Detection-Guided Attention Steering for Vision Language Models

Alan William Zhang [1] **Rui Pan** [1] **Mike Wong** [1] **Ravi Netravali** [1]

Abstract

Modern Vision Language Models (VLMs) have demonstrated remarkable versatility in a wide range of applications, including multimodal reasoning, captioning, and analysis. However, VLMs still struggle with the fundamental tasks of object detection and localization, especially when compared to traditional Convolutional Neural Network (CNN) models. Contemporary VLMs tend to allocate very scarce attention to non-textual inputs such as image tokens and lack the inductive biases present in CNNs, making it difficult for them to identify challenging objects in complex scenes. To address these limitations, we propose a detection-guided dynamic attention steering system that leverages the locality insight from CNNs to efficiently steer a VLM's attention toward more relevant sections of an image. The steering intensity during each inference is dynamically scaled according to the confidence of CNN-generated bounding boxes. Extensive evaluations across multiple state-of-the-art VLMs show substantial improvements on object localization-oriented benchmarks, achieving up to a 4% accuracy gain. Results demonstrate the effectiveness of combining different model architectures to harness their respective strengths for advancing VLM capabilities.

1. Introduction

Vision Language Models (VLMs) have emerged as the state-of-the-art for solving multimodal tasks, ranging from complex reasoning to general visual perception. By leveraging the global semantic capacity of Vision Transformers (ViTs) (Dosovitskiy et al., 2021), VLMs can effectively capture long-range dependencies across flattened sequences of image tokens. Despite their success, VLMs still consistently struggle with detecting and localizing objects due to major structural disadvantages, especially when compared to traditional CNN-based models.

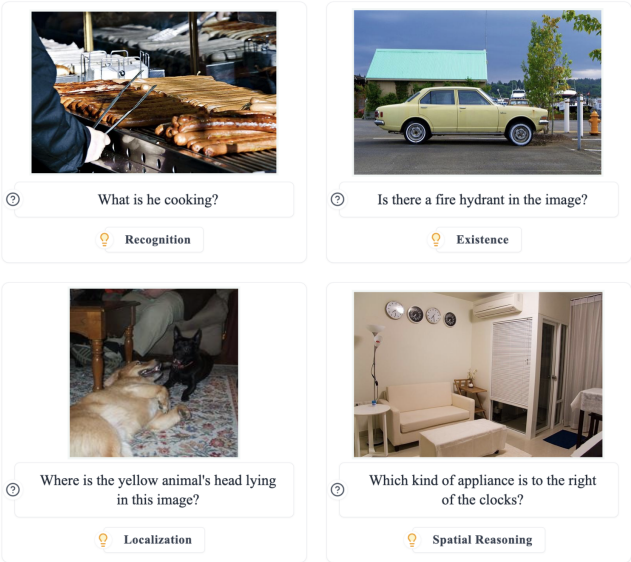
Numerous works have shown that VLMs exhibit a persistent tendency to heavily rely on language priors at inference time rather than concrete visual information (Selvaraju et al., 2019; Vo et al., 2026), resulting in hallucinatory behaviors and assumptions. For example, a VLM may hallucinate the color of an object based on its training distribution rather than grounding its response in the input image provided during inference. Furthermore, vision transformers pre-process 2D visual information into a flattened sequence of image tokens for compatibility with the natural language-based transformer architecture (Vaswani et al., 2017; Dosovitskiy et al., 2021). This causes vision transformers to lack the inductive biases (Cazenavette & Lucey, 2021; Kaya et al., 2025) that are fundamentally inherent to convolutional networks. These structural disadvantages make it exceedingly difficult for VLMs to localize and ground subtle objects hidden within complex or cluttered visual scenes.

To alleviate these architectural limitations, our proposed system introduces a detection-guided dynamic attention steering mechanism for VLMs. This methodology leverages the locality insights native to CNN-based models to guide a VLM's attention toward sections of an image that are most relevant to the text query; directly combating the over-reliance on language priors while also introducing the inductive biases of a convolutional network into a VLM's generation process. The set of target image tokens to steer attention towards will be determined by bounding boxes generated from an auxiliary CNN-based detector. Inspired by previous attention steering techniques used in language-only settings (Wang et al., 2024), our system selectively up-weights the attention scores of a set of target image tokens. The steering intensity applied for each inference is dynamically scaled by the confidence score of the bounding boxes, preventing our system from being excessively influenced by non-confident signals from the CNN detector.

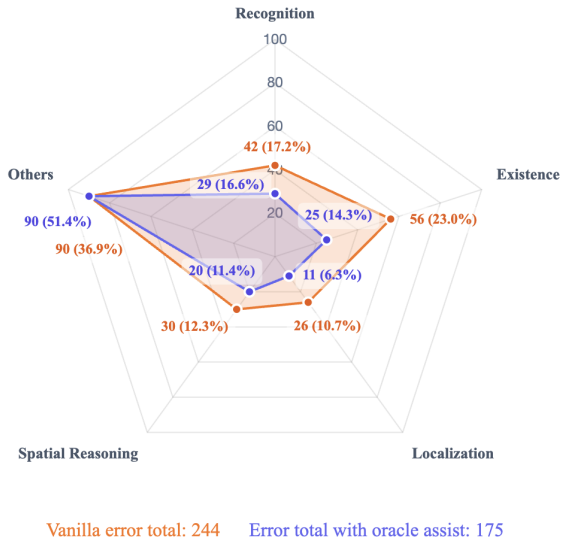
Evaluations conducted across multiple state-of-the-art VLMs, including Qwen3-VL (Bai et al., 2025), InternVL3 (Zhu et al., 2025), and LLaVA-NeXT (Liu et al.,

[1]Computer Science Department, Princeton University, Princeton, New Jersey, USA. Correspondence to: Alan W. Zhang <az0686@princeton.edu>.

Published as a workshop paper at SCALE - ICML 2026



(a) Example queries from each of the four object-centric categories.



(b) Error distribution by categories for vanilla run (orange) and oracle-assisted run (blue).

Figure 1. Error distribution by categories from evaluating a Qwen3-VL-8B model on a random subset of 1200 queries from four popular vision datasets: 300 queries each from POPE (Li et al., 2023), VQA-v2 (Goyal et al., 2017), MMVP (Tong et al., 2024), and GQA (Hudson & Manning, 2019). Object detection/localization centric categories are separated out from the 'Others' category.

2024), reveal substantial improvements on the POPE (Li et al., 2023) and MMVP (Tong et al., 2024) benchmarks, achieving up to a 4.0% accuracy gain for Qwen3-VL on MMVP. By introducing CNN guidance into a VLM's inference pipeline, our system strengthens the foundational visual capabilities of VLMs without requiring additional fine-tuning or retraining.

2. Motivation

2.1. Object Localization Challenges of VLMs

Modern vision language models (VLMs) are routinely evaluated on extensive downstream tasks, such as visual question answering (VQA), visual perception, multi-step spatial understanding, and reasoning. Similar to how humans process optical information, the successful execution of these higher-level cognitive tasks inherently requires the flawless execution of more fundamental, lower-level processes. These foundational processes often involve the precise classification and localization of important objects within a visual medium, acting as a critical prerequisite for upholding the integrity of any subsequent multimodal analysis (Pantazopoulos & Ozyiğit, 2025). However, empirical evidence consistently demonstrates that contemporary VLMs struggle with visual detection and localization.

A deeper investigation into these shortcomings reveals multiple systemic limitations in how VLMs process and utilize vision data during inference. Most notably, prior research identifies a tendency for VLMs to excessively rely on statistical correlations and language priors learned during training rather than grounding their generative responses in the explicit visual evidence presented at inference time (Selvaraju et al., 2019; Vo et al., 2026). For instance, if a VLM's training distribution contains a disproportionate number of images where a queried corn is yellow, the model will learn a strong inclination to generate "yellow" regardless of the corn's actual color in an input image at inference time. This phenomenon remains a significant concern when deploying VLMs in high-stakes, specialized domains such as autonomous driving and medicine (Xie et al., 2025; Wu et al., 2026). Furthermore, this over-reliance on language priors also heavily impairs a VLM's ability to detect expected objects and exacerbates hallucinations of unexpected objects (Pantazopoulos & Ozyiğit, 2025), necessitating external intervention mechanisms that forcibly re-weight internal attention prioritization during multimodal queries.

Recent comparative studies also attribute these localization failures to a fundamental lack of inductive biases in VLMs—such as translation invariance and spatial locality—which are much more apparent in traditional Convolutional Neural Networks (CNNs) (Cazenavette & Lucey, 2021; d'Ascoli et al., 2021; Kaya et al., 2025). For instance, convolutional operations implicitly assume the bias that adjacent pixels exhibit high statistical correlation, while VLMs predominantly utilize Vision Transformers (ViTs) to partition images into flattened sequences of patches, which are then fed sequentially into the prefill stage as image

tokens (Dosovitskiy et al., 2021). Empowered by the self-attention mechanism, this architectural design allows VLMs to model long-range semantic dependencies between all image tokens, regardless of their relative spatial distance. However, this global semantic capacity is achieved at the cost of sacrificing focus on fine-grained local patterns and inductive biases (Dosovitskiy et al., 2021; Liu et al., 2021; Wang et al., 2021). Consequently, VLM attention mechanisms are fundamentally optimized to capture global semantic relations rather than spatially localized features.

2.2. The Impact of Poor Localization

To empirically quantify the detrimental effect of poor localization on overall model accuracy, we evaluated the failure modes of a Qwen3-VL-8B model on a random subset of 1200 queries from four popular vision datasets (POPE, VQA-v2, MMVP, GQA). Figure 1a shows example queries from four main object-centric question categories present in the random subset. After analyzing the results of a vanilla run, we noticed that a significant number of failed queries (63.1%) belonged to one of the object-centric categories; a breakdown is shown in the orange plot of Figure 1b.

To demonstrate the substantial accuracy gains achievable through correcting object-centric errors, for each error case in the four object-centric categories, we appended auxiliary cropped "oracle" images of relevant objects to the query to artificially assist the VLM in accurately detecting and localizing the target object. The system prompt was also appended with: "The second image shows a cropped version of the relevant object". After introducing these changes, the blue plot of Figure 1b shows the number of total errors dropping from 244 to 175, meaning nearly 30% of all baseline failures were corrected. These results show how refining a VLM's object detection and localization capabilities can directly yield non-trivial gains in its overall evaluation accuracy.

2.3. The Potential of CNN-Driven Signals

To further highlight the limitations of VLMs in spatial tasks relative to traditional CNNs, we analyzed the object localization recall rates (the proportion of identified true positives over all true positives) of both model architectures on the RefCOCO-M dataset (Liu et al., 2025). Consisting of granular queries derived from the original RefCOCO dataset (Yu et al., 2016), RefCOCO-M provides higher-fidelity object masks and polygons. We extracted a statistically representative sample of 1,000 visual scenes from the validation split.

For each query, a Qwen3-VL-8B model (VLM) and a YOLOE (Wang et al., 2025) detection/segmentation model (CNN) were provided with an input image and a referring expression (e.g., "oven on top near the cabinet") and were required to output the bounding box coordinates of the object referred to in the expression. The two models were evaluated using the standard Intersection-over-Union (IoU) metric with a threshold of 0.9 – if the outputted bounding box scores an IoU below 0.9, then we consider the object to not be correctly localized.

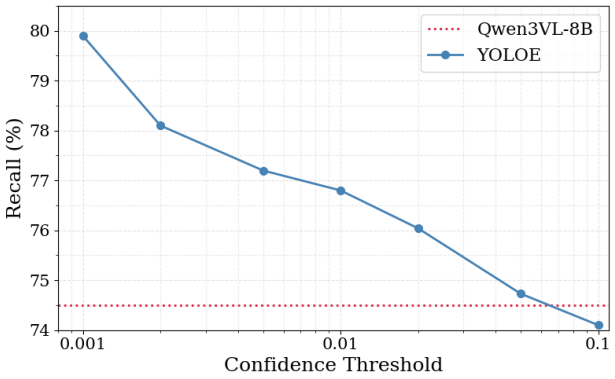


Figure 2. Comparison of recall rate on object localization between Qwen3-VL-8B and YOLOE with a varying confidence threshold on a 1000-image random sample of RefCOCO-M. YOLOE provides a dynamic, tunable recall curve that achieves near 80% at low thresholds, whereas Qwen3-VL-8B maintains a fixed recall of 74.50%.

Since CNNs like YOLOE generally provide a configurable confidence threshold to filter bounding boxes, it is possible to achieve higher recall by lowering this threshold. Figure 2 shows that at moderate thresholds (0.1 to 0.05), YOLOE achieves recall rates that are comparable to Qwen3-VL-8B. However, as the threshold is continually lowered, YOLOE experiences a substantial increase in recall. This structural flexibility allows users to configure CNNs to maximize specific metrics. In the context of maximizing recall, YOLOE can be uniquely customized to identify challenging objects tied to complex referral instructions or text queries. In contrast, VLMs like Qwen3-VL remain bound by a static recall of 74.50%. Ultimately, due to their independence from language priors, stronger inductive biases, and architectural flexibility, we consistently see CNNs exhibiting better object localization signals than contemporary VLMs.

3. Design

3.1. Attention Steering

To introduce external insights and signals into existing models without incurring considerable computational overhead, recent works have preferred inference-time intervention strategies such as attention steering as an alternative to fine-tuning and retraining. Similar to existing works on non-multimodal settings (Zhang et al., 2024; Venkateswaran &

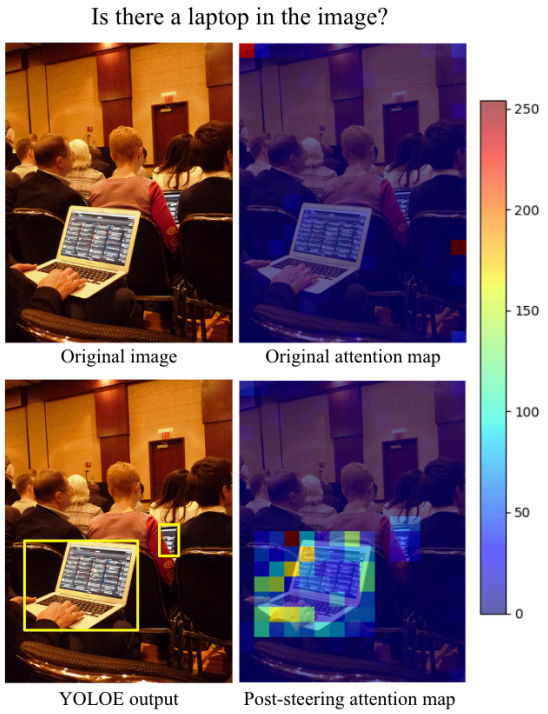


Figure 3. An example visualizing how the attention scores of target tokens are upscaled. The top/bottom right images overlay the attention scores before/after steering, respectively. Each small colored square is an image patch corresponding to one image token. The noun 'laptop' is extracted from the query and detected in the image. All detections are outlined with yellow bounding boxes in the bottom left image.

Contractor, 2026), we define attention steering as the deliberate manipulation or scaling of attention scores for a set of target tokens at inference time. Our system applies a similar intervention strategy to multimodal queries and is motivated by observations from previous work indicating that only a disproportionately low amount of attention is distributed to image tokens containing critical objects of interest (Kang et al., 2025; Tang et al., 2025). Existing research has also attempted to mitigate attention deficits by indiscriminately redistributing attention from visual sink tokens (image tokens with very high attention scores) to non-sink tokens (Kang et al., 2025); our approach improves upon this by utilizing signals from CNNs to better identify more granular steering destinations.

3.2. Integrating CNN Signals

Our system uses bounding boxes generated by an open-vocabulary CNN as spatial signals to identify the set of target image tokens to steer towards. First, important noun phrases are extracted from the text query using the Spacy natural language processing pipeline (Honnibal & Montani, 2017). These noun phrases define the detection classes of an open-vocabulary CNN-based detector. We choose to detect noun phrases instead of isolated nouns or noun lemmas after observing more accurate localization results from open-vocabulary detectors like YOLOE when incorporating adjectives and conditions into the detection classes, such as "red sign" and "jumping man". The CNN detector returns a list of bounding boxes (each associated with a confidence score) when invoked. As shown in Figure 3, the coordinates of each bounding box are then used to determine which image patches overlap with the bounding box. Since each image patch is translated into an image token, the set of target image tokens to steer towards for a particular query is determined by all image tokens that partially or completely overlap with at least one bounding box returned by the CNN detector.

Our system uniformly scales down the attention scores of all non-target image tokens by a steering factor s, where $0 < s \leq 1$. By exploiting the standard downstream normalization procedure, this targeted suppression effectively scales up the attention scores of all target tokens by a proportional factor of $1/s$. Hence, a smaller steering factor corresponds to more aggressive steering, and using $s = 1$ is equivalent to an unmodified inference pass. All non-image tokens, including system and text query tokens, are also included in the set of target tokens to prevent them from being erroneously suppressed by the steering factor. Scaling down non-target image tokens is chosen over scaling up target tokens since the former tends to be more numerically stable while also simplifying the system by leaving the attention scores of non-image tokens untouched.

3.3. Dynamic Confidence-Aware Steering

A critical challenge observed in our experiments when relying on CNN-derived signals is that lowering the confidence threshold of the CNN detector leads to a superlinear proliferation of generated bounding boxes, inevitably increasing the frequency of hallucinated or inaccurate bounding boxes. To counteract this degradation, we introduce a dynamic, per-query steering mechanism that keeps the steering intensity inversely proportional to the confidence scores of CNN-generated bounding boxes. As the confidence score of a proposed bounding box decreases, the original base steering factor s_{base} is exponentially increased via an offset Δs to attenuate the overall steering intensity, thereby mitigating the influence of potentially hallucinated signals. Given a bounding box confidence score $c \in (0, 1]$, the scaling equation for the offset is defined as:

$$\Delta s = 0.9e^{-40c} \tag{1}$$

In practice, our system begins detection with a predefined minimum confidence threshold and accepts the bounding boxes with the highest confidence score. This process is repeated for every detection class (noun phrase). This

selection strategy is chosen after we observed that correct bounding boxes routinely had higher confidence scores than incorrect boxes, even if both were well below the default confidence threshold (0.25 for YOLOE), suggesting that more accurate localization signals will be considered first before inaccurate or hallucinated signals.

Finally, the system adds an offset Δs (calculated using equation 1) to the base steering factor to lower the steering intensity for queries with less confident bounding boxes. The final effective steering factor s that is used is clamped such that $s = \min(1, s_{base} + \Delta s)$. This design ensures that any inaccuracies from the auxiliary CNN are not inadvertently translated into the VLM's final output.

3.4. Retaining VLM Signals and Context

Despite relying on a CNN-based detector for localization, our system still considers the VLM's existing normalized attention distribution when determining the steering targets. To better preserve relevant context undetected by the CNN but emphasized by the VLM, if the VLM is already allocating a substantial amount of attention to a non-trivial region of the input image, all image tokens in that region will be included in the list of target tokens to prevent them from being suppressed.

For each layer in the VLM, we first use Otsu's Method (Otsu et al., 1979) to determine an optimal attention threshold that separates tokens with high attention scores from tokens with low attention scores in that layer. Image tokens that have drastically high normalized attention scores (greater than 0.9, such as sink tokens (Kang et al., 2025)) will severely skew the threshold and are removed from consideration when calculating this threshold. Then, for any cluster of adjacent image tokens that are all above the attention threshold, if that cluster constitutes at least 10% of all image tokens, all tokens in that cluster are added to the list of target image tokens. This percentage can be lowered or increased to capture context with higher recall or precision, respectively.

Hence, in addition to signals from a CNN detector, our system also leverages attention signals from the VLM to verify whether any context is already being attended to. This ensures that any object-centric steering does not obliviously reallocate an excessive amount of attention from parts of an image that provide important context for the VLM to correctly answer a query. A potential enhancement for the future would be to introduce tiered steering intensities and prioritization, such as the ability to choose between steering more towards objects or attention-backed context.

4. Implementation

Our system uses a VLM and a steering module that consists of the Spacy *en_core_web_lg* natural language processing pipeline (Honnibal & Montani, 2017) and YOLOE-11L (Wang et al., 2025), a CNN-based open-vocabulary object detection and segmentation model. The system has a configurable base steering factor of 0.01 and a minimum YOLOE confidence threshold of 0.01. Prior work observed that distinct attention heads and layers within a model often exhibit a disproportionate influence on spatial detection capabilities (Wang et al., 2024); we designed our system with this flexibility in mind and allow for targeted attention steering to specific VLM layers and heads.

5. Evaluation

5.1. Experimental Settings

We evaluated our system using Qwen3-VL-8B (Bai et al., 2025), InternVL3-8B (Zhu et al., 2025), and Llava-NexT-Mistral-7B (Liu et al., 2024) as the VLM backbone, and YOLOE-11-L (∼67MB) as the CNN detector (Wang et al., 2025) on the POPE and MMVP datasets. Table 1 reports the accuracies and F1 scores of each model and compares our system with the baseline. For all experiments, we set the base steering factor to 0.01. The POPE benchmark contains three splits that use different sampling strategies (random, popular, adversarial) to thoroughly test for specific hallucination patterns. The MMVP benchmark consists of pairs of closely related questions and evaluates accuracy in a pairwise manner; a pair of queries is considered correct only if both questions within the pair are answered correctly. Hence, the random-guess baseline accuracy of the MMVP dataset is 25%.

5.2. Results

As shown in Table 1, our system consistently yields performance improvements across all models, demonstrating the value of introducing CNN-based spatial insights into transformer-based vision language models. Most notably on POPE, InternVL3-8B achieves a 2.81% increase in accuracy and a 3.26% increase in F1 score on the random sampling split; on MMVP, Qwen3-VL-8B achieves a 4% increase in accuracy.

In the vanilla runs on the POPE dataset, the majority of errors stemmed from the failure to detect and localize an expected object that was present in the image query (false negative). After our system applies confidence-aware steering, the false negative rate drops substantially; this improvement in recall accounts for the bulk of the observed accuracy and F1 score improvements on the POPE benchmark. An increase in false positives (hallucinations

Table 1. Evaluation results on three similarly sized VLM models of different families. POPE^{R,P,A} denote the random, popular, and adversarial sampling splits for POPE, respectively. Our system improves accuracy of all models on all datasets by up to 4%.

Model	POPE^R		POPE^P		POPE^A		MMVP
	Acc	F1	Acc	F1	Acc	F1	Acc
Qwen3-VL-8B	90.50	89.67	88.33	87.62	87.07	86.45	54.00
+ Ours	**92.93**	**92.53**	**88.80**	**88.50**	**87.30**	**86.97**	**58.00**
InternVL3-8B	87.06	86.36	84.73	84.26	82.53	82.29	29.33
+ Ours	**89.87**	**89.62**	**87.10**	**87.16**	**84.60**	**85.02**	**31.33**
Llava-NeXT-Mistral-7B	89.71	88.89	86.83	85.94	85.40	83.89	26.67
+ Ours	**90.01**	**89.15**	**87.23**	**86.42**	**85.93**	**84.56**	**29.33**

Table 2. Ablation study on Confidence-Aware Steering. Results show how Confidence-Aware Steering account for a significant portion of accuracy gains.

Model	POPE^R		POPE^P		POPE^A		MMVP
	Acc	F1	Acc	F1	Acc	F1	Acc
Qwen3-VL-8B	90.50	89.67	88.33	87.62	87.07	86.45	54.00
+ Confidence-Agnostic Steering	90.93	90.22	88.47	87.23	86.73	85.01	54.67
+ Confidence-Aware Steering (Ours)	**92.93**	**92.53**	**88.80**	**88.50**	**87.30**	**86.97**	**58.00**

of non-existent objects) is also observed, but only to a marginal degree that is eclipsed by the gains in recall. This favorable trade-off serves as a strong testament to the system's robustness against hallucinated or low-confidence bounding boxes in more difficult queries.

As mentioned, the MMVP benchmark consists of contextually similar query pairs and uses a strict pairwise evaluation structure. The increase in accuracy across all models on this benchmark can be attributed to the effectiveness of our system in resolving failure cases. In particular, being able to correctly identify and localize an object allows our system to correct both questions in a previously failed pair. Consequently, instances where our system only corrected one query within a failed pair (thereby yielding no accuracy improvement) were extremely rare. Ultimately, these experimental results validate the effectiveness of our CNN-based steering system in improving the ability of VLMs to localize difficult objects without severely degrading their baseline resistance to hallucinations.

5.3. Ablations and Analysis

To validate the necessity of our system's dynamic confidence-aware steering mechanism, we conduct an ablation study to compare our system's performance with and without confidence-aware steering. Table 2 shows that without this mechanism, the accuracy and F1 score improvements significantly diminish across all datasets. On POPE^A, the lack of confidence-aware steering decreases the accuracy and F1 score even below the baseline. This result further highlights how essential and effective it is for our system to incorporate a robust mechanism for tackling object hallucination.

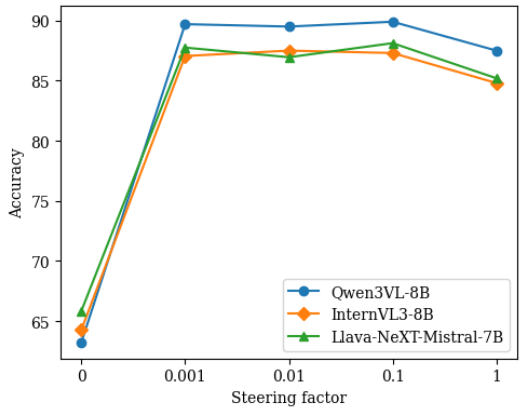


Figure 4. Accuracy of varying base steering factors on a random subset of POPE. Using a steering factor of 1 is equivalent to a vanilla run. Results show our system's robustness against changes to the steering factor.

Figure 4 shows the accuracy of our system on a random subset of 300 queries (100 from each of the three POPE splits) while changing the base steering factor for each VLM. Using a steering factor of 1 is equivalent to a vanilla run without our system. Results demonstrate the robustness of our system to changes in the base steering factor, with minimal accuracy differences observed when scaling the steering factor up or down from 0.01. All three variations still consistently performed better than the baseline. Notably, severe accuracy degradations are observed when the steering factor is set to 0 since all non-target image tokens have their attention scores zeroed out; effectively removing the vast majority of visual information from the decode stage.

6. Related Work

Vision Language Models. Recent advancements in multimodal architectures have established vision language models (VLMs) as the state-of-the-art for tackling generalized visual perception and reasoning tasks. Central to these

modern VLMs is the use of pre-trained Vision Transformers (ViTs) (Dosovitskiy et al., 2021) mostly derived from CLIP (Radford et al., 2021) or SigLIP (Zhai et al., 2023), which process images into localized patches that are each represented by an image token. Major VLM model families such as Qwen-VL, LLaVA, and InternVL (Bai et al., 2023; Liu et al., 2023; Chen et al., 2024) leverage self-attention mechanisms over flattened sequences of image tokens to construct holistic semantic dependencies between textual and visual inputs (Vaswani et al., 2017; Dosovitskiy et al., 2021). In order to maintain spatial awareness, VLMs use techniques like 2D RoPE (Heo et al., 2024) and Spiral RoPE (Liu et al., 2026) to build positional representations and preserve locality information within visual inputs.

Open-Vocabulary CNN Detectors. Historically, the utility of efficient CNN object detectors such as the You Only Look Once (YOLO) series (Redmon et al., 2016) was constrained by their reliance on predefined, closed-set object categories, as they were typically trained on a finite set of classes like the COCO dataset categories (Lin et al., 2014). However, recent breakthroughs in vision-language pre-training have fueled the transition to open-vocabulary detection. Frameworks like YOLO-World (Cheng et al., 2024) and GLIP (Li et al., 2022) bridge this gap by incorporating Re-parameterizable Vision-Language Path Aggregation Networks (Cheng et al., 2024) and spatial-aware contrastive learning (Radford et al., 2021; Li et al., 2022), seamlessly aligning visual representations with open-ended textual embeddings. This new advancement allows CNNs to identify objects using arbitrary natural language phrases in a zero-shot manner.

Attention Steering. A growing body of literature focuses on using inference-time interventions as an efficient way to modulate model behavior without the computationally prohibitive costs of fine-tuning or retraining. In particular, many works have attempted to manipulate or steer internal activations and attention scores to enhance VLMs. For instance, recent studies have explored reallocating attention that is over-saturated in specific tokens to improve focus on the rest of the image (Kang et al., 2025). Other systems sharpen attention onto image regions with higher confidence while widening the attention window to capture more context when confidence is low (Chen et al., 2025), or employ attention-guided image warping to allocate more resolution to image regions that are more query-relevant (Dalal et al., 2025).

7. Conclusion

In this work, we propose a detection-based dynamic attention steering system that addresses the persistent limitations of Vision-Language Models (VLMs) in foundational object detection and localization tasks. Our framework effectively leverages the locality insights of an auxiliary open-vocabulary CNN detector to guide a VLM's attention toward the image sections most relevant to a given text query. To mitigate hallucinations, our system dynamically scales the steering intensity for each query based on confidence signals from the CNN detector. Experimental results reveal substantial improvements in evaluation accuracy and a VLM's ability to detect and localize objects. This work demonstrates the profound effectiveness of combining distinct model architectures to leverage their respective spatial and semantic strengths, paving a path forward for advancing additional capabilities of VLMs in the future.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

LLM/Agent Usage Disclosure

LLMs were used in the process of proofreading and refining this paper.

References

Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., and Zhou, J. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Cazenavette, G. and Lucey, S. On the bias against inductive biases, 2021. URL https://arxiv.org/abs/2105.14077.

Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J. C., Geva, M., He, J., Wu, J., and Li, M. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas, 2025. URL https://arxiv.org/abs/2503.01773.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., Li, B., Luo, P., Lu, T., Qiao, Y., and Dai, J. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://arxiv.org/abs/2312.14238.

Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., and Shan, Y. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16901–16911, 2024. URL https://arxiv.org/abs/2401.17270.

Dalal, D., Vashishtha, G., Mishra, U., Kim, J., Kanda, M., Ha, H., Lazebnik, S., Ji, H., and Jain, U. Constructive distortion: Improving MLLMs with attention-guided image warping, 2025. URL https://arxiv.org/abs/2510.09741.

d'Ascoli, S., Touvron, H., Leavitt, M. L., Morcos, A. S., Biroli, G., and Sagun, L. ConViT: Improving vision transformers with soft convolutional inductive biases. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2286–2296. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/d-ascoli21a.html.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2010.11929.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6904–6913, 2017. URL https://arxiv.org/abs/1612.00837.

Heo, B., Park, S., Han, D., and Yun, S. Rotary position embedding for vision transformer. In *European Conference on Computer Vision (ECCV)*. Springer, 2024. URL https://arxiv.org/abs/2403.13298.

Honnibal, M. and Montani, I. *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing*, 2017. URL https://spacy.io/.

Hudson, D. A. and Manning, C. D. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019. URL https://arxiv.org/abs/1902.09506.

Kang, S., Kim, J., Kim, J., and Hwang, S. J. See what you are told: Visual attention sink in large multimodal models, 2025. URL https://arxiv.org/abs/2503.03321.

Kaya, A., Bilik, I., and Stainvas, I. A comparative study of vision transformers and cnns for few-shot rigid transformation and fundamental matrix estimation, 2025. URL https://arxiv.org/abs/2510.04794.

Li, L. H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.-N., Chang, K.-W., and Gao, J. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10965–10975, 2022. URL https://arxiv.org/abs/2112.03857.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://arxiv.org/abs/2305.10355.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014. URL https://arxiv.org/abs/1405.0312.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2304.08485.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, H., Ren, S., Zhu, T., Wang, P., Xie, C., Yuille, A., Zheng, Z., and Wang, F. Spiral RoPE: Rotate your rotary positional embeddings in the 2d plane, 2026. URL https://arxiv.org/abs/2602.03227.

Liu, J., Wang, W., Zhang, Y., Tang, Y., He, X., Guo, L., Yue, T., and Wang, X. Towards unified referring expression segmentation across omni-level visual target granularities, 2025. URL https://arxiv.org/abs/2504.01954.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer*

Vision (ICCV), pp. 10012–10022, 2021. URL https://arxiv.org/abs/2103.14030.

Otsu, N. et al. A threshold selection method from gray-level histograms. *Automatica*, 11(285-296), 1979.

Pantazopoulos, G. and Ozyiğit, E. B. Towards understanding visual grounding in visual language models. *arXiv preprint arXiv:2509.10345*, 2025. URL https://arxiv.org/abs/2509.10345.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. URL https://arxiv.org/abs/1506.02640.

Selvaraju, R. R., Lee, S., Shen, Y., Jin, H., Ghosh, S., Heck, L., Batra, D., and Parikh, D. Taking a hint: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6091–6100, 2019. doi: 10.1109/ICCV.2019.00619. URL https://arxiv.org/abs/1902.03751.

Tang, F., Huang, Z., Liu, C., Sun, Q., Yang, H., and Lim, S.-N. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4537e2172096d966228d68c428392894-Paper-Conference.pdf.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. URL https://arxiv.org/abs/2401.06209.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03762.

Venkateswaran, P. and Contractor, D. Spotlight your instructions: Instruction-following with dynamic attention steering. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3752–3770, 2026. URL https://arxiv.org/abs/2505.12025.

Vo, A., Nguyen, K.-N., Taesiri, M. R., Dang, V. T., Nguyen, A. T., and Kim, D. Vision language models are biased. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2505.23941.

Wang, A., Liu, L., Chen, H., Lin, Z., Han, J., and Ding, G. Yoloe: Real-time seeing anything, 2025. URL https://arxiv.org/abs/2503.07465.

Wang, W., Xie, E., Li, X., Fan, D.-P., Song, K., Liang, D., Lu, T., Luo, P., and Shao, L. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 568–578, 2021. URL https://arxiv.org/abs/2102.12122.

Wang, Z., Zhang, M., Xu, R., Ye, J., Wang, Y.-F., Qiao, Y., and Duan, Z. Tell your model where to attend: Post-hoc attention steering for llms, 2024. URL https://arxiv.org/abs/2311.02262.

Wu, Y., Yang, Z., Qian, J., Gao, S., Chen, G., Li, Q., Huang, Y.-A., and Huang, Z.-A. Better eyes, better thoughts: Why vision chain-of-thought fails in medicine, 2026. URL https://arxiv.org/abs/2603.06665.

Xie, S., Kong, L., Dong, Y., Sima, C., Zhang, W., Chen, Q. A., Liu, Z., and Pan, L. Are VLMs ready for autonomous driving? an empirical study from the reliability, data and metric perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6585–6597, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Xie_Are_VLMs_Ready_for_Autonomous_Driving_An_Empirical_Study_from_ICCV_2025_paper.pdf.

Yu, L., Poirson, P., Yang, S., Berg, A. C., and Berg, T. L. Modeling context in referring expressions. In *European Conference on Computer Vision (ECCV)*, pp. 69–85. Springer, 2016. URL https://arxiv.org/abs/1608.00272.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.15343.

Zhang, Q., Singh, C., Liu, L., Liu, X., Yu, B., Gao, J., and Zhao, T. Tell your model where to attend: Post-hoc

attention steering for LLMs. In *The 12th International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2311.02262.

Zhu, J. et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.