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), including LLaVA (Liu et al., 2024), PaliGemma (Steiner et al., 2024), Qwen (Bai et al., 2024), and BLIP (Li et al., 2022), are transforming multimodal AI, demonstrating impressive capabilities in both visual understanding and language generation grounded in images. However, object hallucination where models describe nonexistent objects remains a persistent challenge, particularly in high-stakes applications such as healthcare, autonomous systems, and assistive technologies. Such errors not only undermine user trust but can also result in harmful outcomes.

Existing methods for evaluating hallucinations rely heavily on post-hoc analysis of fully generated captions (Li et al., 2023), which is computationally expensive and ill-suited for real-time use. While other approaches attempt to mitigate hallucination during generation or detect it in the generated text (Chen et al., 2024), they cannot predict hallucination risk before decoding begins. This reveals a critical gap: there are few methods that assess a model's propensity to hallucinate based solely on its internal representations.

To address this, we introduce **HALP** (Hallucination Prediction via Probing), a lightweight probing framework that estimates hallucination risk using internal activations prior to caption generation. HALP extracts hidden states at three key stages in the captioning pipeline: after image encoding, after multimodal fusion, and across intermediate decoder layers. A Simple small MLPs is then trained on these representations to predict CHAIR_i (Continuous Hallucination severity), binary hallucination flags.

Our experiments on the COCO 2014 dataset (Lin et al., 2015) demonstrate that early vision encoder outputs are highly predictive of hallucination. For instance, a vision-only probe trained on LLaVA's pooled CLIP embedding achieves an MSE of 0.0455 and Hallucination Existence (AUROC) of 0.750, outperforming probes built on deeper decoder states (MSE $\geq$ 0.0509, AUC $\leq$ 0.665). Similarly, PaliGemma's performance declines as deeper layers are used (MSE rises to 0.8570). Notably, Qwen achieves its best performance at a deeper query layer (MSE 0.03730, AUROC 0.7611). Across all models, HALP enables accurate, real-time hallucination prediction before a single token is generated.

2 Background and Related Work

Object Hallucination in VLMs: Object hallucination occurs when a VLM describes objects not present in the visual input, reducing reliability in sensitive domains like medical imaging or autonomous navigation. These errors often stem from mismatches between language priors and visual grounding, or from annotation biases in training data. Mitigation strategies typically act 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 methods flag hallucinated mentions in generated captions. Evaluation commonly relies on the CHAIR_i metric (Rohrbach et al., 2018a), which measures the ratio of hallucinated object mentions to all mentioned objects and requires full caption generation.

Representation Probing Representation Probing is a diagnostic method where lightweight classifiers or regressors are trained on fixed internal activations to assess whether specific properties are encoded. In NLP, probes have shown that pretrained language models capture part-of-speech tags, syntactic dependencies, and coreference relations at particular layers (Hewitt and Liang, 2019; Marvin and Linzen, 2018). In computer vision, linear probes on convolutional features reveal emergent object detectors in scene-classification networks (Zhou et al., 2015), and techniques like Network Dissection align hidden units with semantic concepts to quantify interpretability (Bau et al., 2017). More recently, probing has been applied to Vision Transformers, where intermediate self-attention and MLP layers encode rich class-specific and scene-level semantics (Chen et al., 2022). These studies highlight probing as a lightweight yet effective tool for mapping task-relevant features in deep networks, motivating its use here to detect hallucination signals early in VLM decoding.

Hallucination Detection Recent work shows that vision and language models often encode hallucination-related signals in their internal states before generation. Orgad et al. (2023) trained probes on hidden activations to predict factual correctness, finding that the most informative signals are concentrated around answer tokens. Zhao et al. (2024) used a lightweight classifier on first-token logits to distinguish safe from unsafe prompts, suggesting models implicitly signal likely errors. Kadavath et al. (2022) introduced self-evaluation strategies, such as adding "none of the above" or using binary questions, to gauge model confidence. These findings support the hypothesis that hallucination cues exist in intermediate representations, motivating pre-generative detection frameworks like HALP.

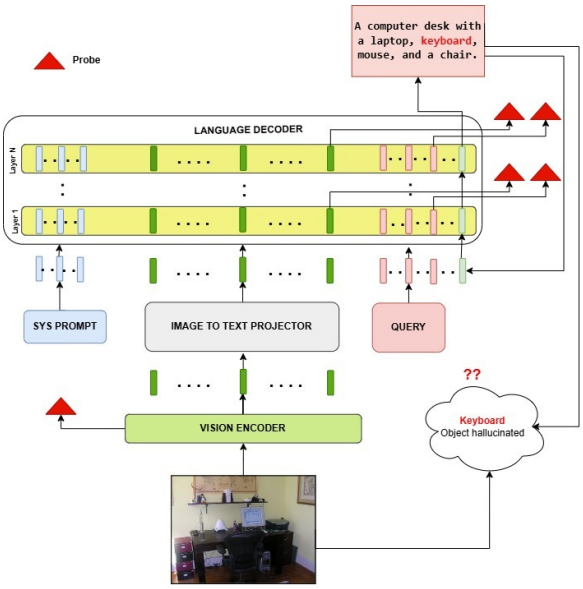

Figure 1: Overview of the HALP pipeline. Visual tokens from the input image are fused with prompt and last query token representations and passed to the decoder. Hidden states are extracted at key stages and used by probes to predict hallucination before caption generation.

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): Proportion of objects 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:

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

Layer-wise Last Vision Token Representations $\mathbf{e}_{f_j}^{(\ell)} \in R^d$: 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).

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

To evaluate HALP's effectiveness across diverse Vision-Language Models (VLMs), we utilized the COCO 2014 dataset (Lin et al., 2015), a widely adopted benchmark for object recognition and captioning.

Models Our experiments included a comprehensive set of state-of-the-art VLMs: LLaVA-v1.5-Vicuna-13b (Liu et al., 2024), PaliGemma-2 3B (Steiner et al., 2024), Qwen 2.5 VL 7B (Bai et al., 2024), and BLIP (Li et al., 2022). Hallucination Metrics: We employed CHAIRi (Rohrbach et al., 2018b) for evaluating object hallucination. This metric was framed as both a Regression task (predicting a continuous hallucination score) and a Binary Classification task (predicting the presence or absence of hallucination based on a threshold).

	Vision Only Representation	Last Vision Token Representation (Layer N)	Last Query Token Representation (Layer N)
Hallucination Degree (MSE) ↓			
LLaVA	0.045	0.050	0.052
PaliGemma	0.085	0.857	0.084
BLIP	0.039	0.043	0.746
Qwen	0.042	0.043	0.040
Hallucination Existence (AUROC) ↑			
LLaVA	0.750	0.632	0.500
PaliGemma	0.732	0.500	0.492
BLIP	0.515	0.500	0.500
Qwen	0.708	0.553	0.761

Table 1: Summary of probe performance on two VLMs. Top: mean-squared error (MSE) for CHAIR_i regression; bottom: ROC–AUC for binary hallucination detection. Layer N refers to the final layer of the language model decoder.

Embedding Extraction Internal VLM representations were extracted from the following key points within each model's architecture:

- **Vision Encoder:** The final output embedding of the VLM's dedicated vision encoder, representing the raw visual features.
- **LLM Decoder:** These were extracted at positions corresponding to the **end of the image token** (multimodal Last Vision Token Representation) and the **end of the query token** (query-conditioned decoder state). The selected layers were Layer 1, n/4, n/2, 3n/4, and n, allowing us to analyze information flow across the decoding process.

Probe Architecture The probe consists of a lightweight, feed-forward Neural Network (NN) with 3 hidden layers and respective dimensions of [1024, 512, 256]. The probes predicts the Hallucination metric (Binary and Continuous from CHAIR_i) by taking hidden LLM decoder states as input.

4.2 Results and Analysis

Our experiments evaluate the performance of lightweight probes on four distinct Vision-Language Models (VLMs): Qwen, BLIP, LLaVa, and PaliGemma. The probes are trained to predict object hallucination by tapping into internal model representations at various depths. Performance is measured using Mean Squared Error (MSE) for continuous hallucination score regression (lower is better) and ROC-AUC for binary hallucination classification (higher is better). The analysis is structured to compare performance across different decoder layers, models, and embedding types.

Layer-wise Analysis We evaluate hallucination existence prediction across five decoder layers (Layer 1, N/4, N/2, 3N/4, N) using AUROC scores (Figure 2). Across models, performance generally improves from the shallowest layer to intermediate

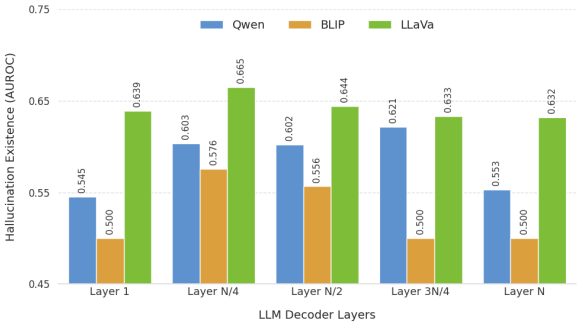

Figure 2: Hallucination Existence (AUROC) for binary hallucination detection. Layer-wise

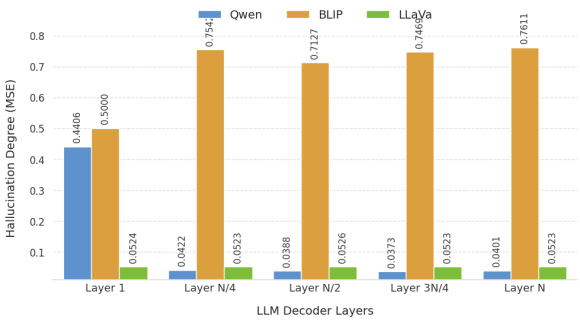

Figure 3: Hallucination Degree (MSE) for regression probes. Model Wise

depths. Layer N/4 consistently yields the highest AUROC scores peaking at 0.665 indicating that mid-level representations are most informative for hallucination detection. Beyond this point, performance plateaus or declines slightly, with minimal gains observed at deeper layers. Notably, early layers show weaker predictive capacity, often near chance level, suggesting that hallucination-relevant features emerge progressively and are most salient in the middle of the decoder.

Token Representation Analysis Figure 4 presents a token-wise comparison of hallucination prediction using image vs. query embeddings across decoder layers. We find that image embeddings consistently yield low MSE (0.038–0.043), while query embeddings produce high MSE (0.713), indicating that BLIP primarily encodes hallucination-irrelevant signals in the query path. This suggests a failure to effectively ground text generation in visual content, highlighting a core limitation in BLIP's architecture.

Model-Wise Hallucination Prediction Analysis We observe that BLIP consistently yields high MSE scores (0.7127), indicating poor modeling of hallucination signals. This underperformance likely stems from its earlier architecture and shallow vision language alignment, which limits its ability to ground text in visual content.

In contrast, Qwen and LLaVa both recent models achieve significantly lower MSE. Their superior performance can be attributed to deeper cross-modal integration, stronger supervision, and training on more diverse datasets. These results highlight the importance of modern architectures and robust alignment for effective hallucination modeling.

5 Conclusion

We presented HALP, a lightweight probing framework for pre-generative prediction of object hallucination in vision language models. By extracting global Vision Only Representations 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 QwenVL-2.5,LLaVA-1.5, BLIP and PaliGemma-2 show that a vision-only probe outperforms deeper, multimodal fusion-based probes in both Hallucination Degree (MSE) and Hallucination Existence (AUROC) prediction, highlighting the limited incremental value and occasional noise introduced by later decoder layers.

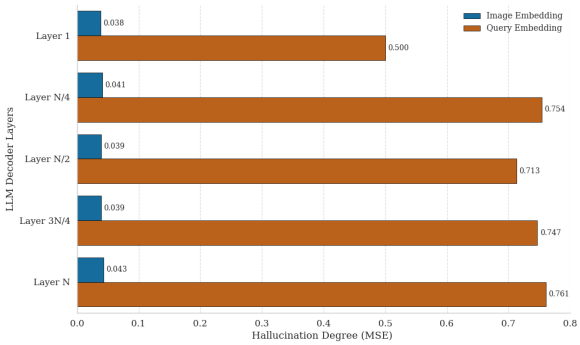

Figure 4: Image vs Query

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.

6 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 undersensitive 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.

7 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 only four open-source VLM architectures (Qwen, LLaVA-1.5, BLIP 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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2024. Qwen2.5-vl technical report.

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

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.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. Preprint. *Preprint*, arXiv:2207.05221. 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, 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.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2023. Llms know more than they show: On the intrinsic representation of llm hallucinations. Preprint. *Preprint*, arXiv:2312.06605. ArXiv:2312.06605.

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. 2018a. 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.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018b. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4035–4045. 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.

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. 2024. The first to know: How token distributions reveal hidden knowledge in large vision-language models. Preprint. *Preprint*, arXiv:2406.13259. ArXiv:2406.13259.

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

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 Mean-Squared Error (MSE) for the regression task and the AUROC score for the binary classification task. The results are organized by the type of internal representation used for the probe: the Vision Only Representation, Last Vision Token Representation and Last Query Token Representation at various decoder layers.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.0504	0.0524
Layer n/4		0.0513	0.0523
Layer n/2	0.0455	0.0523	0.0526
Layer 3n/4		0.0523	0.0523
Layer n		0.0509	0.0523

Table 2: Hallucination Degree (MSE) for probes built on vision only representation, last vision token representation and last query token representations at different decoder layers using LLaVA-v1.5-Vicuna-13b.

LLaVA-v1.5-Vicuna-13b The results for LLaVA highlight that the strongest predictive signal for hallucination is contained within the initial vision encoding. As shown in Tables 2 and 3, both regression and classification performance peak with the vision-only probe and degrade as more decoder-level context is added.

- Regression (Table 2): The vision embedding probe achieved the lowest MSE of 0.0455. The fusion and query embedding probes consistently resulted in higher errors across all layers.
- Classification (Table 3): The vision embedding probe achieved a superior ROC-AUC of 0.750. The fusion embedding probe's performance peaked at Layer 1 (0.665) and then declined, while the query embedding probe showed only modest predictive power in intermediate layers.

PaliGemma-2-2B Similar to LLaVA, the results for PaliGemma indicate that the raw visual features are the most reliable predictors of hallucination. Deeper representations consistently underperform compared to the simple vision-only probe.

- Regression (Table 4): The vision embedding probe yielded the best MSE of 0.0852. Image embedding MSE was significantly worse, especially at deeper layers, while query embedding MSE showed no improvement.
- Classification (Table 5): The vision embedding probe achieved a strong ROC-AUC of 0.732. Both image and query embedding probes performed poorly, with ROC-AUC scores hovering around the 0.5 mark, indicating they are no better than random chance.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.639	0.503
Layer n/4		0.665	0.525
Layer n/2	0.750	0.644	0.621
Layer 3n/4		0.633	0.620
Layer n		0.632	0.500

Table 3: Hallucination Existence (AUROC) scores for binary hallucination detection probes on vision only representation, last vision token representation and last query token representations at different decoder layers using LLaVA-v1.5-Vicuna-13b.

BLIP The results for BLIP present a more nuanced story. While the vision-only probe is not the top performer, the model's query embeddings show strong potential for regression, though its overall classification ability appears limited.

- Regression (Table 6): Probes on query embeddings consistently yielded the best performance, with an MSE as low as 0.0354 at layer 11. This is a notable improvement over the vision-only probe's MSE of 0.0397.
- Classification (Table 7): BLIP's classification performance was modest overall. The best ROC-AUC score was 0.622, achieved by a query embedding at layer 3. The vision-only probe was near chance at 0.515, suggesting the raw visual features in BLIP are not highly discriminative for this task.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.089	0.0857
Layer n/2	0.0852	0.128	0.0849
Layer n		0.857	0.0840

Table 4: Hallucination Degree (MSE) for probes built on vision only representation, last vision token representation and last query token representations at different decoder layers using PaliGemma-2.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.508	0.500
Layer n/2	0.732	0.488	0.500
Layer n		0.500	0.492

Table 5: Hallucination Existence (AUROC) scores for binary hallucination detection probes on vision only representation, last vision token representation and last query token representations at different decoder layers using PaliGemma-2.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.0504	0.0524
Layer n/4		0.0513	0.0523
Layer n/2	0.0455	0.0523	0.0526
Layer 3n/4		0.0523	0.0523
Layer n		0.0509	0.0523

Table 6: Hallucination Degree (MSE) for probes built on vision only representation, last vision token representation and last query token representations at different decoder layers using BLIP.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.639	0.503
Layer n/4		0.665	0.525
Layer n/2	0.750	0.644	0.621
Layer 3n/4		0.633	0.620
Layer n		0.632	0.500

Table 7: Hallucination Existence (AUROC) scores for binary hallucination detection probes on vision only representation, last vision token representation and last query token representations at different decoder layers using BLIP.

QwenVL-2.5-7B stands out as an architecture where query-conditioned decoder states provide a significant boost in predictive performance, surpassing the vision-only probe in both tasks.

- Regression (Table 8): The query embedding at layer 21 achieved the lowest MSE of 0.0373, outperforming the vision probe's MSE of 0.0422.
- Classification (Table 9): Performance steadily improved with query embedding probes at deeper layers, peaking with an ROC-AUC of 0.7611 at layer 27. This is a substantial improvement over the vision probe's 0.7083 and demonstrates the value of query conditioning in the Qwen model.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.04758	0.44061
Layer n/4		0.04452	0.04223
Layer n/2	0.04223	0.04306	0.03883
Layer 3n/4		0.04395	0.03730
Layer n		0.04368	0.04010

Table 8: Hallucination Degree (MSE) for probes built on vision only representation, last vision token representation and last query token representations at different decoder layers using Qwen.

Layer	Vision Only Representation	Last Vision Token Representation	Last Query Token Representation
	Vision Encoder	LLM Decoder	LLM Decoder
Layer 1		0.5449	0.5000
Layer n/4		0.6031	0.7542
Layer n/2	0.7083	0.6019	0.7127
Layer 3n/4		0.6213	0.7469
Layer n		0.5531	0.7611

Table 9: Hallucination Existence (AUROC) scores for binary hallucination detection probes on vision only representation, last vision token representation and last query token representations at different decoder layers using Qwen.

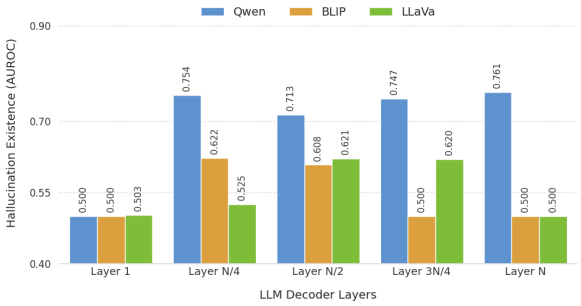

Figure 5: ROC–AUC for binary hallucination detection across LLM Decoder layers.

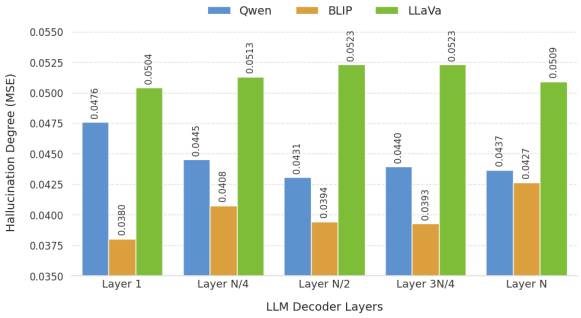

Figure 6: MSE for hallucination degree regression across LLM Decoder layers.

Our experimental results, depicted in Figures 5 through 9, provide insights into HALP's performance across various Vision-Language Models (VLMs) and probe configurations for hallucination detection.

Figure 5 : AUROC from Query Embeddings. This bar chart demonstrates that Qwen consistently achieves the highest AUROC scores (peaking at 0.761) using Query Embeddings, significantly outperforming BLIP (around 0.500) and LLaVA (peaking at 0.665). This indicates strong hallucination predictability from query-conditioned states in Qwen.

Figure 6 : Overall Performance from Vision-Only Probes. Figure presents two subplots evaluating performance using vision-only probes. The left subplot (AUROC) shows LLaVA (0.750) and PaliGemma (0.732) achieve the highest scores, with Qwen (0.7083) also performing well. BLIP

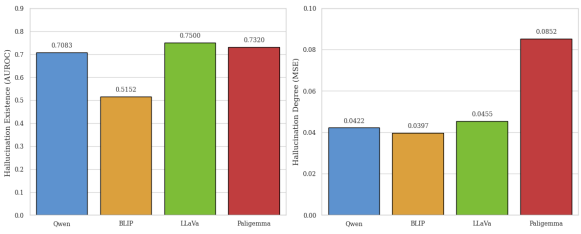

Figure 7: Left Subplot: ROC-AUC for binary hallucination detection using vision-only probes.

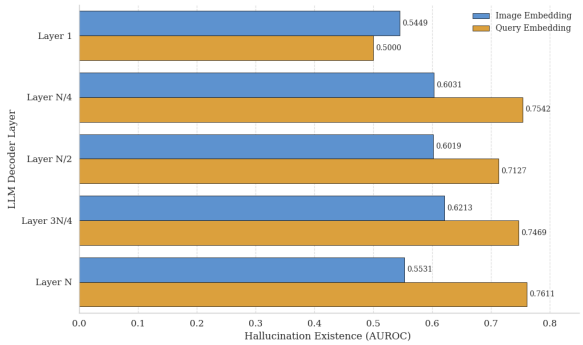

Figure 8: ROC-AUC for binary hallucination detection by LLM Decoder layer for the Qwen model.

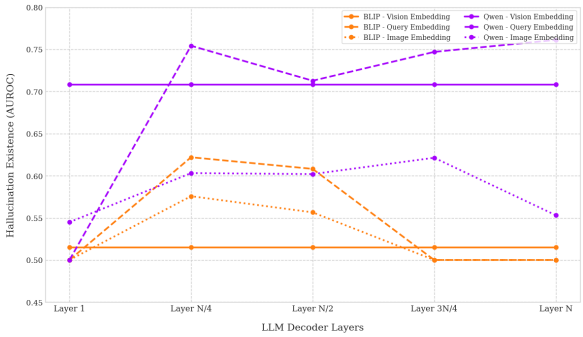

Figure 9: Comparison of AUROC performance for hallucination detection in BLIP and Qwen models, showcasing vision, query, and image embeddings across LLM decoder layers.

(0.5152) performs near chance. The right subplot (MSE) indicates BLIP (0.0397) has the lowest MSE in regression, but this is misleading as its poor ROC-AUC suggests limited discriminative power from its initial visual features.

Figure 7 : MSE from Image Embeddings. This figure illustrates the MSE for hallucination degree regression using Image Embeddings across layers. BLIP generally maintains low MSE values (e.g., 0.0380), while LLaVA shows higher MSE, suggesting varied precision in quantifying hallucination degree from image-conditioned representations across models.

Figure 8 : Qwen's AUROC: Image vs. Query Embeddings. Figure focuses on the Qwen model, directly comparing the AUROC performance between Image Embeddings (blue bars) and Query Embeddings (orange bars) across decoder layers. It clearly shows that Query Embeddings consistently and significantly outperform Image Embeddings for AUROC (peaking at 0.7611 vs. 0.6213). This highlights the dominance of query-conditioned information for hallucination existence prediction in Qwen.

Figure 9 : AUROC Trends for BLIP vs. Qwen. Figure provides a direct line plot comparison of AUROC trends between BLIP (orange lines) and Qwen (purple lines) across different embedding types and decoder layers. Qwen's Vision (solid purple, $\sim$0.708) and Query (dashed purple, peaking over 0.75) embeddings consistently outperform BLIP's (orange lines). BLIP's performance remains modest across all embedding types, reinforcing Qwen's superior ability to capture hallucination signals from diverse internal representations.