the multimodal projection layer $\mathcal{M}$: $\bar{\mathbf{u}} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{u_i}$. This representation captures pure visual information before any LLM and textual integration. Hallucination detection based on this directly probes perception-based signals.

Vision Token Representations (VT) Hidden states from the decoder LLM layer ℓ at the *final position* of the visual token sequence V. These representations capture how visual information is processed and integrated within the multimodal text decoder. We extract these from five strategically selected layers: $\ell \in \{1, \lfloor L/4 \rfloor, \lfloor L/2 \rfloor, \lfloor 3L/4 \rfloor, L\}$, where L is the total number of decoder layers and could be different for different VLMs.

Query Token Representations (QT) Hidden states from the decoder LLM layer ℓ at the *final position* of the query token sequence. Note that this is the final position of the concatenated sequence (V, Q), which encodes the integration of both vision tokens and query text tokens. These representations encode the fully contextualized multimodal information that directly precedes text generation. We extract them from the same five layers as in VT above.

3.4 Probing

For each representation type at a particular layer, we train a lightweight probe individually to assess its potential of hallucination detection before generation. Each probe is a 3-layer MLP with hidden dimensions [512, 256, 128], using ReLU activations. The final layer outputs a binary classification label for b^j for hallucination presence. More details in probe training and hallucination detection evaluation can be found in Section 4.2.

4 Data and Experiments

4.1 Data Construction

We assemble a dataset of 10,000 samples of image and QA pairs from six established VQA benchmarks to ensure broad coverage of vision-language challenges. Detailed data composition can be found in Table 1. Figure 3(a) shows the domain breakdown: Attribute Recognition and Visual Understanding dominate (60% combined), with substantial contributions from Spatial Reasoning (17.7%) and smaller but critical subsets

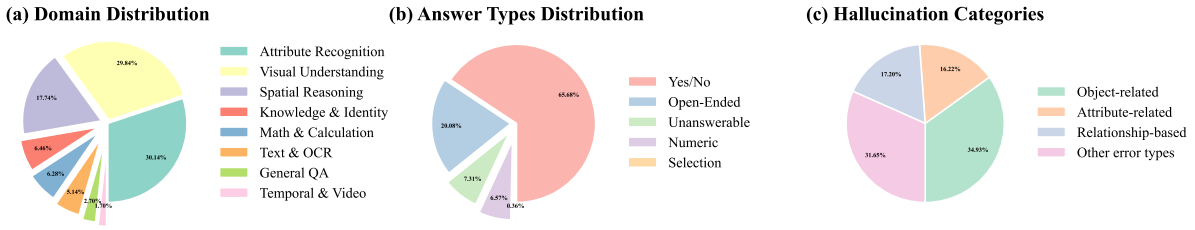

Figure 3: Hallucination detection dataset dstributions of task domains, answer formats, and hallucination types.

Dataset	Focus	Samples
AMBER (An et al., 2023)	discriminative tasks	3,926
HaloQuest (Cheng et al., 2023)	adversarial challenges	2,784
POPE (Li et al., 2023a)	object hallucination	1,230
MME (Fu et al., 2023)	multimodal reasoning	885
HallusionBench (Guan et al., 2023)	visual illusions	617
MathVista (Lu et al., 2023)	mathematical reasoning	558
Total		**10,000**

Table 1: The hallucination detection dataset composition.

Model Name	VF	VT	QT	Average
Gemma3-12B	0.6736	0.5956	0.9349	0.7347
FastVLM-7B	0.6830	0.7028	0.6136	0.6665
LLaVa-Next-8B	0.6108	0.6270	0.9026	0.7135
Molmo-V1-7B	0.6830	0.6867	0.9193	0.7630
Qwen2.5-VL-7B	0.7873	0.6683	0.9150	0.7902
Llama-3.2-11B-Vision	0.7703	0.7377	0.8959	0.8013
Phi4-VL-5.6B	0.6166	0.7738	0.9033	0.7646
SmolVLM2-2.2B	0.7238	0.6894	0.9014	0.7715
Average	0.6935	0.6852	0.8733	0.7507

Table 2: AUROC performance for hallucination detection probes across different vision-language models and different model representations.

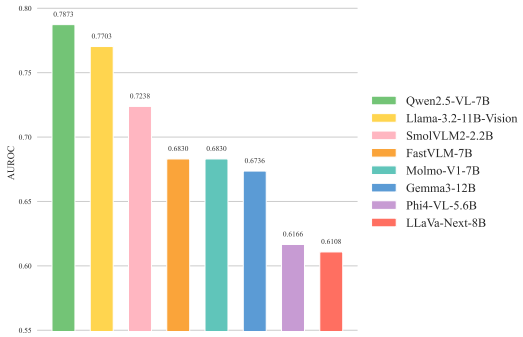

Figure 4: AUROC scores for visual features models.

in Knowledge & Identity, Math & Calculation, Text & OCR, General QA, and Temporal & Video. Figure 3(b) presents the answer format distribution: Yes/No questions comprise two-thirds of the samples, while Open-Ended, Unanswerable, Numeric, and Selection types provide diverse reasoning and generation tasks. Finally, Figure 3(c) categorizes hallucination patterns: Object-related errors are most frequent (34.9%), followed by Other error types (31.7%), Relationship-based failures (17.2%), and Attribute-related misattributions (16.2%). Together, these distributions validate that our dataset captures the variety of tasks and error modes necessary for rigorous hallucination detection evaluation.

4.2 Experimental Setup

Models We quantitatively evaluate HALP's effectiveness across eight contemporary VLMs spanning diverse architectures and parameter scales: Gemma-3 (Steiner et al., 2024), LLaVA-Next-8B (Liu et al., 2024), Llama 3.2 (AI, 2024), Phi-4 (Li et al., 2022), Molmo-7B (Molmo, 2023), Qwen-2.5-VL-7B (Bai et al., 2024), SmolVLM-2B (Team, 2024b) and FastVLM-7B (Team, 2024a).

Post-Generation Hallucination Detection We employ an LLM-as-a-judge framework for automated hallucination assessment follows section 3.2. This automated evaluation maintains consistency across our 10,000-sample dataset while enabling scalable hallucination assessment.

Hallucination Detection Evaluation We use AUROC (Area Under Receiver Operating Characteristic curve) as our primary metric, which provides a threshold-independent measure of binary classification performance. AUROC scores range from 0.5 (random chance) to 1.0 (perfect classification), making them ideal for comparing hallucination prediction across different models and representation types.

Training Configuration Training was performed with the Adam optimizer (learning rate = 1×10^3), a batch size of 32, and 50 epochs. The dataset was split 80/20 for training and validation using stratified sampling to preserve the same proportion of different hallucination type examples in each set, ensuring that the validation fold reflects the overall hallucination type. All experiments use random seed 42 for reproducibility. All experiments are conducted on NVIDIA RTX 4090 GPUs with 24GB VRAM. We use PyTorch 2.0 with CUDA 11.8 for efficient computation. Model representations are extracted using precision fp16 to optimize memory usage while maintaining nu-

merical stability.

5 Results and Analysis

5.1 Main Results

Table 2 summarizes the key results. Across the evaluated models, three consistent patterns emerge.

First, *query token (QT) representations dominate across models*, with seven out of eight systems achieving peak AUROC values between 0.90 and 0.94 when classification is based on deep query states.

Second, *architectural heterogeneity is evident in optimal information encoding*. In some models, visual features already yield strong hallucination detection (Qwen2.5-VL: 0.7873 AUROC; Llama-3.2-11B: 0.7703), suggesting balanced encoding from perception onward. In contrast, models such as LLaVA-Next and Phi4-VL perform poorly with visual features features (0.6108, 0.6166 AUROC), depending more on multimodal integration for robust prediction. *FastVLM-7B illustrates architectural peculiarity*, with maximum AUROC (0.7093) obtained from vision token features rather than from the query position, suggesting a preference for fusion at earlier reasoning stages.

Third, when comparing all representation types, query token vectors consistently provide the most reliable hallucination prediction, with an average AUROC of 0.8898. This outperforms vision tokens (0.7090) and visual features features (0.6955), highlighting that the ideal intervention point for hallucination detection is just after multimodal reasoning but prior to text generation. Vision token features themselves show inconsistent value, plateauing at around 0.70 AUROC, with only Llama-3.2-11B and Phi4-VL exceeding 0.75. Pure vision representations act as a competitive baseline, especially in vision-grounded models, but there remains a notable gap between visual features and query-based performance (ΔAUROC $\approx$ 0.15–0.25), demonstrating the importance of multimodal reasoning components.

Overall, our results establish HALP as a robust framework for detecting hallucination-predictive signals across a range of VLM architectures. Query token probes consistently provide the most effective means for pre-generation assessment, supporting the integration of such methods in safety-critical applications that require rigorous risk evaluation prior to text output.

5.2 Layer-wise Performance of Hallucination Detection

Our layer-wise analysis reveals distinct patterns in how hallucination-predictive signals evolve across decoder depths, with significant architectural variations in optimal extraction layers (Figures 3 and 4. We analyze AUROC performance across five strategic positions: $\ell \in \{1, \lfloor L/4 \rfloor, \lfloor L/2 \rfloor, \lfloor 3L/4 \rfloor, L\}$, where L represents the final decoder layer. Extraction depth for optimal query token performance varies: certain models, such as Gemma3-12B, deliver maximum AUROC at the final layer (0.9349 at L), while others like Molmo-V1 are best at intermediate decoding depths (0.9365 at $L/2$). The analysis demonstrates two fundamentally different progression patterns depending on the representation type and model architecture. Query token representations generally show substantial improvement from shallow to deep layers, while vision token representations exhibit more stable but limited performance across all depths.

Query Token Layer-wise Progression Query token representations demonstrate remarkable layer-wise improvement across most models, with AUROC scores increasing substantially as processing depth increases. Most models achieve their peak performance in deep layers ($L/2$ to L), with dramatic improvements from early layers. For instance, Gemma-3-12B exemplifies the typical progression pattern, with consistent improvement across all layers: $0.7165 \rightarrow 0.8119 \rightarrow 0.9247 \rightarrow 0.9315 \rightarrow 0.9349$, showing how hallucination signals become increasingly concentrated as the model approaches final output generation. This monotonic improvement suggests that multimodal reasoning gradually refines hallucination-relevant features through successive transformer layers.

Vision Token Layer-wise Behavior Vision token representations show markedly different patterns, with generally stable performance across layers and limited overall improvement. Most models plateau around 0.65-0.70 AUROC across all depths, suggesting that visual information processed within decoder layers provides consistent but limited hallucination prediction capability. Notably, Phi4-VL achieves the strongest vision token performance, maintaining 0.77-0.78 AUROC across layers, while FastVLM-7B peaks at Layer $L/2$ (0.7093), aligning with its unique architectural behavior identified in our cross-model analysis.

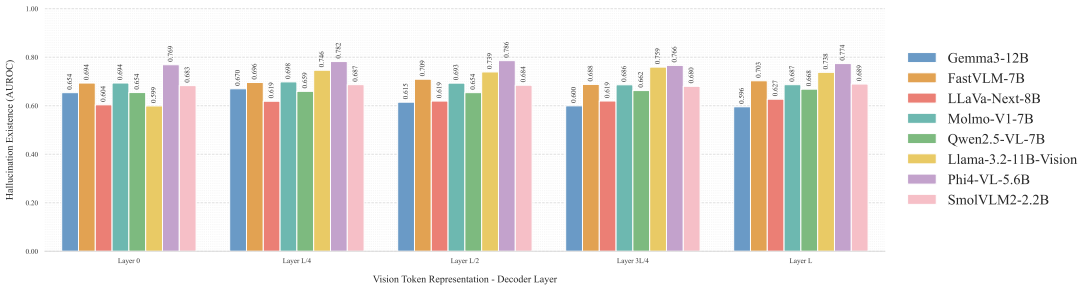

Figure 5: AUROC scores for Vision Token Representation.

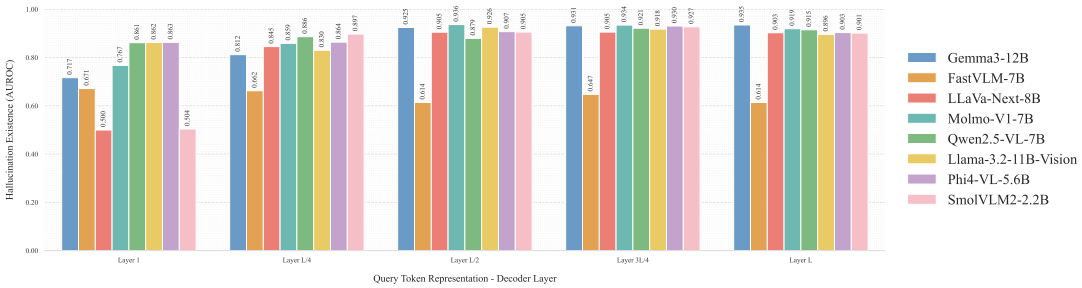

Figure 6: AUROC scores for Query Token Representation.

Layer-wise Optimization Insights The optimal extraction layers vary significantly across models and representation types. For query tokens, Layer $3L/4$ emerges as a consistently strong choice, achieving peak or near-peak performance for most models (Gemma-3: 0.9315, Phi4-VL: 0.9305, SmolVLM: 0.9272). This suggests that the penultimate reasoning stage, just before final output commitment, contains maximal hallucination-predictive information.

Vision tokens show more model-specific optimal layers: FastVLM-7B peaks at Layer $L/2$ (0.7093), Phi4-VL at Layer $L/2$ (0.7858), and Llama-3.2-11B-Vision at Layer $3L/4$ (0.7592). This heterogeneity indicates that different architectures process and preserve visual hallucination signals at varying depths, likely reflecting diverse approaches to visual-textual integration.

These layer-wise patterns provide crucial insights for practical deployment: practitioners can optimize computational efficiency by extracting representations at optimal layers rather than defaulting to final outputs, while the consistent strength of mid-to-deep query token representations offers robust hallucination prediction across diverse architectural choices.

5.3 Hallucination Types and Application Domains

Our analysis (Table 3, Table 4) shows clear variations in hallucination rates and detection performance across error types(Fig. 7) and domains(Fig. 1). Attribute-Related and Other errors are the most frequent and hardest to detect, while Object-Related errors though less common yield strong QT signals (AUROC 0.878). Relationship errors are rare (6.06 %) but notably difficult to predict (AUROC 0.708). Domains such as Temporal & Video (46.9 % error rate, QT AUROC 0.456) and Knowledge & Identity (21.2 %, 0.694) are the weakest areas, whereas Attribute Recognition (7.1 %, 0.863), Text & OCR (6.9 %, 0.855), and Visual Understanding (10.3 %, 0.859) exhibit robust performance. Across all categories, query-token (QT) probes consistently outperform vision-only (VF) representations, with the largest AUROC gaps for Attribute (+0.237) and Relationship (+0.149) errors. These findings underscore QT's effectiveness in flagging hallucinations early and highlight domain-specific challenges for targeted model improvements.

Hallucination Type	Rate (%)	VT AUROC	QT AUROC	ΔAUROC
Attribute-Related	11.92	0.567	0.804	0.237
Object-Related	7.94	0.734	0.878	0.144
Relationship	6.06	0.559	0.708	0.149
Other	13.47	0.755	0.827	0.072

Table 3: Hallucination rates and AUROC performance by error type averaged across models. ΔAUROC shows the performance gain from query token over visual features representations.

These findings underscore QT's effectiveness in flagging hallucinations early and highlight domain-specific challenges for targeted model improvements.

Key Findings Object-Related hallucinations achieve the highest detection performance (0.878

Domain	Rate (%)	QT AUROC
Temporal & Video	46.90	0.456
Knowledge & Identity	21.19	0.694
Math & Calculation	16.79	0.819
Visual Understanding	10.26	0.859
Attribute Recognition	7.14	0.863
General QA	7.04	0.797
Text & OCR	6.94	0.855
Spatial Reasoning	5.85	0.713

Table 4: Hallucination rates and query token AUROC performance by application domain, ordered by hallucination rate.

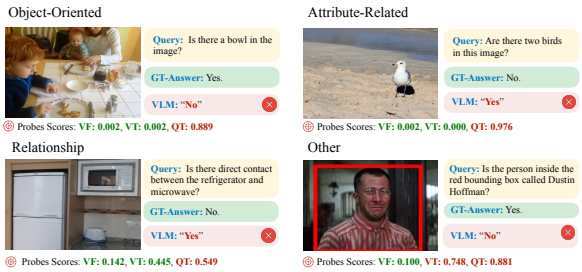

Figure 7: Hallucination Types and Probe Scores To Indicate Hallucination

AUROC) despite moderate frequency, while Attribute-Related errors show the largest improvement from multimodal reasoning (ΔAUROC = 0.237). Temporal & Video tasks represent the most challenging domain with exceptionally high error rates (46.90%) and poor detection performance (0.456 AUROC), indicating fundamental VLM limitations in temporal reasoning. High-risk domains (Temporal & Video, Knowledge & Identity) require enhanced monitoring, while low-risk domains (Attribute Recognition, Text & OCR, Spatial Reasoning) demonstrate reliable performance suitable for automated deployment.

5.4 Qualitative Analysis

Fig. 7 illustrates four hallucination categories Object-Oriented, Relationship, Attribute-Related, and Other along with question, image, prompts, model answers ($\hat{Y}^j$), ground truths (Y^j), and probe scores for vision-only (VF), vision-token (VT), and query-token (QT) representations.

These examples demonstrate how query-token (QT) representations effectively identify hallucination risks across diverse error types before text generation.

5.5 LLM-as-a-Judge Evaluation

LLM-as-a-Judge Validation We adopt LLM-as-a-Judge with GPT-4 (Zheng et al., 2023) to label hallucinations. To verify its reliability, three expert annotators independently labeled a stratified sample of 200 examples, achieving substantial inter-annotator agreement with a Fleiss value of 0.89. The GPT-4 judge matched human consensus on 97% of cases, confirming its suitability for large-scale, automated hallucination labeling in our multi-domain evaluation.

5.6 Practical Implications

The analysis reveals distinct risk profiles for different application contexts. *High-risk domains* (Temporal & Video, Knowledge & Identity) require enhanced monitoring and potentially specialized architectures, while *moderate-risk domains* (Math & Calculation, Visual Understanding) can benefit from standard HALP deployment with appropriate thresholds. *Low-risk domains* (Attribute Recognition, Text & OCR, Spatial Reasoning) demonstrate reliable performance and can support more automated deployment scenarios. However, even in low-risk domains, the substantial performance improvement from query token over visual features probes (average improvement of 0.2-0.3 AUROC) justifies the computational overhead of deeper representation extraction.

These domain-specific patterns provide actionable guidance for practitioners: applications focused on temporal reasoning or factual knowledge retrieval should implement additional safeguards, while visual recognition and OCR tasks can rely more confidently on HALP's pre-generation predictions.

6 Conclusion

We presented HALP, a lightweight framework for pre-generative prediction of hallucinations in vision language models (VLMs). Through extensive experiments across eight modern VLMs and diverse multimodal benchmarks, we examined representations from global visual features to decoder-level vision and query token states within a single forward pass. Our findings show that hallucination risk is detectable before generation and that the most informative signals vary across architectures. The approach enables rapid, real-time hallucination risk assessment without costly decoding. Future work includes integrating probe-based risk scores into decoding-time control for on-the-fly hallucination prevention, adaptive routing, and confidence-aware generation, as well as extending the approach to finer-grained detections.

7 Ethical Considerations

This work addresses hallucination detection in vision-language models, which carries several important ethical implications that must be carefully considered.

Beneficial Applications and Risk Mitigation Our hallucination detection framework serves beneficial purposes by enhancing AI safety and reliability. By enabling early detection of model hallucinations, this research contributes to preventing the spread of misinformation and reducing potential harm from AI-generated false content. The probing approach offers a transparent method to assess model reliability without requiring access to proprietary architectures, supporting broader AI safety research.

Limitations and Misuse Potential While designed for safety enhancement, this technology could potentially be misused. Adversarial actors might attempt to circumvent hallucination detectors or exploit knowledge of detection mechanisms to craft more sophisticated deceptive content. Additionally, over-reliance on automated hallucination detection without human oversight could lead to false confidence in AI system outputs.

Dataset and Evaluation Concerns Our evaluation relies on existing VQA benchmarks, which may contain inherent biases or limitations in representing diverse populations and scenarios. The use of GPT-4 as a judge for hallucination labeling introduces potential bias from the judging model itself, though we employ lenient criteria to minimize false positives. Future work should incorporate more diverse evaluation frameworks and human annotation validation.

Transparency and Reproducibility We commit to open research practices by providing detailed methodology and encouraging reproducible research. However, we acknowledge that computational resource requirements may limit accessibility for some researchers. The internal representation analysis approach requires significant computational infrastructure, potentially creating barriers to widespread adoption.

Broader Impact This research contributes to the responsible development of AI systems by providing tools to assess and improve their reliability. However, we emphasize that hallucination detection should complement, not replace, human judgment in high-stakes applications. Continued research into AI safety, interpretability, and robustness remains essential for the responsible deployment of vision-language models in real-world applications.

8 Limitations

While our probing framework effectively detects a range of hallucination types, several limitations remain. First, the reliance on VQA benchmarks may introduce dataset biases, limiting generalization to other vision–language tasks and real-world scenarios. Second, GPT-4 based judging, despite lenient criteria, can inherit biases and occasional misclassifications from the judge model, affecting label quality. Third, our approach focuses on pre-generation internal signals and does not directly address post-generation mitigation strategies. Fourth, high computational cost for extracting and probing internal representations may hinder deployment in resource-constrained settings. Fifth, the current taxonomy covers major hallucination types but may not capture nuanced cases, such as context-dependent or culturally specific errors, which warrant further study. Finally, our experiments use small-to-medium VLM sizes; performance characteristics and hallucination patterns in larger models remain underexplored.

References

Meta AI. 2024. Llama 3.2 vision: Open vision-language models for mobile and edge ai. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2025-10-06.

Junyang An, Beichen Wang, Haiyang Xu, Jiahuan Liu, Chenyang Wang, Hao Chen, Ruoxi Zhang, Jinguo Shi, and Enhance Chen. 2023. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2024. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. 2022. Adaptformer: Adapting vision transformers for scalable visual recognition. *Preprint*, arXiv:2205.13535.

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

Zhecan Cheng, Chenxi Huang, Wenjia Dai, Yilun Xu, Xin Eric Liu, and Mark Yatskar. 2023. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. *arXiv preprint arXiv:2407.15680*.

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. 2024. From uncertainty to trust: Enhancing reliability in vision-language models with uncertainty-guided dropout decoding. *Preprint*, arXiv:2412.06474.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenrong Qiu, Wei Lin, Jinrui Yang, Xiangyu Zheng, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. 2025. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. *Preprint*, arXiv:2405.15683.

Tianrui Guan, Fuxiao Chu, Zongxia Wang, Longtian Zhou, Jiaqi Mu, Zijun Hu, Zipeng Lin, Xiao Liu, Jing Wang, Jian Yao, and 1 others. 2023. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *arXiv preprint arXiv:2310.14566*.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023a. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Yulia Tsvetkov, Graham Neubig, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. In *Proceedings of the Conference on Language Modeling (COLM)*.

Research Team Molmo. 2023. Molmo: A vision-language model for multimodal tasks. *arXiv preprint arXiv:2305.13624*.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *Proceedings of the Twelfth International Conference on Learning Representations (ICLR)*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

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

Hala Sheta, Eric Huang, Shuyu Wu, Ilia Alenabi, Jiajun Hong, Ryker Lin, Ruoxi Ning, Daniel Wei, Jialin Yang, Jiawei Zhou, and 1 others. 2025. From behavioral performance to internal competence: Interpreting vision-language models with vlm-lens. *arXiv preprint arXiv:2510.02292*.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, and 1 others. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

FastVLM Team. 2024a. Fastvlm: A fast and efficient vision-language model for real-time applications. *arXiv preprint arXiv:2406.13259*.

SmolVLM Team. 2024b. Smolvlm: A lightweight vision-language model. *arXiv preprint arXiv:2407.09538*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Sky CH Wang, Benjamin Van Durme, Jason Eisner, and Chris Kedzie. 2024. Do androids know they're only dreaming of electric sheep? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2015. Object detectors emerge in deep scene cnns. *Preprint*, arXiv:1412.6856.

Appendix

A Model Analysis

This appendix provides a detailed breakdown of the performance metrics for each Vision-Language Model (VLM) evaluated in this study. For each model, we present the AUROC score for the binary classification task. The results are organized by the type of internal representation used for the probe: the Visual Features, Last Vision Token Representation and Last Query Token Representation at various decoder layers

Model Name	VF	VT	QT
Layer 1		0.6538	0.7165
Layer $L/4$		0.6698	0.8119
Layer $L/2$	0.6736	0.6146	0.9247
Layer $3L/4$		0.5995	0.9315
Layer L		0.5956	0.9349

Table 5: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using Gemma3-12B.

Gemma3-12B The results for Gemma3-12B demonstrate a dramatic shift in hallucination predictive power from vision-based to query token representations across decoder layers. As shown in Table 5, query token probes significantly outperform both visual features and vision token approaches.

- **Query Token Dominance**: Query token representations show exceptional performance, with AUROC increasing progressively through deeper layers from 0.7165 (Layer 1) to 0.9349 (Layer n). This suggests that hallucination signals become increasingly concentrated in the final reasoning states before generation.
- **Vision Representations Plateau**: Both visual features (0.6736) and vision token representations (peak: 0.6698 at Layer $L/4$) show moderate but limited predictive power. Vision token performance actually degrades in deeper layers, declining from 0.6698 to 0.5956.
- **Layer-wise Progression**: The most striking finding is the consistent improvement in query token performance with depth, achieving near-perfect detection (0.9349 AUROC) at the final layer, indicating that Gemma3-12B's hallucination signals are most accessible just before text generation begins.

Model Name	VF	VT	QT
Layer 1		0.6935	0.6715
Layer $L/4$		0.6957	0.6623
Layer $L/2$	0.6830	0.7093	0.6139
Layer $3L/4$		0.6879	0.6475
Layer L		0.7028	0.6136

Table 6: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using FastVLM-7B.

FastVLM-7B FastVLM-7B exhibits a unique architectural behavior that diverges significantly from other models in the study. As shown in Table 6, this model demonstrates an unusual pattern where vision token representations outperform both visual features and query token approaches.

- **Vision Token Superiority**: Vision token representations achieve the highest performance, peaking at 0.7093 AUROC in Layer $L/2$. This suggests that FastVLM-7B's hallucination signals are most accessible when visual information is processed within the decoder context at intermediate depths.
- **Query Token Limitations**: Unlike other models, query token representations show consistently poor performance across all layers (0.6139-0.6715), with performance actually degrading in deeper layers. This indicates that hallucination signals do not concentrate in the final reasoning states for this architecture.
- **Architectural Anomaly**: The consistent underperformance of query tokens compared to vision tokens (0.7093 vs 0.6715 best scores) suggests that FastVLM-7B processes multimodal information differently, potentially with weaker integration between visual and textual reasoning pathways. This architectural divergence makes it an outlier among contemporary VLMs.

LLaVA-Next-8B For LLaVA-Next-8B, query token representations again dominate hallucination detection, with AUROC peaking at 0.9053 in Layer

$3L/4$ (Table 7). visual features and vision token representations offer limited predictive power.

- **Query Token Peak**: The best query token performance (0.9053 AUROC) occurs at Layer 3n/4, closely followed by Layer $L/2$ (0.9049). This indicates that hallucination signals for LLaVA-Next-8B are most accessible just before the final decoding stage.

Model Name	VF	VT	QT
Layer 1		0.6037	0.4996
Layer $L/4$		0.6185	0.8453
Layer $L/2$	0.6108	0.6192	0.9049
Layer $3L/4$		0.6194	0.9053
Layer L		0.6270	0.9026

Table 7: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using LLaVA-Next-8B.

- **Vision Token Marginality**: Vision token representations reach a maximum of 0.6270 AUROC at the final layer, showing marginal improvement over the visual features baseline (0.6108).
- **Consistency Across Depths**: Query token performance rises sharply from shallow layers (0.4996 at Layer 1) to deeper layers, highlighting the progressive concentration of hallucination-relevant features in the decoder.

Model Name	VF	VT	QT
Layer 1		0.6936	0.7675
Layer $L/4$		0.6982	0.8588
Layer $L/2$	0.6830	0.6931	0.9365
Layer $3L/4$		0.6862	0.9345
Layer L		0.6867	0.9193

Table 8: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using Molmo-V1-7B.

Molmo-V1-7B Molmo-V1-7B exhibits very strong query token performance, with the highest AUROC of 0.9365 at Layer $L/2$ (Table 8).

- **Peak Query Token**: Query token probe peaks at Layer $L/2$ with AUROC 0.9365, indicating that hallucination signals are most salient in mid-to-late decoder states.
- **Vision Token Moderate**: Vision token representations peak at Layer $L/4$ (0.6982 AUROC) but remain consistently below the query token probe, suggesting moderate predictive power.
- **Stable Visual Features**: The visual features probe (0.6830 AUROC) provides a reasonable baseline but is outperformed by all deeper multimodal representations.

Model Name	VF	VT	QT
Layer 1		0.6543	0.8614
Layer $L/4$		0.6593	0.8863
Layer $L/2$	0.7873	0.6539	0.8794
Layer $3L/4$		0.6625	0.9215
Layer L		0.6683	0.9150

Table 9: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using Qwen2.5-VL-7B.

Qwen2.5-VL-7B Qwen2.5-VL-7B achieves its best hallucination detection performance with both visual features and query token probes. As shown in Table 9:

- **Strong Visual Features Baseline**: The visual features probe reaches 0.7873 AUROC, indicating robust visual grounding.
- **Peak Query Token**: Query token representations peak at Layer $3L/4$ with 0.9215 AUROC, demonstrating the importance of mid-to-deep decoder states.
- **Moderate Vision Token**: Vision token representations obtain a maximum of 0.6683 AUROC at the final layer, showing moderate predictive power relative to other representation types.

Model Name	VF	VT	QT
Layer 1		0.5991	0.8623
Layer $L/4$		0.7463	0.8301
Layer $L/2$	0.7703	0.7392	0.9259
Layer $3L/4$		0.7592	0.9178
Layer L		0.7377	0.8959

Table 10: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using Llama-3.2-11B-Vision.

Llama-3.2-11B-Vision Llama-3.2-11B-Vision demonstrates strong performance across multiple representation types. As shown in Table 10, this model achieves excellent visual features performance while maintaining superior query token detection.

- **Balanced Performance**: visual features probe achieves 0.7703 AUROC, indicating robust visual grounding. Vision token representations peak at Layer $3L/4$ (0.7592), showing competitive intermediate performance.
- **Query Token Peak**: Query token representations achieve maximum performance at Layer $L/2$ (0.9259 AUROC), demonstrating strong hallucination signal concentration in mid-deep decoder states.
- **Vision-Grounded Architecture**: The strong visual features baseline suggests that Llama-3.2-11B-Vision encodes substantial hallucination-relevant information at the visual processing stage, making it less dependent on multimodal reasoning for detection.

Model Name	VF	VT	QT
Layer 1		0.7689	0.8629
Layer $L/4$		0.7821	0.8638
Layer $L/2$	0.6166	0.7858	0.9072
Layer $3L/4$		0.7662	0.9305
Layer L		0.7738	0.9033

Table 11: Hallucination Detection (AUROC) for probes built on visual features, vision token, and query token representations at different decoder layers using Phi4-VL.

Phi4-VL Phi4-VL exhibits unique behavior with strong vision token performance alongside excellent query token detection. Table 11 shows distinctive architectural characteristics.

- **Vision Token Strength**: Vision token representations perform exceptionally well, peaking at Layer L/2 (0.7858 AUROC), suggesting effective visual-contextual integration within the decoder.
- **Query Token Excellence**: Query token representations achieve maximum performance at Layer 3L/4 (0.9305 AUROC), demonstrating strong hallucination prediction capability.
- **Visual Features Limitation**: The relatively weak visual features performance (0.6166 AUROC) indicates that Phi4-VL relies heavily on multimodal integration for hallucination detection, rather than pure visual features.

B Dataset Composition and Characteristics

Our comprehensive evaluation dataset consists of 10,000 carefully selected image-question pairs that provide a balanced representation of multimodal reasoning challenges and hallucination patterns. The following tables present the detailed breakdown across three key distributional characteristics.

Main Domain	Task Categories	Count
Attribute Recognition	→ States (wet/dry, clean/messy) → Colors, numbers, counts → Physical properties, actions	3,014
Visual Understanding	→ Object recognition → Existence questions → Visual challenges, illusions	2,984
Spatial Reasoning	→ Object relationships → Positioning, proximity → Spatial configurations	1,774
Knowledge & Identity	→ Celebrities, landmarks → Scenes, artwork, common knowledge	646
Math & Calculation	→ Numerical reasoning → Table interpretation	628
Text & OCR	→ Text reading, OCR → Translation, code understanding	514
General QA	→ General VQA	270
Temporal & Video	→ Video sequences → Temporal ordering	170
Total		**10,000**

Table 12: Domain type distribution showing main categories and their specific task subcategories.

Domain Type Distribution The domain distribution as seen in Table 12 reveals a balanced focus on

core visual-language capabilities, with **Attribute Recognition** and **Visual Understanding** together comprising nearly 60% of the dataset. This emphasis on fundamental visual perception tasks ensures robust evaluation of VLMs' basic competencies. The inclusion of specialized domains like **Temporal & Video** (170 samples) and **Math & Calculation** (628 samples) provides targeted assessment of advanced reasoning capabilities, while the substantial representation of **Spatial Reasoning** (1,774 samples) evaluates models' understanding of object relationships and positioning—a critical capability for hallucination detection.

Main Category	Specific Error Types	Count
Object-Related (3,493)	→ Object Hallucination	2,810
	→ Object Existence/Count	683
	→ Visual Challenge (Complex)	699
	→ Scene Hallucination	250
	→ Action Hallucination	123
Attribute-Related (1,622)	→ Attribute Hallucination (State)	1,169
	→ Attribute Hallucination (Number)	280
	→ Attribute Hallucination	163
	→ Attribute Hallucination (Color)	10
Relationship (1,720)	→ Relationship Hallucination	1,720
Other (3,165)	→ General VQA	670
	→ Reasoning/Calculation	407
	→ Unanswerable/Insufficient Context	355
	→ Identity Hallucination	340
	→ Temporal Hallucination	170
	→ Text/OCR Hallucination	151
Total		**10,000**

Table 13: Hallucination type distribution with grouped main categories and specific error subcategories.

Hallucination Type Distribution The hallucination taxonomy as seen in Table 13 demonstrates a comprehensive coverage of error patterns encountered in VLM outputs. **Object-Related** errors dominate with 3,493 instances, reflecting the fundamental challenge of object hallucination in vision-language models where fabricated objects (2,810 cases) represent the most common failure mode. The substantial representation of **Relationship** hallucinations (1,720 cases) highlights the difficulty models face in accurately describing spatial and logical relationships between visual elements. **Attribute-Related** errors (1,622 instances) predominantly involve state misattributions (1,169 cases), indicating models struggle with dynamic properties rather than static characteristics like color. The **Other** category encompasses diverse failure modes including reasoning errors and contextual misunderstandings, ensuring comprehensive evaluation coverage.

Answer Category	Definition	Count
Yes/No	→ Binary answers: yes, no → Simple true/false responses	6,568
Open-Ended	→ Descriptive text answers → Detailed explanations	2,008
Unanswerable	→ "Can't tell"/"Don't know"	731
Number	→ Numeric answers, counts → Quantitative responses	657
Selection	→ Multiple choice options → "The one on the right", etc.	36
Total		**10,000**

Table 14: Answer type distribution showing response categories and their specific characteristics.

Answer Type Distribution The answer format distribution as showcased in Table 14 reflects a pragmatic balance between evaluation complexity and real-world applicability. The predominance of **Yes/No** questions (6,568 samples, 65.7%) provides clear ground truth for binary classification while enabling efficient large-scale evaluation. **Open-Ended** responses (2,008 samples, 20.1%) evaluate free-form generation capabilities and complex reasoning, representing scenarios where models must provide detailed explanations. The inclusion of **Unanswerable** questions (731 samples, 7.3%) tests models' ability to recognize limitations and avoid overconfident responses—a critical capability for reducing hallucinations. **Numeric** responses (657 samples, 6.6%) focus on quantitative reasoning precision, while **Selection** tasks (36 samples, 0.4%) provide targeted evaluation of spatial and referential understanding.

This carefully balanced dataset composition enables robust evaluation of hallucination prediction across the full spectrum of VLM capabilities, ensuring that HALP's performance metrics reflect real-world deployment scenarios and diverse error patterns encountered in multimodal reasoning tasks.

C Model-Specific Hallucination Patterns

Table 15 reports the percentage of hallucinated outputs in four error categories Attribute-Related, Object-Related, Relationship, and Other for each of eight Vision–Language Models. Lower percentages represent stronger resistance to that error mode. Notably, Qwen2.5-VL-7B achieves the lowest attribute-related (4.87 %) and relationship (0.87 %) rates, while FastVLM-7B minimizes object-related hallucinations (5.81 %). The "Other" cate-

Hallucination Type	count	FastVLM-7B	Gemma3-12B	LLaVA-Next-8B	Llama-3.2-11B	Molmo-V1	Phi4-VL	Qwen2.5-VL-7B	SmolVLM2-2.2B
Attribute-Related	3493	7.58%	11.84%	19.73%	9.58%	14.30%	8.51%	**4.87%**	20.47%
Object-Related	1622	**5.81%**	10.42%	8.93%	7.27%	10.91%	6.80%	6.38%	8.33%
Relationship	1720	6.74%	5.12%	15.29%	4.91%	7.44%	3.73%	**0.87%**	2.62%
Other	3165	13.52%	13.08%	14.66%	14.46%	14.66%	16.85%	**9.13%**	10.21%

Table 15: Hallucination rates by basic hallucination type across eight Vision–Language Models. Bold values indicate the lowest hallucination rate per category.

gory is best constrained by Qwen2.5-VL-7B (9.13 %). These results highlight each model's relative strengths and weaknesses in pre-generation hallucination detection.

D Data Construction

We assemble the 10,000 sample evaluation set by sampling from six established VQA benchmarks. Figure 3(a)–(c) summarize its composition:

1. Domains (Figure 3(a)): Attribute Recognition (30.1%), Visual Understanding (29.8%), Spatial Reasoning (17.7%), Knowledge & Identity (6.5%), Math & Calculation (6.3%), Text & OCR (5.1%), General QA (2.7%), and Temporal & Video (1.7%).

2. Answer Formats (Figure 3(b)): Yes/No (65.7%), Open-Ended (20.1%), Unanswerable (7.3%), Numeric (6.6%), and Selection (0.4%).

3. Hallucination Categories (Figure 3(c)): Object-related (34.9%), Other errors (31.7%), Relationship-based (17.2%), and Attribute-related (16.2%).

These distributions ensure broad coverage of visual and reasoning challenges. We then employ GPT-4 as an LLM judge (Zheng et al., 2023) with lenient criteria: flagging hallucinations only when the model invents nonexistent entities, contradicts the ground truth, or provides entirely incorrect information, but not for accurate paraphrases or added detail. The resulting labels $b^j = \text{LLM}_{\text{judge}}(\hat{Y}^j, Y^j, Q^j)$ serve as targets for our probes (Section 5.5).

E LLM as a Judge

To obtain reliable hallucination labels, we employ GPT-4 as the judge model (Zheng et al., 2023). Formally,

$$b^j = \text{LLM}_{\text{judge}}(\hat{Y}^j, Y^j, Q^j),$$

where the judge compares the model-generated response $\hat{Y}^j$ against the ground truth Y^j in the context of query Q^j. We use a lenient flagging strategy:

- Flag as hallucinated ($b^j = 1$) only if:
  - The response mentions entities/objects not present in the image.
  - The response contradicts the ground truth factually.
  - The response provides completely incorrect information.
- Do *not* flag ($b^j = 0$) if:
  - The response paraphrases the ground truth accurately.
  - The response adds correct details beyond the ground truth.
  - The response uses different wording but conveys the same meaning.

The prompt template given to GPT-4 is:

```
You are an expert judge for vision-language
tasks.
Compare the generated answer:
"<generated_answer>"
against the ground truth: "<ground_truth>"
for the question: "<question>".

Use the following lenient criteria:
- Flag as hallucination (output "1") only if
the model invents nonexistent objects,
 contradicts the ground truth, or provides
  completely incorrect information.
- Otherwise, output "0".

Return only "0" or "1".
```

F Comparative Analysis: Gemma-3-12B vs. FastVLM-7B

Comparing Gemma-3-12B and FastVLM-7B (refer Figure 6) reveals contrasting architectural strategies for encoding hallucination signals. Gemma-3-12B demonstrates the canonical progression pattern, with query tokens showing dramatic improvement from Layer 1 (0.7165) to Layer L (0.9349),

while vision tokens actually decline from 0.6538 to 0.5956. This suggests that Gemma-3-12B concentrates hallucination detection capabilities in late-stage multimodal reasoning while sacrificing pure visual signal strength.

FastVLM-7B exhibits the opposite behavior: vision tokens maintain relatively strong performance (0.6935 $\rightarrow$ 0.7028) with optimal performance at Layer $L/2$ (0.7093), while query tokens show minimal improvement and even slight decline (0.6715 $\rightarrow$ 0.6136). This architectural divergence indicates that FastVLM-7B preserves visual hallucination signals throughout processing rather than concentrating them in final reasoning states, possibly reflecting different training objectives or fusion strategies.