Detecting Vision-Language Model Hallucinations before Generation

Anonymous ACL submission

Abstract

Object hallucination is a significant challenge that undermines the reliability of the Vision Language Model (VLM). Current methods for evaluating hallucination often require computationally expensive complete sequence generation, making rapid assessment or large-scale analysis difficult. We introduce HALP (HALlucination Prediction via Probing), a novel framework to efficiently estimate a VLM's propensity to hallucinate objects without requiring full caption generation. HALP trains a lightweight probe on internal VLM representations extracted after image processing but before autoregressive decoding. HALP offers a new paradigm for efficient evaluation of VLM, a better understanding of how VLMs internally represent information related to grounding and hallucination, and the potential for real-time assessment of hallucination risk.

1 Introduction

Vision-Language Models (VLMs) (Bordes et al., 2024) are transforming multimodal AI, they have demonstrated remarkable capabilities in understanding and generating language about visual scenes. However, their propensity to generate factually incorrect or "hallucinated" content, especially describing non-existent objects, is a major impediment to their reliability and trustworthiness in critical applications (e.g., medical, autonomous systems). This erodes user trust and can lead to harmful outcomes.

Current VLM evaluation, particularly for object hallucination, heavily relies on post-hoc analysis of fully generated outputs (Li et al., 2023). Other approaches focus on mitigating hallucination during or after generation or detecting it in generated text (Chen et al., 2024). While important, these don't address the need for efficient, pre-generative prediction of a model's likelihood to hallucinate for a given input. This limits rapid model iteration, large-scale analysis of internal states, and real-time risk assessment. There's a gap in methods that can forecast hallucination propensity from early signals within the model. The central hypothesis is that these internal VLM states may harbor predictive signals of potential object hallucinations even before a full caption is decoded.

We propose HALP (HALlucination Prediction via Probing), a framework to train lightweight "probes" directly on these internal VLM states. The goal is to efficiently predict whether a VLM is likely to hallucinate for a given image, and to what extent, without needing to generate the entire output sequence.

In this work, we introduce **HALP** (HALlucination Prediction via Probing):

- **State extraction**: We tap hidden activations at three key points in the captioning pipeline—(i) the end-of-image token (post-visual encoding), (ii) the end-of-query token (after multimodal fusion), and (iii) intermediate layers across the decoder.
- **Probe design**: We train simple linear classifiers or small MLPs on these activations to predict (a) a continuous hallucination severity score, (b) a binary hallucination flag, and (c) the likelihood of specific common objects being hallucinated.

Our experiments on COCO 2014 images data show that nearly all hallucination-predictive information is contained in the raw vision encoder output. On LLaVA-1.5, a 3-layer MLP probe trained solely on the pooled CLIP embedding achieves an MSE of 0.0455 for CHAIR$_i$ regression and a ROC–AUC of 0.75 for binary detection—both superior to probes built on the model's final multimodal fusion or query-conditioned embeddings (MSE $\geq$ 0.0509, AUC $\leq$ 0.665). We observe the same pattern on PaliGemma-2 (vision-only

MSE=0.0852, AUC=0.732), demonstrating that later decoder layers add noise rather than new signals for hallucination forecasting. These findings underscore the power of early vision representations for real-time hallucination risk assessment.

2 Background and Related Work

Object Hallucination in VLMs: Object hallucination occurs when a VLM describes objects absent from the visual input, undermining reliability in domains such as medical imaging or autonomous navigation. Such errors stem from mismatches between language priors and visual grounding, as well as annotation biases in training datasets. Mitigation strategies operate during or after generation, including 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). Post-hoc detection methods flag hallucinated mentions in generated captions. Evaluation predominantly uses the CHAIR$_i$ metric (Rohrbach et al., 2018), defined as the ratio of hallucinated object instances to all objects mentioned in a caption, and requires full output generation.

Probing: Probing is a diagnostic methodology wherein lightweight classifiers or regressors are trained on fixed internal activations of a neural network to test whether those activations encode specific properties. In NLP, probes have revealed that pretrained language models systematically encode part-of-speech tags, syntactic dependencies, and coreference relations at particular layers by training linear classifiers on token- or sentence-level hidden states (Hewitt and Liang, 2019; Marvin and Linzen, 2018). In computer vision, linear probes applied to convolutional activations reveal the spontaneous emergence of object detectors in scene-classification networks (Zhou et al., 2015), and have been used to systematically quantify unit interpretability by aligning individual hidden units with semantic concepts via Network Dissection (Bau et al., 2017). More recently, linear probing has been adopted to evaluate and analyze Vision Transformer representations, demonstrating that intermediate self-attention and MLP layers encode rich class-specific and scene-level semantics (Chen et al., 2022). These works illustrate that probing offers a lightweight yet powerful tool for charting where and how task-relevant features emerge in deep architectures, motivating our use of probes to detect hallucination signals early in VLM decoding.

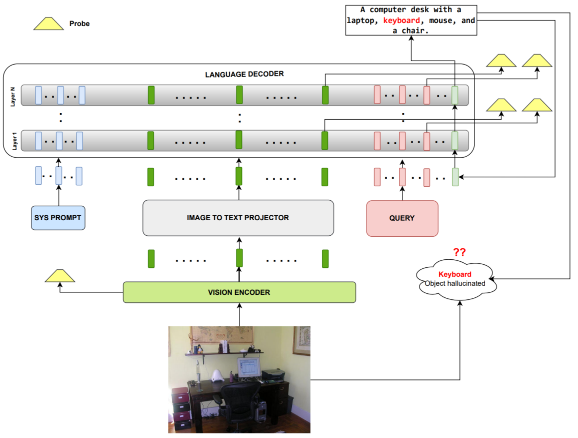

Figure 1: Overview of the HALP probing pipeline. An input image is first encoded into visual tokens by the vision encoder and projected into the language embedding space by the multimodal connector. These embeddings, together with a system prompt and task-specific query tokens, are fed into the LLM decoder. Hidden states are then extracted at three strategic positions—the end of the visual token sequence, the end of the query sequence, and selected intermediate decoder layers—and passed to lightweight probes that predict object hallucination before full caption generation.

3 HALP: HALlucination Prediction via Probing

3.1 Preliminaries: VLM Architecture

A vision–language model (Liu et al., 2023) consists of three core components in sequence: first, a vision encoder (Radford et al., 2021) decomposes the input image into a set of continuous feature vectors, or "visual tokens," capturing patch-level visual information; next, a multimodal connector which maps those visual tokens into the same embedding space as the language model, enabling joint reasoning over vision and text; finally, a Transformer-based LLM decoder (Team et al., 2024) consumes the fused visual embeddings—optionally alongside task-specific query tokens—and autoregressively generates the target text. It is precisely the hidden activations at various positions within this encoder–connector–decoder pipeline that we tap for our hallucination-prediction probes.

3.2 Post-Generation Hallucination

A caption is generated by processing an image along with a prompt in VLM. Once the caption

$\hat{c}_j$ is generated, the presence and degree of hallucination are assessed based on two key indicators:

Continuous metric (a_j): Defined as the proportion of objects in the generated caption that are considered hallucinated.

Binary indicator (b_j): A binary value indicating whether hallucination occurred (1) or not (0).

3.3 Extracting Internal Representations for Probing

To forecast hallucination risk before any tokens are generated, we extract three classes of vectors from a single forward pass of the VLM on image I_j:

1. **Global vision embedding** $\mathbf{e}_{v_j} \in R^d$: the pooled output of the vision encoder, which summarizes the primary visual features of I_j.

2. **Layer-wise fusion embeddings** $\mathbf{e}_{f_j}^{(\ell)} \in R^d$: for each selected decoder layer $\ell \in L$, we record the hidden state at the position immediately following the projected visual tokens. This vector captures how the model's attention mechanism has integrated image features with any preceding text (e.g., system prompts).

3. **Query-conditioned decoder states** $\mathbf{h}_{q_j}^{(\ell)} \in R^d$: from the same layers $\ell \in L$, we also extract the hidden state at the final query token—i.e. just before autoregressive generation begins—to capture the fused multimodal context that guides the forthcoming caption.

We then concatenate all of these—$\mathbf{e}_{v_j}$, $\{\mathbf{e}_{f_j}^{(\ell)}\}_{\ell \in L}$, and $\{\mathbf{h}_{q_j}^{(\ell)}\}_{\ell \in L}$—into a single feature vector x_j. Our lightweight probe is trained on $\{(x_j, y_j)\}$, where y_j is the ground-truth hallucination metric, enabling pre-generation prediction of hallucination propensity.

4 Experiments and Results

4.1 Experimental Setup

Dataset: We employ the COCO 2014 dataset (Lin et al., 2015). Each image is annotated with up to five human-written captions and instance segmentation masks covering 80 common object categories.

Models: We evaluate two state-of-the-art open-source VLMs

LLaVA-1.5 (Liu et al., 2024) combines a CLIP-based vision encoder with a Vicuna 1.5b language model, fused via a two-layer MLP projection and cross-attention module.

PaliGemma-2 (Steiner et al., 2024) integrates a SigLIP vision backbone with the Gemma 2 LLM through a learned cross-attention connector.

Post-Generation Hallucination Metrics: To measure object-level hallucination, we compute:

CHAIR$_i$: (Rohrbach et al., 2018) the ratio of hallucinated object mentions to all mentioned objects in a generated caption.

Binary indicator: a Boolean flag set to 1 if CHAIR$_i > 0$, indicating any hallucination.

Hidden-State Extraction Layers: We probe hidden states at five key decoder depths:

$$\ell \in \{0,\ 1,\ \lfloor N/2 \rfloor,\ N-2,\ N-1\},$$

where N is the total number of Transformer blocks in the LLM decoder. These layers capture early, middle, and late decoding dynamics to assess when hallucination signals emerge.

Probe Architectures We train two families of probes, each a 3-hidden-layer MLP:

- *Vision-only probe* P_v: input is the pooled vision embedding $\mathbf{e}_{v_j}$.

- *Layer-wise probes* P_ℓ for each $\ell \in L$: input is the concatenation of $\mathbf{e}_{v_j}$ and all fusion/query states up to decoder layer ℓ.

4.2 Results and Analysis

	Vision Embedding	Image Embedding (Layer N)	Query Embedding (Layer N)
		Regression (MSE)	
LLaVA	0.0455	0.0509	0.0523
PaliGemma	0.0852	0.8570	0.0840
		ROC–AUC	
LLaVA	0.750	0.632	0.500
PaliGemma	0.732	0.500	0.492

Table 1: Summary of probe performance on two VLMs. Top block: mean-squared error (MSE) for continuous CHAIR$_i$ regression; bottom block: ROC–AUC for binary hallucination detection. Layer N defined as last layer in the Language Model decoder.

Across both regression and classification tasks, the simplest "vision-only" probe—using only the pooled encoder output—consistently outperforms probes built on deeper decoder representations (refer table 2 and 3 in A.1). For the continuous hallucination severity prediction, the vision-only probe achieves a mean-squared error of 0.0455, whereas probes based on the end-of-image embeddings incur higher errors (0.0504–0.0523) and query-based probes perform no better (0.0523–0.0526). This indicates that the raw visual features alone capture nearly all of the information needed to estimate hallucination severity, and that adding successive layers of multimodal fusion slightly degrades regression accuracy.

Similarly, for binary hallucination detection, the vision-only probe attains the highest ROC-AUC (0.75). Image-level probes peak modestly at decoder layer 1 (AUC = 0.665) before declining to approximately 0.63 by layer 30, while query-based probes start near chance ($\approx$ 0.50 at layer 0), improve to around 0.62 at the mid-layers, then drop again. These trends show that although early fusion layers introduce some discriminative signal, they never surpass the straightforward vision embedding—and later decoding stages actually dilute it.

Table 1 shows that our findings generalize across two distinct VLM architectures. LLaVA achieves substantially lower regression error (MSE = 0.0455) and higher classification accuracy (ROC–AUC = 0.75) when using only the vision encoder's pooled embedding, compared to PaliGemma (MSE = 0.0852, AUC = 0.732). Probes built on the final decoder "image" embedding hurt performance for both models (LLaVA: MSE 0.0509, AUC 0.632; PaliGemma: MSE 0.8570, AUC 0.500), and query-embedding probes offer no benefit (LLaVA AUC = 0.50; PaliGemma AUC = 0.492). In other words, the pure vision-based probe is the most reliable across architectures, while deeper multimodal fusion and query conditioning consistently degrade both regression and classification performance. This reinforces the conclusion that the primary hallucination-predictive signal resides in the vision encoder outputs, with little to no incremental gain—and often added noise—from later decoder stages. In summary, almost all hallucination-predictive signal is already present in the vision encoder's output. Subsequent multimodal fusion and decoding layers introduce more noise than benefit for the specific task of pre-generation hallucination forecasting.

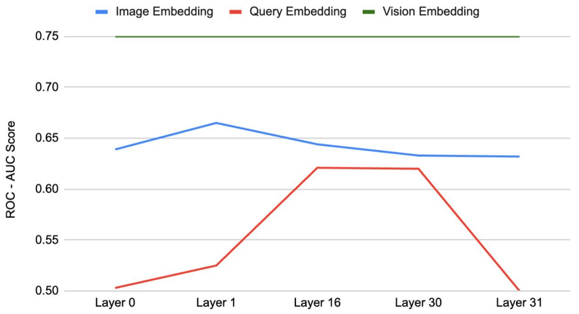

Figure 2: ROC–AUC for binary hallucination detection as a function of decoder layer. The green line shows the vision-only probe (constant at 0.75), the blue line shows probes built on the end-of-image embedding at each layer, and the red line shows probes on the end-of-query embedding.

5 Conclusion

We presented HALP, a lightweight probing framework for pre-generative prediction of object hallucination in vision–language models. By extracting global vision embeddings and layer-wise fusion representations from a single forward pass, and training simple MLP probes, we demonstrated that most hallucination-predictive information is already encoded in the vision encoder's outputs. Our experiments on LLaVA-1.5 and PaliGemma-2 show that a vision-only probe outperforms deeper, multimodal fusion-based probes in both continuous (MSE) and binary (ROC-AUC) hallucination prediction, highlighting the limited incremental value—and occasional noise—introduced by later decoder layers.

These findings have two main implications. First, they enable rapid, real-time hallucination risk assessment without expensive autoregressive decoding. Second, they suggest that future mitigation strategies might focus on refining the vision encoder's grounding signals rather than modifying the decoder. In future work, we plan to extend HALP to additional hallucination metrics (e.g. attribute or relation errors), evaluate its generalization across diverse VLM architectures and domains, and integrate probe outputs into decoding-time correction mechanisms for on-the-fly hallucination prevention.

Ethical Considerations

Our work focuses on detecting and predicting object hallucinations in vision–language models (VLMs) by probing internal representations. While HALP itself does not generate novel content, its deployment may influence downstream applications that rely on VLM outputs—for example, in healthcare, autonomous vehicles, or assistive technologies. An overly aggressive hallucination flag could result in false alarms, causing unnecessary intervention or eroding user trust, whereas an under-sensitive probe could fail to catch critical errors. We therefore advocate for human-in-the-loop validation in high-stakes domains and recommend threshold calibration based on application requirements. Additionally, our probe is trained on COCO data, which may contain demographic or cultural biases in image selection and caption annotations; these biases could propagate into hallucination predictions. We encourage future practitioners to evaluate HALP's performance on diverse, representative datasets and to apply bias-mitigation techniques when extending the framework to real-world systems.

Limitations

First, HALP's efficacy depends on the quality and diversity of the training set: we use COCO 2014, which covers a limited set of object categories and visual scenarios. Our continuous CHAIR$_i$ proxy and binary flag capture only object-level hallucinations and do not account for errors in attributes, relations, or higher-order semantics. Second, we evaluate on two open-source VLM architectures (LLaVA-1.5 and PaliGemma-2); results may not generalize to much larger or proprietary models with different fusion mechanisms or decoding strategies. Third, our probe requires access to intermediate hidden states, which may not be exposed by closed-source APIs or edge-deployed models. Finally, HALP predicts hallucination risk but does not itself correct or mitigate errors; integrating probe outputs into a feedback loop for on-the-fly correction remains future work.

References

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network dissection: Quantifying interpretability of deep visual representations. *Preprint*, arXiv:1704.05796.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, and 22 others. 2024. An introduction to vision-language modeling. *Preprint*, arXiv:2405.17247.

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.

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.

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.

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.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

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.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 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.

Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, Siyang Qin, Reeve Ingle, Emanuele Bugliarello, Sahar Kazemzadeh, Thomas Mesnard, Ibrahim Alabdulmohsin, Lucas Beyer, and Xiaohua Zhai. 2024. Paligemma 2: A family of versatile vlms for transfer. *Preprint*, arXiv:2412.03555.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, and 89 others. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

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

A Appendix

A.1 Results

Layer	Vision Embedding	Fusion Embedding	Query Embedding
Layer 0	0.0455	0.0504	0.0524
Layer 1		0.0513	0.0523
Layer 16		0.0523	0.0526
Layer 30		0.0523	0.0523
Layer 31		0.0509	0.0523

Table 2: Regression mean-squared error (MSE) for probes built on vision, fusion, and query embeddings at different decoder layers using LLaVA-v1.5-Vicuna-13b.

Layer	Vision Embedding	Fusion Embedding	Query Embedding
Layer 0	0.750	0.639	0.503
Layer 1		0.665	0.525
Layer 16		0.644	0.621
Layer 30		0.633	0.620
Layer 31		0.632	0.500

Table 3: ROC–AUC scores for binary hallucination detection probes on vision, fusion, and query embeddings at different decoder layers using LLaVA-v1.5-Vicuna-13b.

Layer	Vision Embedding	Image Embedding	Query Embedding
Layer 0	0.0852	0.089	0.0857
Layer 1		0.128	0.0849
Layer 16		0.857	0.0840

Table 4: Regression mean-squared error (MSE) for probes built on vision, fusion, and query embeddings at different decoder layers using PaliGemma-2.

Layer	Vision Embedding	Image Embedding	Query Embedding
Layer 0	0.732	0.508	0.500
Layer 1		0.488	0.500
Layer 16		0.500	0.492

Table 5: ROC–AUC scores for binary hallucination detection probes on vision, fusion, and query embeddings at different decoder layers using PaliGemma-2.