Mechanisms of Object Localization in Vision–Language Models

Supplementary Material

6. Dataset

We provide additional details on dataset construction, task prompting, and representative examples.

6.1. Dataset Filtering Details

We evaluate on the COCO validation split [18] with label corrections from [28], and apply the following filtering steps to improve annotation quality and satisfy the requirements of our experimental setup.

1. **Object size.** We discard objects occupying less than 0.4% or more than 60% of the image area. Extremely small objects provide insufficient visual detail, while very large objects dominate the field of view.
2. **Resolution.** Images with a minimum side length below 200 px are removed to ensure sufficient spatial resolution.
3. **Uniqueness filtering.** We retain only images that contain a single instance of the queried object category in order to avoid multiple valid answers for the same query.

6.2. Prompts and Input Template

For each image–object pair, we query the model either to localize a target object or to classify the objects present in the image.

Input template. All models receive inputs in the following format:

$$\{\texttt{System}\}\ \{\texttt{Image}\}\ \{\texttt{Task}\}$$

where `System` is the model-specific system prompt prepended to every query, `Image` refers to the embedded visual tokens, and `Task` is the task-specific instruction described below. Given this combined input, the model generates its response autoregressively.

Localization. For object localization, the model is prompted to predict the bounding box of a given target class:

`Please provide the bounding box coordinates of the {class}.`

where `{class}` is the name of the target object.

Classification. For object classification, we instruct the model to enumerate all objects present in the image while restricting the answer space to the predefined category set:

`List all objects in the image. Choose only from {class}$_1$, {class}$_2$, ...`

This formulation avoids the need to handle a large variety of free-form answers, which is particularly problematic for COCO due to its broad and diverse category set.

Alternative binary formulation for classification. A natural alternative is to pose a binary query per category:

`Is there a {class} in the image?`

However, this formulation substantially increases the incidence of object hallucinations. To quantify this effect, we measure the false positive rate

$$\mathrm{FPR} = P(\mathrm{detected} = \mathrm{True} \mid \mathrm{has\_object} = \mathrm{False})$$

on the manipulated dataset described in Section 2.2, where the queried object is absent. Across InternVL3-5 8B and LLaVA-1.5 7B/13B, the binary query yields consistently higher FPRs (0.56–0.70) than the list-based formulation (0.27–0.45). For this reason, all main experiments use the list-based prompt.

6.3. Image Examples

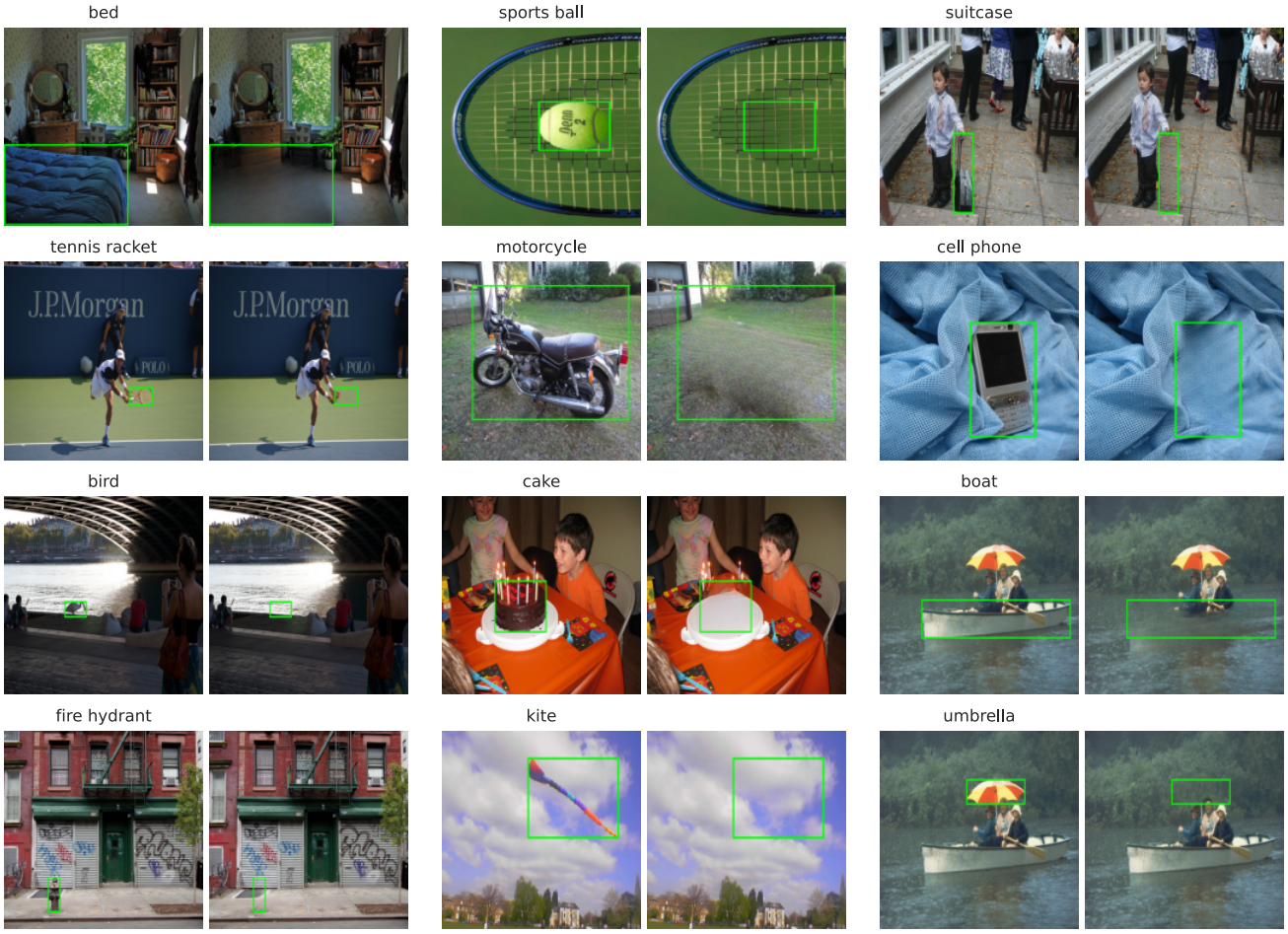

Figure 6. **Dataset examples.** Example images from the COCO validation set where exactly one object per image is removed (highlighted by its original bounding box) and the background is filled using an inpainting strategy [30]. This procedure allows us to filter the dataset for potential hallucinations: if the model can still detect the removed object purely from contextual cues, it undermines the validity of our grounding analysis.

7. Ablation Study

7.1. Visualization of Masking

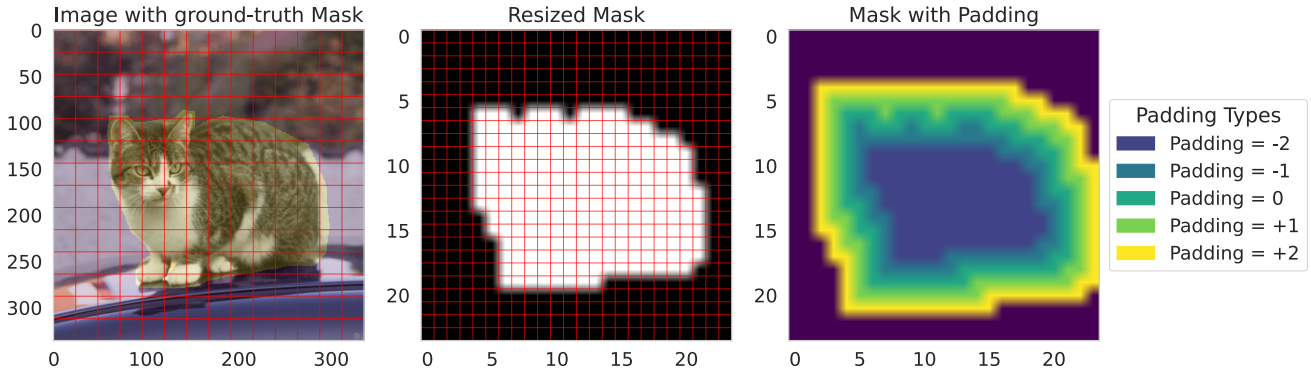

Figure 7. **Visualization of object mask for the ablation experiment**. Left: the original 336×336 input image in pixel space with an annotated object mask. This mask is mapped onto the 24×24 token grid of the vision transformer, where a token is selected if it has any pixel overlap with the original mask. Right: examples of padding applied to the token mask. Negative padding removes adjacent tokens and shrinks the ablated region, while positive padding adds neighboring tokens and expands it.

7.2. Standard Deviations of Random Ablation Experiment

Models:		LLaVa 7B		LLaVa 13B		InterVL3.5 8B	
Ablation Strategy	Token (%)	Loc. (%)	Cls. (%)	Loc. (%)	Cls. (%)	Loc. (%)	Cls. (%)
Baseline	*0*	35.34	58.10	46.98	65.30	72.64	83.30
Random (3 seeds)	*1*	35.52 ± 0.21 ↑0.2	57.98 ± 0.18 ↓0.1	46.74 ± 0.08 ↓0.2	64.86 ± 0.40 ↓0.4	72.32 ± 0.23 ↓0.3	83.44 ± 0.51 ↑0.1
	4	35.57 ± 0.09 ↑0.2	57.35 ± 0.18 ↓0.8	45.99 ± 0.39 ↓1.0	64.68 ± 0.23 ↓0.6	72.30 ± 0.39 ↓0.3	83.33 ± 0.09 ↑0.0
	8	35.09 ± 0.08 ↓0.2	56.58 ± 0.29 ↓1.5	45.25 ± 0.08 ↓1.7	63.89 ± 0.14 ↓1.4	71.71 ± 0.21 ↓0.9	83.10 ± 0.35 ↓0.2
	16	33.71 ± 0.50 ↓1.6	56.39 ± 0.65 ↓1.7	43.76 ± 0.16 ↓3.2	63.92 ± 0.24 ↓1.4	70.46 ± 0.20 ↓2.2	83.02 ± 0.39 ↓0.3
	24	31.92 ± 0.23 ↓3.4	55.72 ± 0.52 ↓2.4	41.95 ± 0.29 ↓5.0	63.51 ± 0.29 ↓1.8	68.81 ± 0.42 ↓3.8	82.83 ± 0.47 ↓0.5
	32	30.43 ± 0.13 ↓4.9	55.40 ± 0.34 ↓2.7	39.88 ± 0.33 ↓7.1	62.54 ± 0.52 ↓2.8	66.88 ± 0.47 ↓5.8	82.62 ± 0.56 ↓0.7
	48	25.65 ± 0.29 ↓9.7	54.80 ± 0.83 ↓3.3	34.44 ± 0.16 ↓12.5	61.34 ± 0.66 ↓4.0	59.03 ± 0.39 ↓13.6	81.69 ± 0.33 ↓1.6

Table 4. **Performance after token ablation.** The baseline corresponds to the model without any token ablation and serves as a reference for randomly ablated tokens. This table additionally reports one standard deviations across three random seeds for the random ablation experiments. See Table 1 for the complete results.

7.3. Additional Dataset

We provide additional results on the Pascal VOC dataset [12]. We applied the same pre-processing pipeline as desribed in Section 2.2 , which results in 1.957 objects across 1.585 images. We observe consisent results as describes in Section 3.1.

Models:	LLaVA 7B		LLaVA 13B		InternVL 8B	
Task:	Loc.	Cls.	Loc.	Cls.	Loc.	Cls.
Baseline	52.41	73.41	63.68	82.97	85.78	97.06
Object	**9.52**	**31.37**	**18.63**	**54.78**	**26.35**	**52.94**
Gradients	22.63	57.23	20.83	74.39	27.00	89.34
Random	50.41	72.96	61.08	84.52	83.76	97.30

Table 5. Token ablation results on the PascalVOC dataset [12]. We compare object-token ablation with ablation of an equal number of high-gradient and randomly selected tokens (3 seeds average).

7.4. Object Extension Experiment

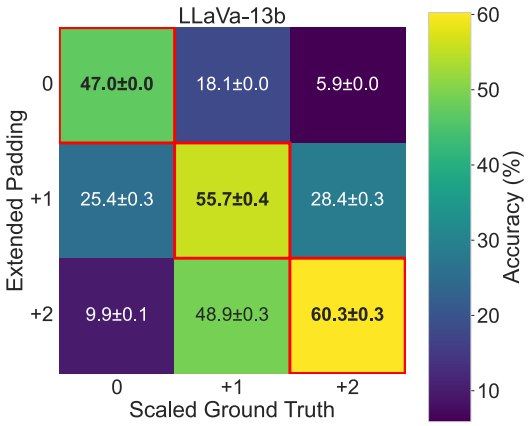

Figure 8. **Object extension additional Results.** Alignment between predicted and scaled ground-truth bounding boxes under object padding. Each cell shows the mean accuracy between predictions obtained with a given padding level and ground-truth boxes scaled by different amounts. Diagonal entries correspond to matching padding and scaling levels, indicating how well the predicted box size adapts to the artificially enlarged object. Standard deviations are annotated.

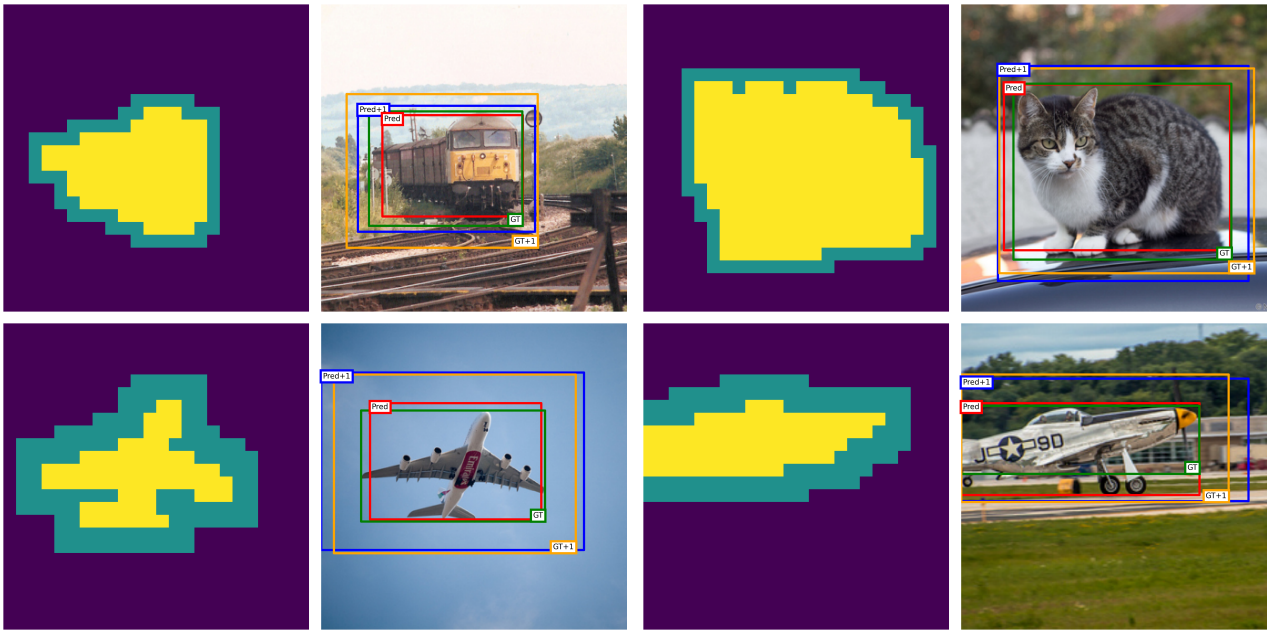

Figure 9. **Examples of the object extension experiment.** Each image shows the input with its mask in pixel space. The yellow region indicates the original mask, while the green region denotes the padding p added by sampling tokens from the object. The top row corresponds to $p = 1$ and the bottom row to $p = 2$. We display both the predicted and ground-truth bounding boxes for the original and the extended object. The predicted boxes expand consistently with the mask, suggesting that the model containerizes object tokens to define spatial boundaries.

7.5. Ablation Results for Global and Local Views

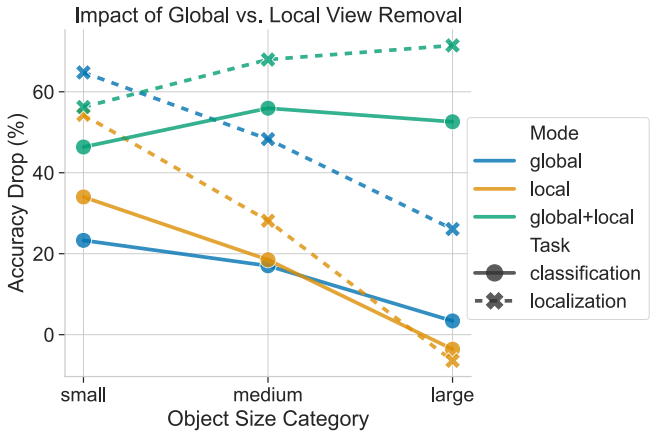

Figure 10. Performance drop for classification and detection when ablating global, local, or both image views across object sizes. We follow the COCO [18] convention and group targets according to their bounding-box area: *small* ($< 32^2$), *medium* ($32^2 \leq$ area $\leq 96^2$), and *large* ($> 96^2$).

8. Positional Information

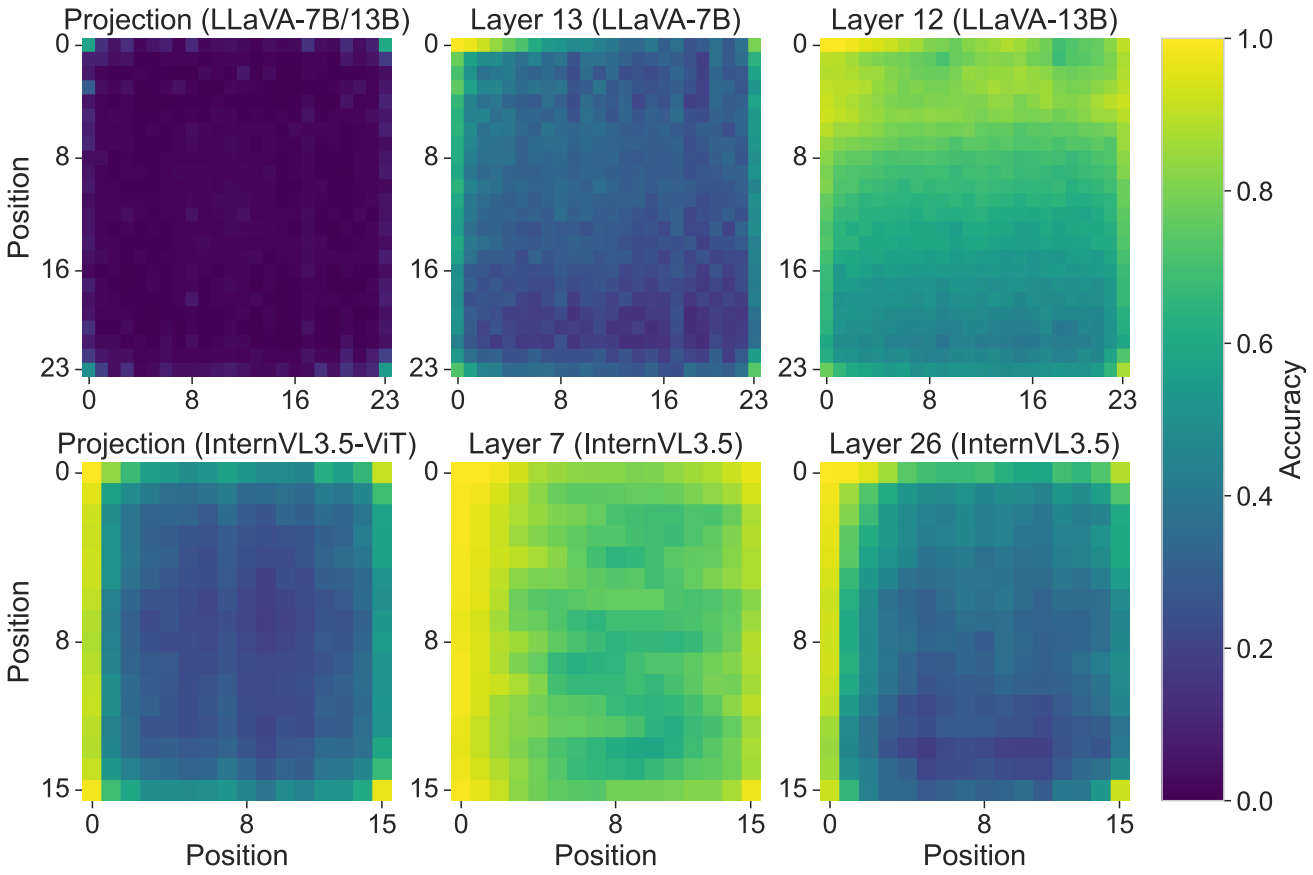

Figure 11. **Heatmap visualizations of positional decoding accuracy at selected stages of LLaVA.** Each heatmap shows the probability of correctly predicting the position of a visual token in the 24×24 grid. The multimodal projection retains positional information mainly at the four corners, effectively marking the image boundaries needed to infer its dimensions. Accuracy then increases within the LLM, becoming highest in the early-mid-layers (layer 13 in LLaVA-7B, layer 12 in LLaVA-13B and layer 7 for InternVL3.5). Corners and boundary regions remain the most reliably recovered, indicating that they serve as anchors for reconstructing the global spatial layout.

9. Attention Blocking

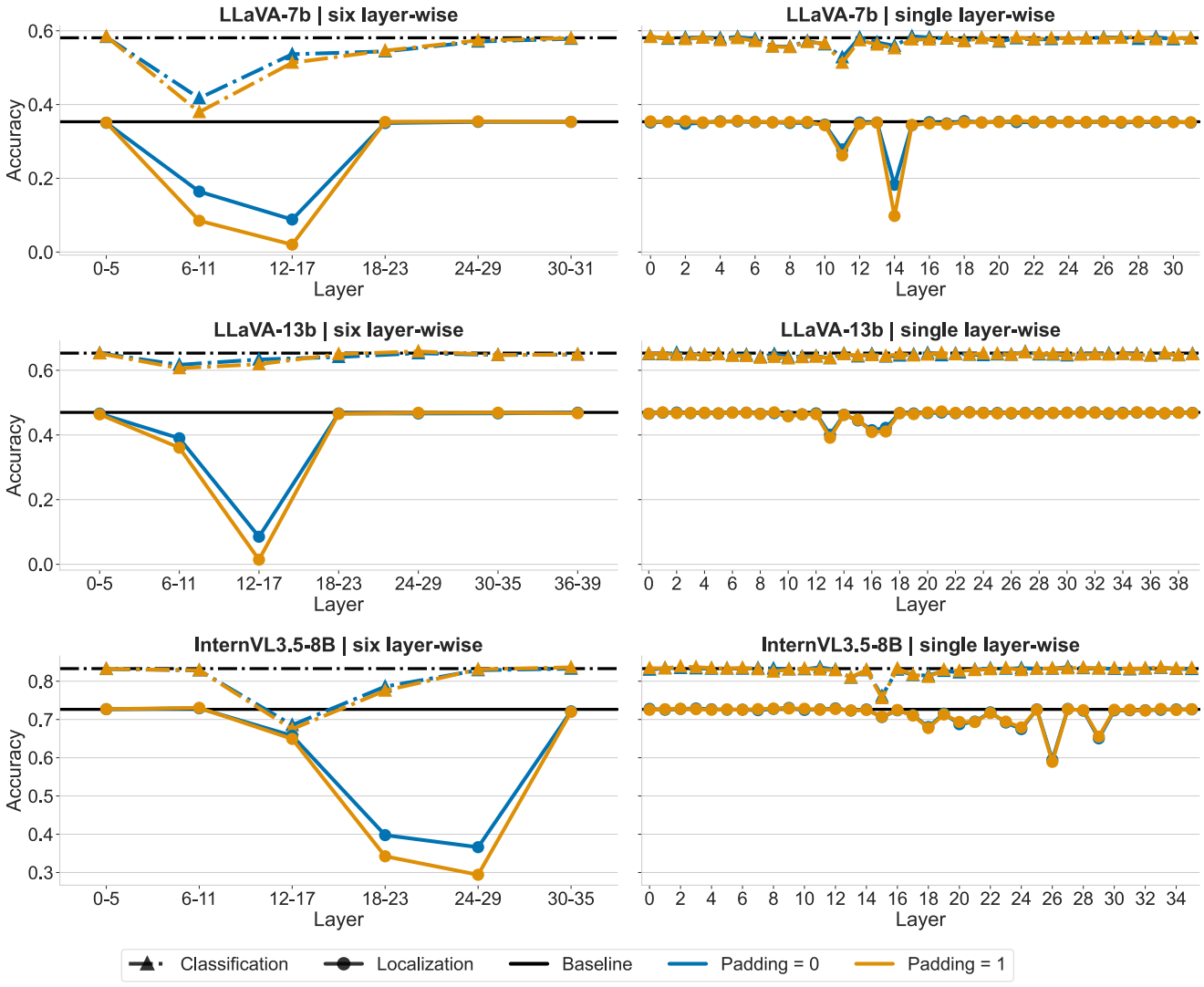

Figure 12. **Attention blocking across layers.** We measure classification and localization accuracy when blocking attention from post-image tokens to object tokens, either in groups of six layers (left) or one layer at a time (right). Localization accuracy drops sharply in early–mid layers for LLaVA models and in mid–late layers for InternVL, while classification remains largely stable across the network. Blocking attention to padding tokens causes an additional decrease in localization performance.

10. Causal Mediation Analysis

10.1. Visualization of CMA Method

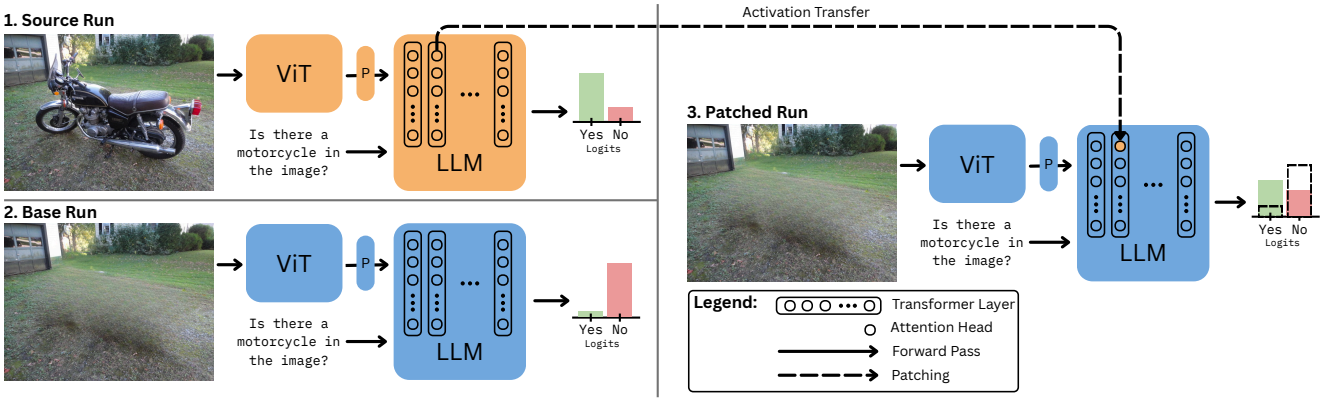

Figure 13. **Causal mediation via activation patching.** We compare three model runs: (1) the source run, where the object is present and the model produces the correct answer; (2) the base run, where the object is removed and the model fails; and (3) the patched run, where we transfer hidden activations from a selected attention head in the source run into the base run. Improvements in the patched prediction indicate that the transferred head carries task-relevant information. This procedure is repeated head-wise to quantify each head's causal contribution.

10.2. Additional Results for LLaVa-13b

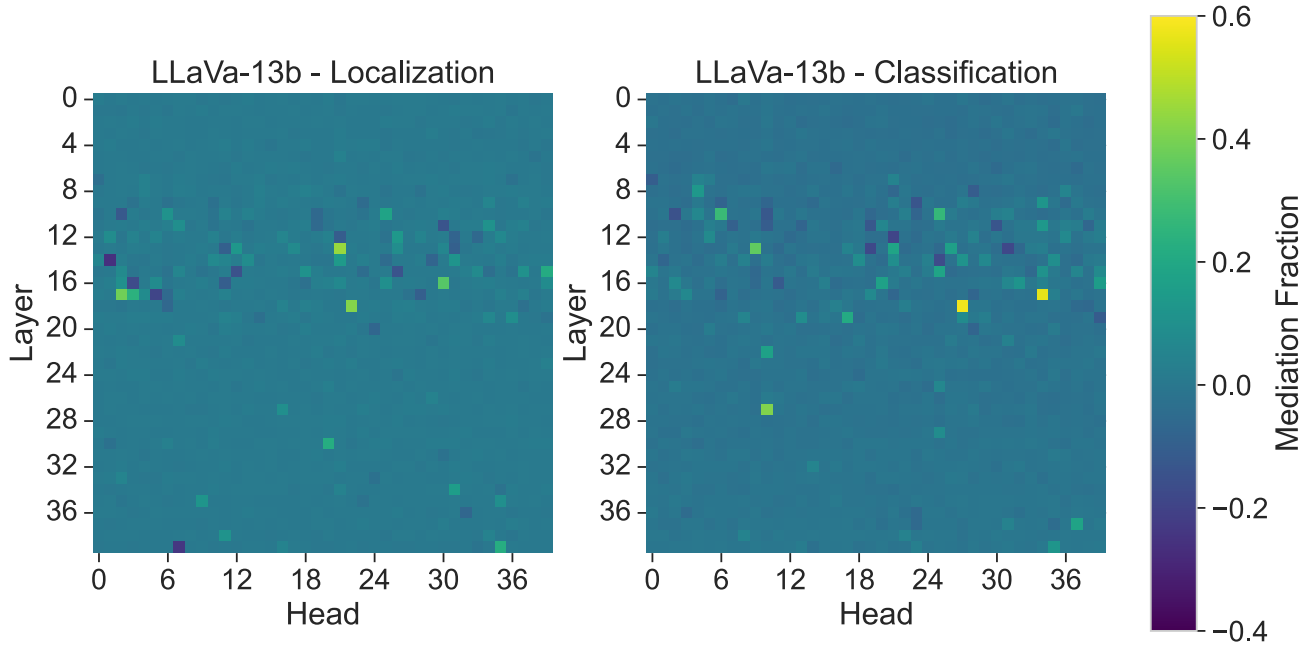

Figure 14. **Causal Mediation Analysis for LLava13b.** Mediation Fraction scores for every attention head across all layers, shown separately for the detection and classification tasks.

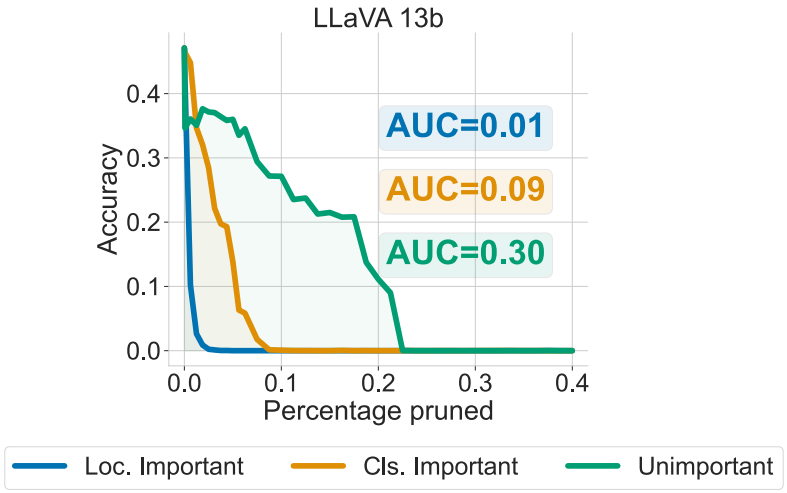

Figure 15. Localization accuracy under cumulative head ablation. Attention heads are ranked by their mean MF and progressively removed. Normalized AUC scores enable method comparison.