Mechanisms of Object Localization in Vision–Language Models

Timothy Schaumlöffel[1,2] Martina G. Vilas[1] Gemma Roig[1,2]
[1]Goethe University Frankfurt, Germany [2]The Hessian Center for AI, Germany
https://github.com/t9s9/vlm-loc-mechanisms

Abstract

Visually-grounded language models (VLMs) are highly effective in linking visual and textual information, yet they often struggle with basic classification and localization tasks. While classification mechanisms have been studied more extensively, the processes that support object localization remain poorly understood. In this work, we investigate two representative families, LLaVA-1.5 and InternVL-3.5, using a suite of mechanistic interpretability tools, including token ablations, attention knockout, and causal mediation analysis. We find that localization is driven by a containerization mechanism in which object-aligned tokens define the spatial extent of the object, while the semantic arrangement of tokens within those boundaries is largely irrelevant to the predicted box. Only a very small set of attention heads mediates the causal effect for both classification and localization, concentrating in early–mid layers for LLaVA and mid–late layers for InternVL. The two tasks share some early processing but ultimately depend on largely distinct specialized heads. Overall, we provide the first layer- and head-level account of localization in VLMs, revealing narrow computational pathways that can guide future model design and grounding objectives.

1. Introduction

Visually-grounded Language Models (VLMs) combine a pre-trained vision encoder with a large language model (LLM), typically refined through vision-language instruction tuning. The visual encoder extracts grid-level features from an image, a multimodal adapter maps them into the language embedding space, and the resulting tokens are processed jointly with text by the LLM. This architecture allows VLMs to link visual and textual inputs and has enabled strong performance on tasks such as visual question answering, captioning, and open-ended reasoning about images [1, 8, 10, 19].

Despite these advances, VLMs continue to struggle with core vision tasks. They often misclassify or fail to accurately localize objects [31, 38]. While the mechanisms underlying classification have been studied [21, 38], much less is known about localization and detection. Closing this gap is important because most VLMs inherit visual features from CLIP [25], which was trained with global image-text supervision and struggles with the pixel-level precision required for localization and detection [3, 27, 39]. Yet VLMs can still answer queries that require identifying and locating objects, suggesting that these models build spatial structure from weakly grounded visual representations. This raises the question of how the mechanisms enabling localization and detection emerge in VLMs.

In this paper, we present an initial mechanistic study of object localization in VLMs. We combine token-level ablations, controlled perturbations of visual representations, positional decoding, attention knockout, and causal mediation analysis to probe how information relevant for localization is encoded and transformed inside the model. Our main findings are:

1. **Grounding through containerization.** Localization information is directly encoded in the visual tokens. The model groups these tokens into *containers* that define object boundaries, largely independent of the spatial arrangement of semantics within the object boundaries.
2. **Multi-view integration of spatial and semantic cues.** In architectures with global and local views, the global view carries the dominant spatial signal for localization, while local high-resolution crops primarily refine classification, especially for small objects. The two views provide complementary, rather than redundant, evidence.
3. **Implicit spatial layout learning.** The LLM infers the two-dimensional structure of the image from the one-dimensional token sequence: residual positional signals at the multimodal projection and strong corner anchors are sufficient for the model to reconstruct approximate row boundaries and a grid-like layout.
4. **Sparse, task-critical attention heads.** A very small number of attention heads mediate the causal effect for both classification and localization. In LLaVA models, these heads emerge predominantly in the *early–mid layers*, whereas in InternVL they appear in the *mid–late*

layers. Despite partial overlap in early processing, the dominant heads for the two tasks are largely disjoint, yet localization causally depends on classification-critical heads, revealing a sequential mechanism in which object identification precedes spatial grounding.

2. Method

We start by introducing the model architectures, the dataset, and task definitions used throughout this work.

2.1. Visually Grounded Language Models

We study vision–language models that follow the ViT $\rightarrow$ MLP $\rightarrow$ LLM paradigm, where a visual encoder extracts patch features, an MLP projects them into the language space, and a large language model (LLM) generates the output. We choose two representative VLMs that instantiate this paradigm at different levels of architectural complexity: LLaVA-1.5 [19], a simple and interpretable baseline, and InternVL-3.5 [33], a state-of-the-art variant incorporating token compression and multi-view processing.

LLaVA-1.5 employs a CLIP ViT-L/14 [25] visual backbone and Vicuna LLM [7] connected by a two-layer MLP adapter. Images are padded to square shape and resized to 336^2 px. The backbone outputs 24×24 patch embeddings, which are directly mapped into the LLM embedding space without spatial aggregation. This one-to-one token mapping makes LLaVA a simple, interpretable baseline for analyzing visual–linguistic alignment. We analyze the two available versions: LLaVA-7B and LLaVA-13B, which use Vicuna-7B/13B [7] as the language backbone, respectively.

InternVL-3.5 uses a custom, contrastively pre-trained InternViT-300M backbone [6] and a Qwen3 LLM [34], linked by a two-layer MLP adapter. It introduces two key architectural extensions that distinguish it from LLaVA: (i) *Pixel Shuffle*: Each 2×2 block of visual tokens from the backbone is merged into a single token before projection into the text space using a learned compression. This reduces the number of tokens by a factor of four while preserving local spatial structure. (ii) *Dynamic High-Resolution Processing*: Input images are split into a variable number of 448^2 px tiles that are processed independently by the visual backbone. The number of tiles is chosen dynamically based on the image's aspect ratio and size, enabling more detailed processing. In parallel, a globally resized 448^2 px thumbnail provides coarse context. We refer to the high-resolution tiles as local views and the thumbnail as the global view. All local and global tokens are concatenated and passed to the LLM. After compression, each crop produces 16×16 visual tokens. In our experiments, we cap the number of local tiles at six to reduce computational cost. We study InternVL-3.5 8B, which uses a Qwen3-8B language model.

2.2. Dataset

To ensure that our analyses isolate the visual evidence used by VLMs, we construct a carefully curated dataset derived from the COCO validation split [18]. This requires correcting annotation inaccuracies and filtering images to remove ambiguous or low-quality cases prior to evaluation.

Base Dataset and Filtering We use the COCO validation split and ensure that none of the evaluated models were trained on these samples. Because the split contains multiple annotation issues (e.g., missing objects and coarse masks), we first apply the semi-automatic correction procedure of Singh et al. [28]. We then apply a small set of quality filters to remove extremely small or dominant objects, low-resolution images, and ambiguous cases with multiple valid targets. A detailed description of the filtering steps is provided in Appendix 6.1. After filtering, the dataset contains 6.403 object annotations across 3.560 images.

Object-Removed Control Set Contextual cues can lead to *hallucinated detections*, where models predict the presence of an object solely from background context. To control for this effect, we construct an auxiliary object-removed variant of the dataset. For each image, the target object is removed and the missing region is inpainted using LaMa [30], which reconstructs background structure with high realism. We retain only those image pairs for which a model correctly identifies the object in the original image but fails to do so in the inpainted counterpart. This ensures that subsequent analyses rely on real object evidence rather than contextual correlations. Examples of the inpainted dataset are provided in Appendix Figure 6.

Because this procedure results in model-dependent subsets, we take the intersection across all three models, yielding 2.248 object annotations across 1.720 images as our final probing subset.

2.3. Task

We evaluate models on two complementary visual tasks: Classifying object presence in the image and localizing its position by providing the bounding box coordinates. For each task, we design a different prompt for the same image.

Localization. The model is prompted to predict the bounding box coordinates of a target object. The predicted bounding boxes are parsed and compared against ground-truth annotations using the intersection-over-union (IoU) metric. Performance is measured as the success rate, defined as the proportion of samples where IoU exceeds thresholds of 0.5, 0.7, and 0.9. The final localization score is obtained by averaging over these three thresholds.

Classification. The model is prompted to list all objects present in the image, restricted to COCO's category set.

A prediction is counted as correct if the ground-truth class name appears anywhere in the model's response. Performance is reported as the proportion of correctly classified instances. We adopt a list-based formulation over a binary alternative to reduce object hallucinations.

For more details, we refer to Appendix Section 6.2.

3. Experiments

3.1. Visual Information Ablation

We conduct an ablation study to investigate the contribution of visual input tokens to the performance of the VLMs on the classification and localization tasks.

Method. We ablate visual information at the LLM input, i.e., after the multimodal projection but before positional encodings and autoregressive processing. To remove image-specific content while preserving domain-consistent embedding statistics, we replace the original visual token embeddings with a global average visual embedding computed once over the ImageNet [11] validation set.
We evaluate four token selection strategies for ablation:

i) *Object Tokens*: We project the object mask onto the image token grid and include all tokens that overlap with it by at least one pixel. To probe for boundary sensitivity and context dependence, we shrink or dilate the mask by 1 or 2 token padding. For InternVL models, this procedure is applied to both the local high-resolution and the global thumbnail views of the object. A visualization and details of the masking procedure are provided in the appendix Figure 7.

ii) *Register Tokens*: Global image features are hypothesized to be encoded in register tokens [9]. We therefore select those tokens whose embedding norms exceed two standard deviations above the mean.

iii) *Integrated Gradients*: We identify the image tokens most relevant to the model's decision by computing Integrated Gradients [29] with respect to the correct class logits (for classification) or bounding box coordinates (for localization). Tokens are ranked by their attribution magnitude, and the top-k highest-gradient tokens are selected as the most influential ones.

iv) *Random Tokens*: As a control, we randomly select k image tokens, repeat the process with three different seeds, and report the mean and standard deviation.

Results. As Table 1 shows, across all three models, both localization and classification tasks rely on the information encoded in object tokens. Ablating these tokens results in a significantly larger performance decline compared to removing an equal number of tokens either randomly or via gradient-based selection.Localization is more affected than classification: removing object tokens reduces localization performance below 10% accuracy, while classification still succeeds in 20–30% of cases. Positive padding around the object further amplifies the effect, while maintaining the original boundaries through negative padding has minimal influence on performance. These findings indicate that the essential information for both tasks resides within the object boundaries.

3.1.1. Object Containerization

Next, we investigate the mechanisms by which the model encapsulates objects to generate bounding boxes. To test this, we artificially expand the ground-truth object mask by adding p layers of surrounding tokens. Concretely, we randomly duplicate tokens from within the original object and copy them into the adjacent padding region. This procedure increases the spatial extent of the object while disrupting its structure: the added area is filled with misplaced but object-related features (e.g., eye-related tokens may appear below a mouth in a face). We then measure whether the predicted bounding box expands accordingly across ten random sampling seeds and report the outcome in Figure 1.

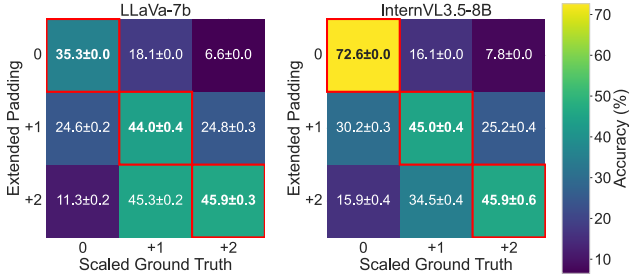

Figure 1. Alignment between predicted and scaled ground-truth bounding boxes under object padding. Each cell shows the mean accuracy between predictions obtained with a given padding level and ground-truth boxes scaled by different amounts. Diagonal entries correspond to matching padding and scaling levels, indicating how well the predicted box size adapts to the artificially enlarged object. Standard deviations are annotated. LLaVA-13B results are shown in the appendix Figure 8.

Across all architectures, the predicted bounding boxes scale consistently with the artificially enlarged objects, as reflected by the strong diagonal alignment: predictions from padding = 1 inputs achieve the highest accuracy with the +1-scaled ground-truth boxes, while padding = 2 inputs best align with the +2-scaled boxes. The results indicate that localization depends mainly on the presence of object-related tokens within the spatial region rather than on their semantically coherent arrangement. We show qualitative examples in appendix Figure 9.

To further support this claim, we shuffle the image tokens

Table 1. Performance after token ablation. The baseline corresponds to the model without any token removal and serves as a reference for ablations targeting the object mask (with varying padding), highest-gradient tokens, random tokens, and register tokens. We report both absolute accuracy and the corresponding drop relative to the baseline. The average proportion of removed tokens is indicated as a percentage of all image tokens; for InternVL, if applicable, we report the number of removed tokens for the (global, local) views separately. Similar amounts of removed tokens are highlighted in **bold**. Standard deviations across random seeds are provided in appendix Table 4.

Models:		**LLaVA 7B**		**LLaVA 13B**		**InternVL3.5 8B**		
Ablation Strategy	Token (%)	Loc. (%)	Cls. (%)	Loc. (%)	Cls. (%)	Token (%)	Loc. (%)	Cls. (%)
Baseline	*0*	35.34	58.10	46.98	65.30	*0*	72.64	83.30
– 2 Padding	*1*	34.71 ↓0.6	57.78 ↓0.3	47.09 ↑0.1	65.21 ↓0.1	*(1,2)*	72.73 ↑0.1	83.41 ↑0.1
– 1 Padding	*3*	31.73 ↓3.6	55.29 ↓2.8	43.74 ↓3.2	65.48 ↑0.2	*(4,5)*	72.35 ↓0.3	81.85 ↓1.5
Object	***8***	**5.92 ↓29.4**	**19.44 ↓38.7**	**7.37 ↓39.6**	**31.41 ↓33.9**	***(13,10)***	**11.27 ↓61.4**	**33.19 ↓50.1**
+ 1 Padding	*14*	0.73 ↓34.6	11.25 ↓46.9	0.59 ↓46.4	15.12 ↓50.2	*(23,14)*	3.77 ↓68.9	23.71 ↓59.6
+ 2 Padding	*21*	0.34 ↓35.0	10.59 ↓47.5	0.28 ↓46.7	12.54 ↓52.8	*(34,18)*	2.02 ↓70.6	20.73 ↓62.6
Integrated Gradients	*1*	31.54 ↓3.8	52.67 ↓5.4	38.80 ↓8.2	60.36 ↓4.9	*1*	62.46 ↓10.2	81.49 ↓1.8
	4	22.14 ↓13.2	48.40 ↓9.7	23.50 ↓23.5	54.72 ↓10.6	*4*	43.64 ↓29.0	77.49 ↓5.8
	8	**15.84 ↓19.5**	**43.95 ↓14.2**	**13.09 ↓33.9**	**49.24 ↓16.1**	*8*	30.13 ↓42.5	74.69 ↓8.6
	16	8.75 ↓26.6	37.28 ↓20.8	5.66 ↓41.3	44.80 ↓20.5	*16*	**16.21 ↓56.4**	**71.00 ↓12.3**
	32	2.79 ↓32.6	29.05 ↓29.1	1.39 ↓45.6	35.72 ↓29.6	*32*	6.44 ↓66.2	64.86 ↓18.4
	48	0.86 ↓34.5	32.25 ↓25.9	0.37 ↓46.6	28.65 ↓36.6	*48*	3.23 ↓69.4	59.52 ↓23.8
Random (3 seeds)	*1*	35.52 ↑0.2	57.98 ↓0.1	46.74 ↓0.2	64.86 ↓0.4	*1*	72.32 ↓0.3	83.44 ↑0.1
	4	35.57 ↑0.2	57.35 ↓0.8	45.99 ↓1.0	64.68 ↓0.6	*4*	72.30 ↓0.3	83.33 ↑0.0
	8	**35.09 ↓0.2**	**56.58 ↓1.5**	**45.25 ↓1.7**	**63.89 ↓1.4**	*8*	71.71 ↓0.9	83.10 ↓0.2
	16	33.71 ↓1.6	56.39 ↓1.7	43.76 ↓3.2	63.92 ↓1.4	*16*	**70.46 ↓2.2**	**83.02 ↓0.3**
	32	30.43 ↓4.9	55.40 ↓2.7	39.88 ↓7.1	62.54 ↓2.8	*32*	66.88 ↓5.8	82.62 ↓0.7
	48	25.65 ↓9.7	54.80 ↓3.3	34.44 ↓12.5	61.34 ↓4.0	*48*	59.03 ↓13.6	81.69 ↓1.6
Register	*1*	35.07 ↓0.3	59.48 ↑1.4	46.53 ↓0.5	64.95 ↓0.4	*(4,4)*	72.64 ↑0.0	83.41 ↑0.1

within the object mask directly at the LLM input. As shown in Table 2, localization performance drops only slightly under this perturbation compared to a full shuffle of all image tokens. Notably, object classification is unaffected by either shuffling procedure, confirming that the observed effect is specific to localization.

Together, these findings suggest that the model employs a form of **containerization**, in which tokens collectively define the spatial extent of an object, largely independent of their internal semantic structure.

3.1.2. Contribution of Global and Local Views

To understand how the InternVL model distributes semantic and spatial information across its two visual pathways, we ablate the object-aligned tokens in either the global resized view or the fine-grained local view. Table 3 summarizes the resulting change in accuracy for localization and classification. Removing object tokens from only one view causes a moderate drop, whereas ablating both views simultaneously leads to a substantially larger decline (see Tab. 1) for both tasks. This indicates that the model integrates complementary information from the global and local pathways.

Table 2. Performance under input token shuffling perturbations. We report localization and classification accuracy for two conditions: (i) shuffling all image tokens, and (ii) shuffling image tokens within the object mask. Results are averaged across three seeds.

Mode	**LLaVA 7B**	**LLaVA 13B**	**InternVL3.5 8B**
	Localization		
Baseline	35.34	46.98	72.64
Full	1.90 ± 0.10 ↓33.4	0.30 ± 0.00 ↓46.7	4.70 ± 0.20 ↓67.9
Object	35.30 ± 0.20 ↓0.0	45.20 ± 0.20 ↓1.8	37.70 ± 0.40 ↓34.9
	Classification		
Baseline	58.10	65.30	83.30
Full	62.42 ± 0.62 ↑4.3	67.04 ± 0.33 ↑1.7	81.19 ± 0.39 ↓2.1
Object	58.44 ± 0.39 ↑0.3	65.21 ± 0.25 ↓0.1	83.15 ± 0.40 ↓0.1

The effect is particularly pronounced for localization: removing the global object tokens reduces accuracy by -36.4%, while local ablation yields a smaller decline of -9.7%. Classification shows the same trend but with reduced magnitude (-9.5% vs. -6.6%). These results identify the global pathway as the primary carrier of spatial grounding, with local high-resolution cues providing additional semantic detail. Padding amplifies this effect. In

single-view ablations, increasing the masked region leads to only minor additional degradation because the intact view still provides sufficient object evidence. In two-view ablations, however, padding removes the remaining semantic and spatial cues in both pathways, resulting in much larger drops. This dissociation indicates that each view can compensate for moderate damage to the other, but neither can compensate when both are impaired, demonstrating strong synergy rather than redundancy between the global and local representations.

Table 3. Extended ablation experiment for separate (local and global) views of the **InternVL** architecture.

	Strategy	Token (%)	Loc. (%)	Cls. (%)
	Baseline	*0*	72.64	83.30
Local	– 1 Padding	*5*	72.41 ↓0.2	82.52 ↓0.8
	Object	***11***	**62.93** ↓**9.7**	**76.65** ↓**6.6**
	+ 1 Padding	*15*	63.20 ↓9.4	77.45 ↓5.9
Global	– 1 Padding	*5*	72.58 ↓0.1	83.14 ↓0.2
	Object	***14***	**36.20** ↓**36.4**	**73.80** ↓**9.5**
	+ 1 Padding	*24*	25.86 ↓46.8	72.06 ↓11.2

We complement these findings as shown in appendix Figure 10 by breaking down the performance drop by object size. For single-view ablations, the degradation decreases systematically with object size. Small objects are highly sensitive to the removal of either view, especially for localization (global: -64.8%; local: -54.3%). Medium objects show reduced but still substantial dependency. Large objects are comparatively robust: global removal still harms localization (-26.1 %), whereas local removal can even slightly improve performance (e.g., +6.5 %), suggesting redundancy or noise in fine-detail tokens for large objects.

In summary, we conclude that the global view supplies the essential spatial signal for localization, while the local view primarily supports the classification of small objects. The two pathways therefore provide complementary evidence whose integration enables robust semantic and spatial reasoning. These insights suggest that the number of crops could be adapted dynamically to the task, potentially allowing for more efficient model configurations.

3.2. Position Encoding

Before examining how object information propagates through the network, we first assess how much of the spatial structure needed for localization is preserved across the model's layers.

Method. To evaluate how positional information remains identifiable throughout the model's processing hierarchy, we train a separate linear classifier for each model layer, including the multimodal projection, to predict the position of every image token in the input grid. We predict each spatial axis independently. Classifiers are trained for 10 epochs on 50,000 ImageNet images and evaluated on 10,000 images.

Results. In Figure 2, we observe that positional information in both visual backbones is decodable from early layers but largely vanishes by the final layers, consistent with prior findings that contrastively trained ViTs trade spatial precision for semantic abstraction over depth [14]. In contrast, in the LLM, positional identifiability is initially low. However, it increases rapidly and peaks around layer 12 for LLaVA-7B, layer 13 for LLaVA-13B and layer 7 for InternVL3.5-8B. The multimodal projection retains strong signals for the four image corners (see app. Fig. 11), which appear sufficient for the LLM to infer approximate row boundaries ("line breaks") across the token sequence. Tokens aligned with these inferred line breaks are predicted with higher probability than other positions (Fig. 2, right), suggesting that the model uses them as structural anchors when reconstructing spatial layouts.

3.3. Localizing Task Processing

Next, we investigate the location of task-specific processing in the language model. Specifically, we examine whether classification and localization rely on shared or distinct components by progressively narrowing the analysis from layer groups to individual attention heads.

3.3.1. Attention Knockout

Method. Our next experiments aim to identify where in the network the visual information required for the task is extracted and processed. We apply the *attention knockout* technique [13, 21], which blocks attention and thereby prevents communication between tokens. Unlike [21], we eliminate attentions from all tokens following the image tokens to the object tokens, effectively removing any information extracted from the object tokens in those layers. We combine layers in groups of four and block all attention heads within each group, as well as across all layers, as a global baseline. We evaluate the resulting performance drop on classification and localization tasks using our filtered COCO subset.

Results. Our results are shown in Figure 3. For the LLaVA models, blocking attention to the object significantly decreases performance in the early to mid layers of the model, while perturbing attention flow in later layers leaves performance largely unaffected. In contrast, for InternVL, the strongest decline occurs in the middle to later layers.
Across all models, both classification and localization rely on early shared layers, after which localization depends on

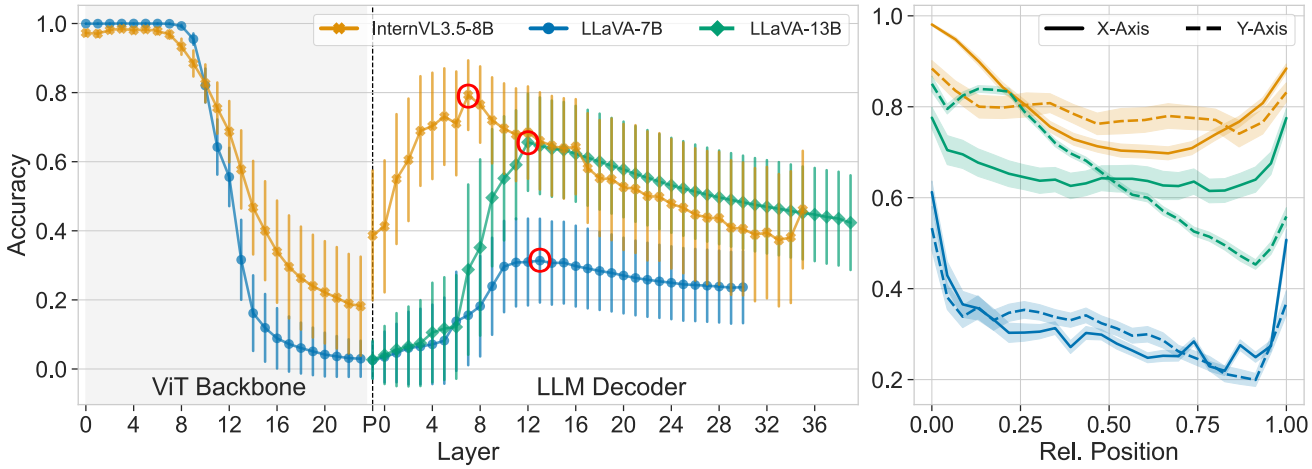

Figure 2. Positional decoding results. Left: average position accuracy per layer for visual backbone (0-23), the multimodal projection (P) and three different language models (0-39). The maximum value per language model is marked with a red circle. Right: per-position accuracy at maximum layer of LLM, showing higher accuracy at the image corners.

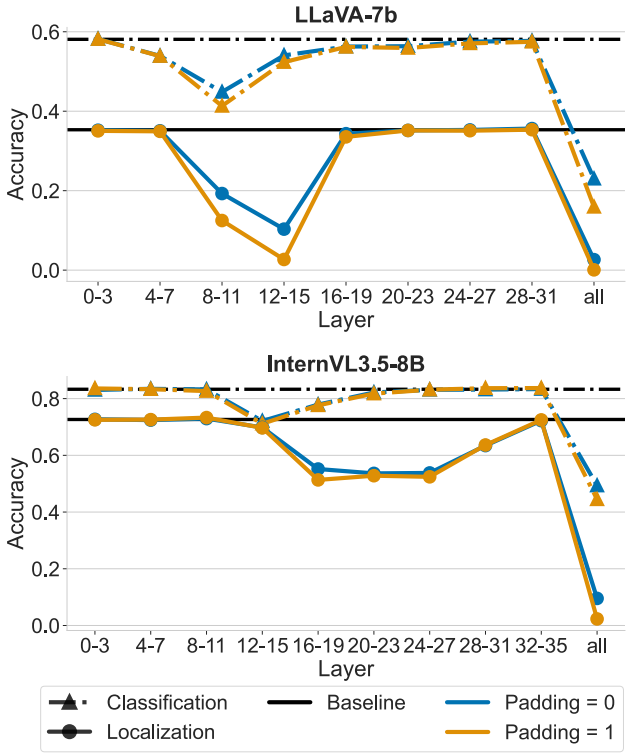

Figure 3. Performance after attention knockout. We block attention from image tokens to object tokens across grouped layers. The curves show classification and localization accuracy and compares to accuracy of the unmodified model. Additional results can be found in appendix Section 9.

additional task-specific processing. This pattern suggests a two-step processing mechanism: the model first identifies the object, then localizes it. Moreover, for the LLaVA models, the layers with the largest localization decline align with those that retain strong positional information (see Sec. 3.2). Compared to layer-wise knockouts (app. Fig. 12 right), ablating groups of layers amplifies the effect, particularly for classification, as further shown in Figure 12 (left), where we group six consecutive layers. This suggests that the model accumulates object-related information across several layers rather than relying on layer-specific mechanisms. The following section builds on these findings by identifying task-specific heads through causal mediation analysis.

3.3.2. Causal Mediation Analysis

In the previous section, we presented evidence that the mechanisms underlying localization and classification are concentrated within a narrow region of the language model's processing pipeline. Using causal mediation analysis (CMA) [20, 32, 35], we now aim to pinpoint in which model components these mechanisms reside. CMA enables us to estimate the causal contribution of an embedding at a specific computational block within the model.

Method. We apply CMA using activation patching to identify which attention heads causally contribute to solving the visual task. For each example, we construct two versions of the *same* image: a source image, where the relevant object is present, and a base image, where the object has been removed using a diffusion-based inpainting model (see Sec. 2.2). We use 50 images from this curated subset. Separately for each attention head, we extract the hidden activations from the source run and patch them into the forward pass of the base run, yielding the counterfactual output $y^{\star}$. As in our attention-blocking experiments (Sec. 3.3.1), we patch all activations associated with the prompt tokens. An overview of the setup is shown in Figure 13.
All outputs are evaluated under teacher-forcing using token-

level perplexity. Since fixed template tokens (e.g., brackets in a bounding-box answer) are predicted with near-deterministic probability and would artificially dominate perplexity, we mask these tokens out when computing the score. Let P_{base}, P_{src}, and P_{patched} denote the perplexity of the base, source, and patched runs, respectively. We quantify the causal contribution of a component using the Mediation Fraction (MF):

$$\text{MF} = \frac{P_{\text{base}} - P_{\text{patched}}}{P_{\text{base}} - P_{\text{src}}}. \quad (1)$$

MF measures how much of the performance gap between base and source is closed by the patched component: an MF close to 1 indicates that the patched head fully mediates the task-relevant information; MF ≈ 0 implies no contribution; and MF < 0 indicates misleading or interfering effects.
For classification, we use the binary query (see Sec. 6.2). The list formulation generates object names in an arbitrary order, which spreads probability mass across many tokens. This leads to per-class perplexity differences that are too noisy for reliable mediation estimates. In contrast, the binary query produces a single decisive token, which makes it well suited for activation patching. Although it shows a higher hallucination rate in open evaluation, this is not a problem in our setting. The paired source–base design ensures that changes in perplexity reflect causal effects.

Results. Figure 4 reports the Mediation Fraction (MF) for every attention head across all layers. Consistent with the attention-blocking experiment (Sec. 3.3.1), the causal mediation analysis locates the core processing region in the early–mid layers for LLaVA and in the mid–late layers for InternVL. For LLaVA, most non-zero MF values cluster around layers 11–16 for both localization and classification, whereas for InternVL the dominant heads appear in layers 16–22. Outside these ranges, nearly all heads exhibit MF scores close to zero, indicating that the vast majority of attention heads are not causally involved in the task.

In addition, the distribution of MF scores across heads is highly sparse for both tasks. Only a small number of attention heads exhibit large mediation values, indicating that the bulk of task-relevant information is carried by a compact set of specialized heads. The vast majority of heads contribute negligibly, with MF values near zero, suggesting that they neither facilitate nor interfere with the task. The two tasks differ in how this information is distributed: localization relies on a more concentrated subset of heads, whereas classification engages a broader set of heads dispersed across earlier and intermediate layers. Despite these differences, we do observe a limited number of shared heads. Among the top-10 heads ranked by MF, only two heads in LLaVA and one in InternVL are shared across both tasks; even when expanding to the top-50 heads, the overlap remains limited: 20 for LLaVA-7B, 15 for LLaVA-13B, and 18 for InternVL. This pattern indicates that while some early computations are reused, the dominant mediators of localization and classification are largely distinct.
Finally, we observe a substantial number of negative MF values, especially in InternVL, suggesting that certain activations from the source image are incompatible with those of the base image. In these cases, patching increases perplexity, indicating that these heads encode counter-evidence or contextual signals that conflict with the ground-truth output.

Overall, these results reveal that VLMs implement visual reasoning through highly selective, low-redundancy pathways: only a handful of attention heads mediate nearly the entire causal effect, and different tasks rely on different subsets of these components, reflecting functional specialization within the model's internal computation.

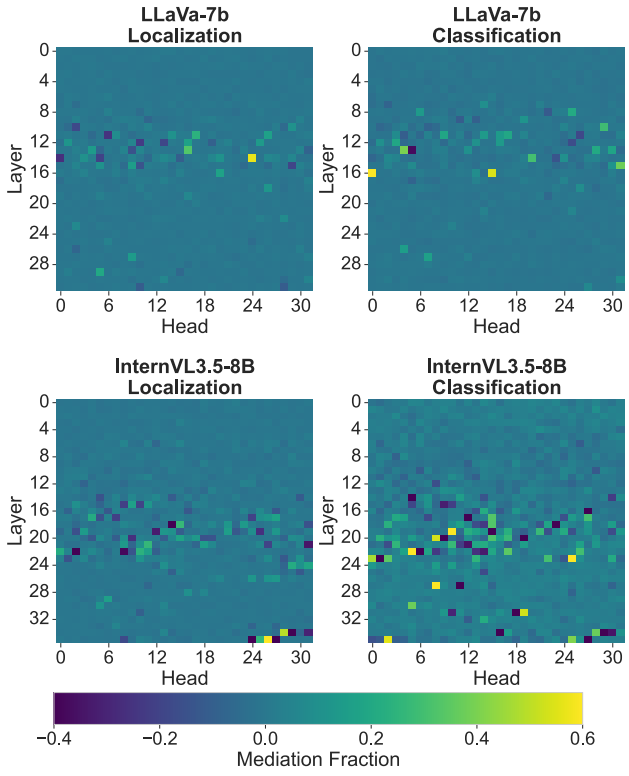

Figure 4. Mediation Fraction (MF) scores for every attention head across all layers, shown separately for the localization and classification tasks. For LLaVA-13B results, refer to appendix Figure 14.

3.3.3. Head Ablation Analysis

CMA reveals a small number of attention heads with high MF scores per task. To determine whether these heads are necessary rather than merely correlated with performance,

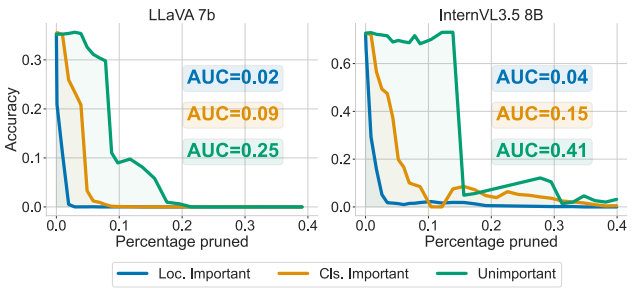

Figure 5. Localization accuracy under cumulative head ablation. Attention heads are ranked by their mean MF and progressively removed. Normalized AUC scores enable method comparison.

we progressively removed them in a cumulative ablation study.

Method. For each task, we rank all attention heads according to their mean MF score and divide them into two groups: task-critical heads (those with the highest MF scores) and low-importance heads (near zero MF heads). We then perform cumulative ablations on both groups by successively removing a proportion of the total number of heads and measuring the resulting localization accuracy. Ablation is implemented by setting the output of the selected heads to zero during the forward pass. Three settings are evaluated: ablation of localization-critical heads, ablation of classification-critical heads and ablation of low-importance heads as a control baseline.

Results. As shown in Figure 5, ablating task-critical heads leads to a substantially larger drop in localization accuracy than removing an equal number of low-importance heads. This confirms that the CMA ranking identifies components that are causally necessary for the task rather than simply being correlated with performance. In contrast, removing unimportant heads results in only a slight decline in performance, even when a large fraction of them is removed, indicating that task-relevant information is concentrated in a small subset of attention heads. Despite the limited overlap between classification-critical and localization-critical heads, ablating classification-critical heads still causes a strong degradation in localization accuracy. This indicates that localization depends on intermediate object-identification representations rather than operating independently. Together with the attention-knockout results (Sec. 9), these findings provide causal evidence for a sequential processing mechanism in which the model first identifies the object and subsequently determines its spatial extent using a smaller set of specialized heads.

4. Related Work

Prior mechanistic interpretability studies of VLMs primarily examine high-level reasoning, VQA behavior, or hallucinations [2, 15, 16, 22]. Li et al. [17] traces object representations across layers and modalities to mitigate hallucinations, while Yu and Lee [36] uses layer-wise probing to reveal a stage-wise processing hierarchy in VLMs. We build on these approaches but focus specifically on where and how localization information is extracted and transformed, using causal mediation analysis at the attention-head level.

A second thread of research documents systematic failure modes across a range of VLM capabilities, including single-object classification [38], multi-object classification [4], counting [26], visual search [4], and spatial reasoning [5]. These studies reveal significant gaps in how VLMs ground semantic and spatial information, but they do not examine object localization or identify where in the model spatial representations emerge.

In parallel, several recent works aim to improve grounding and localization capabilities in VLMs through architectural changes, additional training strategies, or specialized datasets [1, 10, 23, 24, 37]. While these models demonstrate enhanced performance on grounding benchmarks, their internal computational mechanisms remain largely unexplored.

5. Discussion and Limitations

Our findings shed light on the fundamental mechanisms through which VLMs capture and encode spatial structure. Our experiments reveal that positional information is reconstructed in the LLM, rather than relying on positional information encoded in the visual backbone. Localization depends on a containerization process in which object tokens collectively define spatial boundaries, while the internal spatial or semantic arrangement of these tokens plays only a minor role. Moreover, causal mediation analysis shows that only a very small number of attention heads are responsible for both tasks, emerging in the early–mid layers for LLaVA and in the mid–late layers for InternVL, with classification and localization relying on largely distinct sets of specialized heads. Cumulative head ablation further reveals that localization causally depends on classification-critical heads, providing evidence consistent with a sequential processing mechanism in which the model first identifies the object and subsequently determines its spatial extent. These results refine our understanding of VLMs and open directions, including targeted head fine-tuning or grounding-aware attention supervision.

Our study uses a filtered COCO subset where images contain multiple objects but only one per queried category. This allows us to isolate foundational spatial mechanisms, and the same approach can be naturally extended to more complex images and tasks. We apply CMA to attention heads and analyze fixed models, leaving other components and training dynamics unexplored. Extending this framework to segmentation, relational grounding, video, and additional architectures is a promising direction for future work.

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft: DFG project 5368 ("Abstract REpresentations in Neural Architectures (ARENA)") and DFG project 539642788, RO 6458/5-1 ("Learning from the Environment Through the Eyes of Children (LEECHI)"). We gratefully acknowledge support from The Hessian Center For Artificial Intelligence and Goethe-University (NHR Center NHR@SW) for providing the computing and data-processing resources needed for this work.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. 1, 8

[2] Samyadeep Basu, Martin Grayson, Cecily Morrison, et al. Understanding information storage and transfer in multi-modal large language models. *Advances in Neural Information Processing Systems*, 37:7400–7426, 2024. 8

[3] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 1

[4] Declan Iain Campbell, Sunayana Rane, and Tyler Giallanza. Understanding the limits of vision language models through the lens of the binding problem. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 8

[5] Shiqi Chen, Tongyao Zhu, and Ruochen Zhou. Why is spatial reasoning hard for VLMs? an attention mechanism perspective on focus areas. In *Forty-second International Conference on Machine Learning*, 2025. 8

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, et al. 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*, pages 24185–24198, 2024. 2

[7] Wei-Lin Chiang, Zhuohan Li, Zi Lin, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023. 2

[8] Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023. 1

[9] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *The Twelfth International Conference on Learning Representations*, 2024. 3

[10] Matt Deitke, Christopher Clark, Sangho Lee, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 1, 8

[11] Jia Deng, Wei Dong, Richard Socher, et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[12] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 88(2): 303–338, 2010. 3, 4

[13] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, 2023. 5

[14] Dongsheng Jiang, Yuchen Liu, Songlin Liu, et al. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023. 5

[15] Nicholas Jiang, Anish Kachinthaya, Suzanne Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. 8

[16] Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14549–14558, 2025. 8

[17] Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 31645–31653, 2026. 8

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, et al. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 1, 5

[19] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1, 2

[20] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2022. Curran Associates Inc. 6

[21] Clement Neo, Luke Ong, Philip Torr, et al. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 1, 5

[22] Vedant Palit, Rohan Pandey, Aryaman Arora, and Paul Pu Liang. Towards vision-language mechanistic interpretability: A causal tracing tool for blip. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2856–2861, 2023. 8

[23] Georgios Pantazopoulos and Eda B Özyiğit. Towards understanding visual grounding in visual language models. *arXiv preprint arXiv:2509.10345*, 2025. 8

[24] Zhiliang Peng, Wenhui Wang, Li Dong, et al. Grounding multimodal large language models to the world. In *The*

Twelfth International Conference on Learning Representations, 2024. 8

[25] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2

[26] Sunayana Rane, Alexander Ku, and Jason Michael Baldridge. Can generative multimodal models count to ten? In *ICLR 2024 Workshop on Representational Alignment*, 2024. 8

[27] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 139–156. Springer, 2024. 1

[28] Shweta Singh, Aayan Yadav, Jitesh Jain, et al. Benchmarking object detectors with coco: A new path forward. 2024. 2, 1

[29] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[30] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, et al. Resolution-robust large mask inpainting with fourier convolutions. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3172–3182, 2022. 2

[31] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, et al. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578, 2024. 1

[32] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. 6

[33] Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2

[34] An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report, 2025. 2

[35] Yukang Yang, Declan Campbell, Kaixuan Huang, Mengdi Wang, Jonathan Cohen, and Taylor Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. *Proceedings of Machine Learning Research*, 267:70515–70549, 2025. 6

[36] Zhuoran Yu and Yong Jae Lee. How multimodal LLMs solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Second Conference on Language Modeling*, 2025. 8

[37] Hao Zhang, Hongyang Li, Feng Li, et al. Llava-grounding: Grounded visual chat with large multimodal models. In *European Conference on Computer Vision*, pages 19–35. Springer, 2024. 8

[38] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, et al. Why are visually-grounded language models bad at image classification? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 8

[39] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 1