# An Existential Crisis: Investigating "Existence Bias" in VLM Spatial Reasoning

**Kaushik Karthikeyan** [*1] **Tengerleg Enkhtuvshin** [*1] **Leyla Yaayladere** [*1] **Sarah Verreault** [*1]

## Abstract

Vision–Language Models perform well at recognizing objects but struggle to grasp spatial relations. We analyze `LLaVA-1.5-7B` on a controlled synthetic 2D dataset using logit lens, linear probing, and attention diagnostics. On balanced binary relational questions, the model exhibits a strong *yes bias*. Logit lens and occlusion analyses show this bias depends on visual context rather than language priors alone. Linear probing shows that the correct relational label becomes linearly decodable by mid-layers, despite the final output contradicting this internal information. Attention patterns indicate reliance on object existence rather than relational cues, suggesting an underlying *existence bias*. We distinguish *yes bias* as the observable tendency to answer "*yes*" and *existence bias* as the mechanism driving this behavior. The code of our project is publicly available.[1]

## 1. Introduction

Vision–Language Models (VLMs) extend large language models by enabling joint reasoning over images and text. While VLMs have demonstrated impressive accuracy in identifying entities within images, their ability to understand relationships between those entities remains a significant challenge. Hou et al. (2025) found that apparent success often stems from a reliance on pre-learned knowledge rather than a genuine understanding of the relationships themselves. This gap between entity recognition and relational understanding remains poorly understood at a mechanistic level. This project aims to bridge this gap by addressing the following ***Research Question***:

*Why do VLMs succeed at entity recognition but fail at relational reasoning?*

To answer this, we analyze how relational decisions are formed internally in a vision–language model under controlled visual conditions. Our main contributions are:

1. A controlled 2D synthetic dataset for spatial relational reasoning with precise annotations and balanced question formats.
2. Empirical evidence of a strong *yes bias* in binary relational reasoning, characterized via accuracy.
3. Layer-wise diagnostics (logit lens, linear probing) showing when answer information becomes decodable and how this contrasts with the model's final behavior.
4. Quantitative and qualitative attention analysis: CoM and IoU metrics reveal object grounding, while attention masking experiments highlight persistent false positives driven by *existence bias*.

## 2. Related Work

**Spatial Relational Reasoning.** Kamath et al. (2023) introduced controlled benchmarks and found that state-of-the-art VLMs perform poorly on spatial relational queries. Tong et al. (2024) demonstrate persistent positional and relational deficiencies, showing that scaling alone does not yield reliable improvements.

**Attention Analysis.** Chen et al. (2025) show that simply increasing attention to image tokens does not improve VLM performance; instead, effective models adaptively modulate visual attention based on context and confidence. Liu et al. (2025) argue that while VLMs encode visual evidence in deeper layers, they often fail to translate perception into reasoning and decision making. However, these studies do not isolate binary spatial verification under controlled visual complexity while jointly comparing logit-level dynamics, linear decodability, and attention alignment. Our work focuses on this controlled setting to connect failure modes to layer-wise signals and attention behavior.

## 3. Models and Methodology

**Model.** We analyze `LLaVA-1.5-7B` (Liu et al., 2023), which encodes an image into patch-level visual tokens via a CLIP-based vision encoder and projects them into the Vicuna-1.5 language embedding space using a lightweight MLP, concatenated with text tokens for joint decoding.

**Task and Data Generation.** We construct a controlled synthetic 2D dataset to isolate spatial relational reasoning from linguistic and dataset biases. Images are rendered at $336 \times 336$ on a $24 \times 24$ grid (patch size $14 \times 14$), matching LLaVA-1.5's visual token granularity. The dataset contains

[1] github.com/sarah-verr/DeepLearningProject

five complexity levels (0–4) that increase object count (2–4), object scale (single- to multi-patch), and spatial interactions (Appendix A.2). Each scene is annotated with ground-truth bounding boxes, patch indices, and pairwise relations. We generate balanced yes/no relational questions across {left, right, above, below}, alongside entity-recognition, relational-attribute, and directional question variants (Appendix A.3).

**Prompt Design.** Prompts explicitly constrain the answer space (e.g., *yes/no* for binary verification) to reduce formatting variance and linguistic ambiguity (Appendix B).

**Mechanistic Tools.** We use two complementary layer-wise tools. **Logit lens** projects intermediate hidden states into the output vocabulary using the model's unembedding matrix. Belrose et al. (2025) shows that early-layer projections can be misaligned with the final decoding space, motivating cautious interpretation of outputs. **Linear probing** trains a per-layer logistic-regression classifier on hidden states (using the final input token that conditions the answer) to test whether the ground-truth label is linearly decodable at each layer.

**Attention Analysis.** We analyze attention patterns to test whether LLaVA attends to visually relevant regions when answering spatial relations. For each binary relational question, we identify the *relation token* (e.g., *left/right/above/below*), the *subject token* and the *object token* in the prompt and treat it as the source token. We extract its attention over image tokens at every layer and head, and analyse the following metrics:

- **Image-vs-text attention fraction:** the fraction of attention assigned to image versus text tokens.
- **Attention entropy:** spatial dispersion of attention across the image.
- **Center-of-Mass (CoM) distance** and **Intersection over Union (IoU):** spatial alignment between attention maps and ground-truth object regions.

Exact computation procedure is described in Appendix E.1.

## 4. Results

### 4.1. Task Performance and Failure Modes

We evaluated `LLaVA-1.5-7B` on four question types to measure entity recognition and relational reasoning. The model performs relatively well on entity recognition questions across all complexity levels, while performance on relational tasks varies by response type (Table 1).

Across all question types, accuracy peaks at Level 2, where the two objects are larger and well separated. Performance drops at level 0, suggesting a potential merging effect where the model perceives two adjacent single-patch objects as a single entity. This interpretation is supported by frequent "1" responses to object-count queries at Level 0 (Table 3).

A clear contrast emerges across task formats. Binary Existence Questions reach 76.6% accuracy overall, while Binary Relational Questions remain near-random chance 50%. Performance improves when moving away from yes/no response format. Directional Relational Questions show a bias toward certain relation types, with consistently higher accuracy for *right* and *above*. Notably, Relational Attribute Questions, which combine entity recognition with relational reasoning in a multiple-choice format, achieve substantially higher accuracy than their random baseline. This suggests that mixing entity recognition with relational reasoning, and moving away from binary choices, improves performance.

We focus our remaining analysis on Binary Relational Questions to understand why this specific format fails. Both Binary Entity Recognition and Binary Relational Questions exhibit a bias toward predicting "*yes*" but this bias is significantly more pronounced in relational questions, as shown in the confusion matrices (Figure 5). Understanding this differential *yes bias* is central to explaining the model's relational reasoning failure.

To test whether this behavior depends on visual input, we evaluate controlled occlusions. With a blank image, the model predicts "no" in most cases (Appendix Figure 6b). When either the subject or object mentioned in the question is removed, the model's responses vary with scene complexity. Under occlusion, the "False Positives" rate decreases by 50-60% in lower-complexity scenes, but this improvement drops significantly to only 5% decrease in higher-complexity scenes (Table 5). Next, we compare these logit-level dynamics with linear decodability of the same representations.

### 4.2. Logit Lens

We apply logit lens to estimate the layer-wise probability of predicting "yes" (Figure 1). Across all complexity levels, the logit-lens signal exhibits a strong *yes bias* in most layers. The probability assigned to "yes" increases sharply in the mid layers, while "no" remains suppressed for most of the network.

### 4.3. Linear Probing Results

Probe accuracy increases sharply around layer 16 across all levels, followed by a stable plateau (Figure 2).The plateau accuracy range ($\sim$80%) exceeds the model's performance on the same task by around 30%.

### 4.4. Attention Results

We analyze attention patterns to examine how visual information is accessed during relational reasoning and how this

*Table 1.* Relational Reasoning performance by complexity level and relation type. Exact Match Accuracy (%). †Relational Attribute Questions have an even split of 4- and 9-option MC questions. The expected random guess accuracy is $\frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{9} \approx 18\%$

| Question Answer Set | Uniform Random Baseline | Synthetic Data Levels | | | | | Relation Types | | | | Overall |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Level 0 | Level 1 | Level 2 | Level 3 | Level 4 | Left | Right | Below | Above | |
| **Binary Existence Questions** *(Yes/No choices)* | 50.0 | 80.0 | 81.2 | **84.4** | 68.1 | 69.4 | NA | NA | NA | NA | **76.6** |
| **Relational Binary Questions** *(Yes/No choices)* | 50.0 | 51.3 | 51.3 | **54.0** | 51.9 | 50.6 | 52.4 | 50.5 | 51.9 | 50.5 | 51.3 |
| **Relational Attribute Questions** *(Multiple choice (MC): 4 or 9 options)* | **18.1**† | 56.3 | 54.4 | **77.7** | 65.7 | 52.1 | 61.7 | 65.1 | 64.4 | 59.6 | **62.7** |
| **Relational Direction Questions** *(2-way directional: Left/Right or Above/Below choices)* | 50.0 | 52.5 | 67.5 | **75.0** | 67.6 | 67.3 | 60.6 | **82.6** | 44.7 | **81.3** | 67.4 |

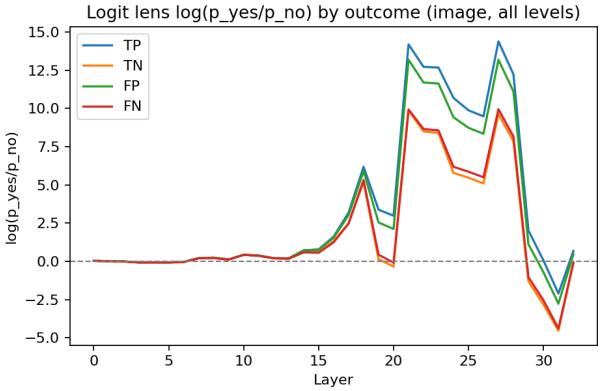



*Figure 1.* Mean per-layer log-odds of yes vs no, computed from softmax mass over yes/no tokens; curves are split by TP/TN/FP/FN, with the zero line marking equal preference. Positive y-values indicate yes preference.

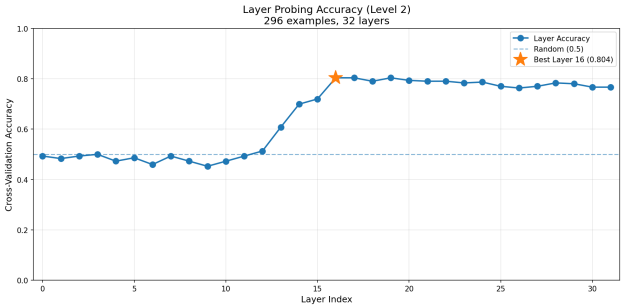



*Figure 2.* Layer-wise probing accuracy for the yes–no task, averaged over all level 2 questions. Accuracy rises sharply at layer 16, where it peaks (0.804), and then plateaus.

varies across layers. Our findings confirm established phenomena Kaduri et al. (2025): text-image attention allocation peaks in early layers (Appendix Figure 9a, Section E.2), while attention entropy over image tokens starts high in early layers and decreases in intermediate layers (Appendix Figure 9b, Section E.3).

**Object Grounding Signals.** To assess how attention targets the referenced entities during relational reasoning, we analyze subject- and object-conditioned grounding signals. The CoM heatmaps (Appendix Fig. 13, Fig. 14) show that attention shifts toward the mentioned entity, with the strongest grounding emerging in layers 11–15. The IoU maps exhibit the same mid-layer trend (Appendix Fig. 16, Fig. 17), with a small subset of heads achieving higher overlap with the ground-truth object mask.

**Attention Masking.** To test the functional impact of spatial attention, we mask attention on the region opposite the queried relation relative to the subject (Appendix 11). In low-complexity scenes, this yields a ∼10% accuracy increase (Appendix Table 6), mirroring our occlusion results (Appendix Table 5). In high-complexity scenes, this effect vanishes, likely because the presence of remaining objects continues to trigger the *yes bias*.

## 5. Discussion

**Terminology.** *Yes bias* denotes the observable tendency to answer "yes" on binary queries (manifesting as elevated false positives). *Existence bias* denotes our proposed mechanism: the model treats the presence of relevant entities as sufficient evidence for the queried relation.

Our results show that failures in binary relational reasoning arise not from missing visual information but from how `LLaVA-1.5` converts this information into a yes or no decision. The model detects and localizes objects reliably, yet it struggles to reject relations, which produces a consistent *yes bias*. Scenes containing any shapes increase the likelihood of predicting "yes," even when the queried relation is false.

Occlusion experiments support this interpretation. With no shapes present, the model predicts mainly "no" (Figure 6). When multiple objects are present, removing the subject or object often does not reduce the "yes" rate (Table 5), hinting that the model treats unrelated shapes as evidence for the relation.

Layer-wise analyses further clarify how this bias forms. Linear probing reveals that the correct relational label is linearly decodable in mid-to-late layers (Figure 2) even when the model answers incorrectly, indicating a disconnect between internal representations and final decisions. This aligns with Yu & Lee (2025), where late layers in LLaVA prioritize output formatting rather than semantic integration. However, linear decodability does not imply that the model causally uses this information to form its final decision.

Complementing the layer-wise results, the CoM and IoU metrics show that the model reliably localizes both the subject and object once they are referenced, but this grounded information is not consistently integrated into the final relational decision. Even when both entities are correctly attended, the model often defaults to interpreting their presence as evidence for the relation, consistent with an underlying existence bias.

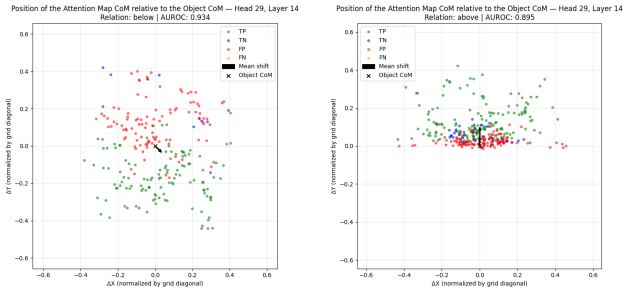
Spatial attention masking yields only small improvements, indicating a persistent *yes bias* in which the mere presence of multiple objects, rather than their spatial configuration, induces affirmative predictions.

To investigate the source of false positives, we analyzed attention CoM vectors relative to the object. Specifically, we examined the vector from the object to the attention CoM to determine where the model focuses when predicting relational answers. While AUROC values across heads and layers suggest that attention direction relative to the object predicts answer correctness (Figure 12), closer inspection reveals that this effect is largely driven by an entity existence bias.

False positives often occur when attention is directed toward the subject opposite the queried relation. In such cases, even though the model attends to a spatially incorrect region, the presence of the subject is sufficient to trigger a "yes" prediction. This suggests that the observed separability primarily reflects entity detection rather than true spatial relational reasoning (Figure 3a). After applying the attention mask, attention clusters around the object CoM, indicating that the model relies on the object itself to produce affirmative responses.

Overall, the model perceives spatial relations correctly at the representation level but fails to reliably reject incorrect relations in cluttered scenes. This *yes bias* suggests that improving VLM relational reasoning will require mechanisms that explicitly enforce subject-object grounding and relational verification rather than relying on general object presence as evidence.

**Limitations.** Our findings are based on a controlled synthetic 2D dataset designed to isolate spatial relations, so the results may not fully transfer to natural images with richer visual semantics. We study a single VLM, and it is unclear




(a) Without masking: false positives cluster on the side opposite the queried relation, showing the model attends to the subject yet predicts "yes."

(b) With masking: false positives cluster around the object, indicating the model predicts "yes" by attending to the object itself.

*Figure 3.* Existence bias in attention. False positives arise from attending to the wrong side (a) or directly to the object (b), showing reliance on object presence rather than relational reasoning.

how broadly *existence bias* generalizes across architectures or training regimes. Our attention masking intervention is coarse and applied uniformly across layers/heads. The mask is a heuristic, not a learned mask. A more targeted or learned intervention may yield different effects. Finally, we focus on binary relational verification rather than complex compositional or multi-relation reasoning. Our findings are only diagnostic and correlational.

## 6. Conclusion

We presented a controlled study of spatial relational reasoning in `LLaVA-1.5-7B` using a synthetic 2D dataset. Despite strong entity recognition, the model performs near chance on balanced binary relational verification, exhibiting a pronounced *yes bias*: a tendency to answer "yes" regardless of the true spatial configuration. We provide evidence consistent with an underlying *existence bias*, in which the presence of objects in the scene is treated as evidence that the queried relation holds, even when the specific subject–object relation is false. Occlusion experiments support this view, as removing the referenced entities reduces false positives only when no other objects remain, indicating that object presence rather than relational structure drives affirmative predictions.

Layer-wise analyses show that mid-layer representations contain linearly decodable information about the correct relational label, but this signal does not consistently propagate to the final output. Spatial attention masking yields only modest improvements, suggesting that improving attention alone is insufficient to overcome the underlying existence bias. Overall, these results suggest that, although hidden representations contain sufficient relational information, *existence bias* dominates the final output, limiting reliable spatial relational reasoning.

## References


Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens, 2025. URL https://arxiv.org/abs/2303.08112.

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. In Singh, A., Fazel, M., Hsu, D., Lacoste-Julien, S., Berkenkamp, F., Maharaj, T., Wagstaff, K., and Zhu, J. (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 9910–9932. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/chen25cr.html.

Hou, Y., Gile dereli, B., Tu, Y., and Sachan, M. Do vision-language models really understand visual language?, 2025. URL https://arxiv.org/abs/2410.00193.

Kaduri, O., Bagon, S., and Dekel, T. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, June 2025.

Kamath, A., Hessel, J., and Chang, K.-W. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023. URL https://arxiv.org/abs/2310.19785.

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.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning, 2023.

Liu, Z., Chen, Z., Liu, H., Luo, C., Tang, X., Wang, S., Zeng, J., Dai, Z., Shi, Z., Wei, T., Dumoulin, B., and Tong, H. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms, 2025. URL https://arxiv.org/abs/2510.17771.

Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. URL https://arxiv.org/abs/2401.06209.

Yu, Z. and Lee, Y. J. How multimodal llms solve image tasks: A lens on visual grounding, task reasoning, and answer decoding, 2025. URL https://arxiv.org/abs/2508.20279.

# A. Synthetic Data

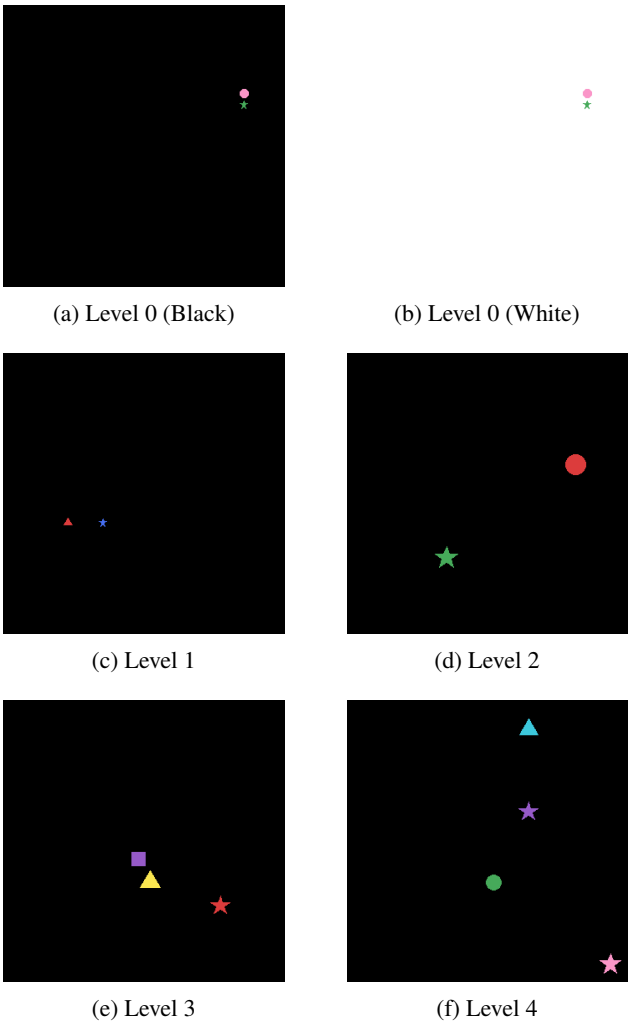
## A.1. Image Background

Each scene is rendered on both black and white backgrounds to verify that model attention grounds to entities rather than background artifacts. No significant difference is observed between black and white backgrounds (Table 4).

## A.2. Complexity Level Specifications

Our dataset comprises five levels of increasing visual complexity:

- **Level 0:** Two small objects (single $14 \times 14$ px patch each), positioned in horizontally or vertically adjacent grid cells. Objects are strictly centered within cells to establish baseline spatial relations (e.g., "*left of*" or "*above*").
- **Level 1:** Two small objects that can be placed anywhere in the same row or column, but not necessarily adjacent. Tests whether the model can track relations across distances.
- **Level 2:** Two larger objects spanning multiple patches (up to $4 \times 4$ grid cells), positioned anywhere in the image, not constrained to same row/column. Tests scale invariance and diagonal spatial reasoning.
- **Level 3:** Three objects with properties similar to Level 2. Introduces multi-object scenes while maintaining focus on pairwise spatial relationships.
- **Level 4:** Four objects that may touch or overlap. Tests robustness to visual complexity and occlusion while querying the same spatial relations.



(a) Level 0 (Black) (b) Level 0 (White) (c) Level 1 (d) Level 2 (e) Level 3 (f) Level 4

*Figure 4.* Synthetic data samples (Levels 0–4). Each row displays increasing complexity in object count and spatial configuration.

## A.3. Question-Answer Examples and Distribution

**Example Question-Answer Pairs.** Given an image containing a *pink circle* at the top and a *green star* at the bottom (Figure 4a), the following question types are generated:

**Entity Recognition:** These questions verify the model's capability to identify object presence and attributes. This serves as a sanity check and reveals how recognition varies across complexity levels.

- Yes/No: "*Is there a pink circle?*" → "*Yes*"
- Yes/No: "*Is there a blue square?*" → "*No*"
- Attribute: "*What color is the star?*" → "*Green*"
- Attribute: "*What shape is the pink object?*" → "*Circle*"
- Attribute: "*How many objects are present in the image?*" → "*2*"

**Binary Relational Reasoning:** These test relational reasoning in its simplest form with balanced *yes*/*no* questions. This format minimizes linguistic ambiguity to isolate the model's ability to verify spatial relationships.

- "*Is the pink circle above the green star?*" → "*Yes*"
- "*Is the green star above the pink circle?*" → "*No*"

**Attribute Relational Reasoning:** These probe whether the model can combine entity recognition with spatial context by querying attributes based on location.

- "*What is the color of the object below the pink circle?*" → "*Green*"
- "*What is the shape of the object above the green star?*" → "*Circle*"

**Directional Relational Reasoning:** These directly query the spatial direction between entities. Questions focus on either vertical or horizontal axes independently and do not currently cover diagonal relations (e.g., *top-right*).

– "*Where is the pink circle in relation to the green star?*" → "*Above*" (*vertical*)

– "*Where is the green star in relation to the pink circle?*" → "*Below*" (*vertical*)

Table 2 details the distribution of ground-truth spatial relationships across the three relational question types. All sets are balanced across the four directions (*left*, *right*, *above*, *below*), and binary questions maintain a strict *50% yes – 50% no* distribution.

*Table 2.* Distribution of ground-truth spatial relationships across the three Question-Answer sets.

| Question Answer Set | Relationship Types | | | |
|---|---|---|---|---|
| | **Left** | **Right** | **Above** | **Below** |
| **Binary Relational Questions** *(Yes/No questions)* | 852 | 852 | 824 | 824 |
| **Relational Attribute Questions** *(Multiple choice)* | 292 | 284 | 316 | 304 |
| **Relational Direction Questions** *(2-way directional)* | 426 | 426 | 412 | 412 |

## B. Prompting

All prompts are designed to constrain model outputs to predefined vocabularies, ensuring consistent single-word answers. We verified through manual inspection of model outputs that LLaVA-1.5 reliably follows these constraints and produces exactly one word from the specified options. Given the brittleness of prompt engineering, we adopt straightforward templates to minimize linguistic confounds, avoiding chain-of-thought prompting, few-shot examples, or other complex prompting strategies.

All prompts are formatted using the processor's `apply_chat_template` method following the official `LLaVA-1.5-7B` implementation[2] , ensuring proper conversation structure for the model.

### B.1. Entity Recognition Prompts

*Existential Yes/No:* Questions about object presence are answered with "yes" or "no":

```
Answer the following question about
the image with either "yes" or "no".
QUESTION: Is there a pink circle?
```

[2]https://huggingface.co/llava-hf/llava-1.5-7b-hf

*Attribute Questions:* Questions about color, shape, or count are constrained to specific vocabularies:

```
Allowed colors: "red", "blue", "green",
"yellow", "purple", "cyan", "orange",
"pink" or "lime".
Answer the following question with
exactly one option from the allowed
colors list.
QUESTION: What color is the star?
```

For shape queries, allowed options are "square", "circle", "triangle", or "star". For counting queries, allowed numbers are "1", "2", "3", or "4".

### B.2. Binary Relational Reasoning Prompts

Binary relational questions constrain answers to "yes" or "no":

```
Use the spatial layout in the image
to answer the following relational
reasoning question with either "yes"
or "no".
QUESTION: Is the green star above
the pink circle?
```

### B.3. Attribute Relational Reasoning Prompts

These prompts combine spatial context with attribute queries, constraining outputs to the same color/shape vocabularies as Entity Recognition:

```
Allowed colors: red, blue, green,
yellow, purple, cyan, orange, pink,
lime.
Use the spatial layout to answer the
following question with exactly one
option from the allowed colors list.
QUESTION: What is the color of the
object below the pink circle?
```

### B.4. Directional Relational Reasoning Prompts

Directional questions are constrained to either vertical or horizontal options (not both simultaneously). *Vertical questions:*

```
Allowed relationships: "above" or
"below".
Use the spatial layout in the image
to answer the following question with
exactly one option from the allowed
relations list.
QUESTION: Where is the pink circle
in relation to green star?
```

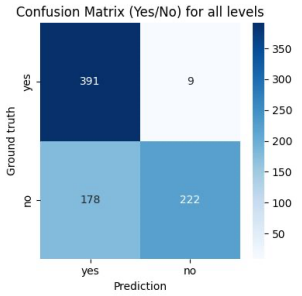
## C. Preliminary Analysis

*Table 3.* Entity Recognition performance by complexity level. Exact Match Accuracy (%). Random baselines: Existence 50%, Color ∼11%, Shape 25%, Number 25% (9 color, 4 shape, 4 number choices).

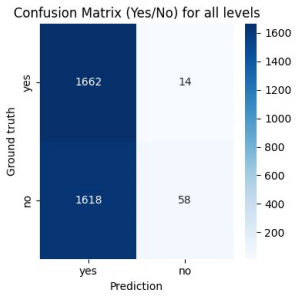
| Complexity Level | Question Type | | | |
|---|---|---|---|---|
| | Existence | Color | Shape | Number |
| Level 0 | 80.0 | 44.2 | 58.8 | 2.5 |
| Level 1 | 81.2 | 52.1 | 58.8 | 15.0 |
| Level 2 | 84.4 | 63.2 | 68.8 | 60.0 |
| Level 3 | 68.1 | 66.7 | 68.8 | 30.0 |
| Level 4 | 69.4 | 64.3 | 75.0 | 60.0 |
| **Overall** | 76.6 | 58.5 | 66.0 | 33.5 |

*Table 4.* Accuracy based on colour of background, indicating background color does not impact performance

| Question Answer Set | Background | |
|---|---|---|
| | Black | White |
| **Binary Existence Questions** *(Yes/No choices)* | 80.0 | 81.2 |
| **Binary Relational Questions** *(Yes/No questions)* | 51.7 | 50.9 |
| **Relational Attribute Questions** *(Multiple choice)* | 61.5 | 63.9 |
| **Relational Direction Questions** *(2-way directional)* | 66.9 | 67.8 |

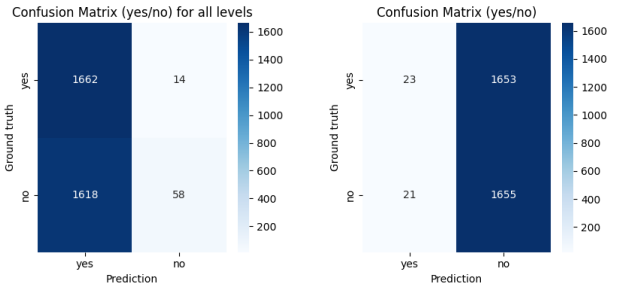



(a) Binary (Yes/No) Existence Questions




(b) Binary (Yes/No) Relational Questions

*Figure 5.* Confusion matrices highlighting the scale of the *yes bias*. In existence questions (a), the model retains a degree of discriminative ability. In contrast, relational questions (b) show a near-total collapse, with the model producing False Positives in ∼ 96.5% of negative samples, defaulting to an affirmative response regardless of the actual spatial configuration.




(a) Input images contain shapes; the model predominantly predicts 'yes'.

(b) Input images contain no shapes (plain black or white images); the model predominantly predicts 'no'.

*Figure 6.* Comparison of confusion matrices illustrating the effect of object presence on model predictions. (a) When shapes are present, the model mainly predicts 'yes', indicating reliance on visual cues. (b) When no shapes are present, the model mainly predicts 'no', showing that the previously observed 'yes' bias is not driven solely by language priors.

## D. Linear Probing

For each transformer layer $\ell$, we extract the hidden state of the final input token, which conditions the model's next-token prediction. Let $h^{(\ell)} \in \mathbb{R}^d$ denote this representation. We train a separate logistic-regression probe $f^{(\ell)} : \mathbb{R}^d \to \{0, 1\}$ at each layer, using $h^{(\ell)}$ as input and the ground-truth yes/no label as the supervision target.

To obtain robust estimates of linear decodability, we evaluate each probe using $k$-fold cross-validation (with $k = 5$). The classifier is optimized with a cross-entropy loss and L2 regularization within each fold. The reported accuracy for each layer is the mean across folds.

This procedure yields a layer-wise estimate of how well the correct answer can be recovered from the model's hidden states. Results for probing on the level-2 dataset are shown in Figure 2.

## E. Attention Analysis

### E.1. Computation of Attention Metrics

We consider three types of source tokens—source, target, and the relational phrase—depending on the aspect of the model being analyzed (Figure 7). The relational phrase is used to study how the model tracks spatial relationships expressed in the query. When the source phrase tokenizes into multiple tokens, we average the corresponding attention maps. The subject and object source tokens are used to analyze how the model grounds attention to the referenced entities in the image.

*Table 5.* False Positive Data analysis: Per-level yes-rate for occluded FP-filtered data. Original FP baseline is 100%. Accuracy improvement is computed as 1 − overall yes rate relative to the original False Positive baseline. The goal is to decrease False Positives to 0%.

| Level | Overall Yes Rate (Avg) | Subject Removed | Object Removed |
|---|---|---|---|
| Level 0 | 0.4093 ↓ 0.5907 | 0.4125 ↓ 0.5875 | 0.4062 ↓ 0.5938 |
| Level 1 | 0.4219 ↓ 0.5781 | 0.4625 ↓ 0.5375 | 0.3812 ↓ 0.6188 |
| Level 2 | 0.5051 ↓ 0.4949 | 0.5676 ↓ 0.4324 | 0.4426 ↓ 0.5574 |
| Level 3 | 0.7978 ↓ 0.2022 | 0.7913 ↓ 0.2087 | 0.8043 ↓ 0.1957 |
| Level 4 | 0.9521 ↓ 0.0479 | 0.9523 ↓ 0.0477 | 0.9518 ↓ 0.0482 |

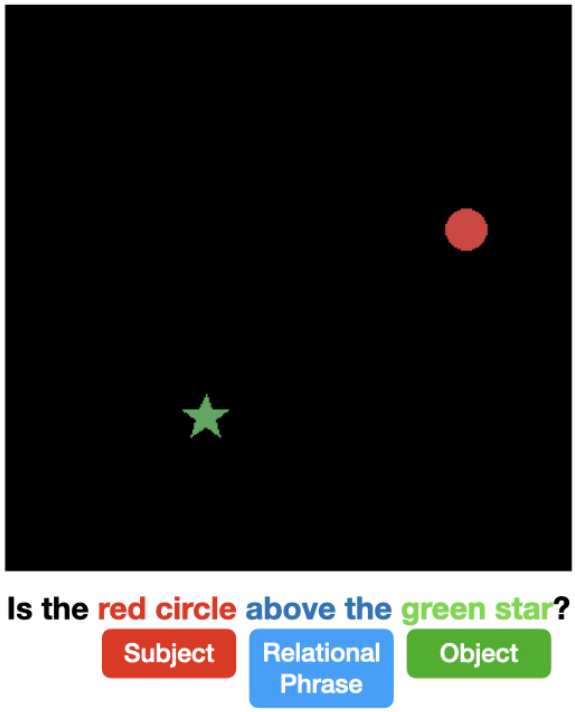



*Figure 7.* Overview of the three types of source tokens used to retrieve attention maps for computing attention metrics.

The resulting attention maps are then used to compute the attention metrics defined in Section 3, as summarized below:

1. **Attention Fraction** We use the softmax-normalized attention weights and sum the attention assigned to all image tokens. This quantity is divided by the total attention mass to obtain the fraction of attention placed on the image. The fraction of attention placed on text is computed as 1− image attention fraction.

2. **Attention Entropy** The attention weights over image tokens are re-normalized to form a probability distribution restricted to the image tokens. We then compute the entropy of this distribution.

3. **Center-of-Mass (CoM) Distance** The attention CoM is computed as a weighted average of the 2D patch coordinates, using the attention weights. The object CoM is computed by assigning equal weight to all image patches covered by the object mask and averaging their 2D coordinates. The CoM shift is defined as the vector from the object CoM to the attention CoM, capturing the spatial alignment between attention and the relevant object. The distance is the magnitude of this vector. An example visualisation is shown in Figure 8.

4. **Intersection-Over-Union (IoU)** We measure localization by computing the Intersection-over-Union (IoU) between a thresholded attention map and the ground-truth object mask. Attention weights over image patches are first normalized, and the top-$k\%$ of mass is selected to form a binary attended region. Let $A$ denote this attention mask and $G$ the ground-truth patch mask. IoU is then computed as:

$$\text{IoU} = \frac{|A \cap G|}{|A \cup G|}.$$

A higher IoU indicates better alignment of the model's attention with the true object location.

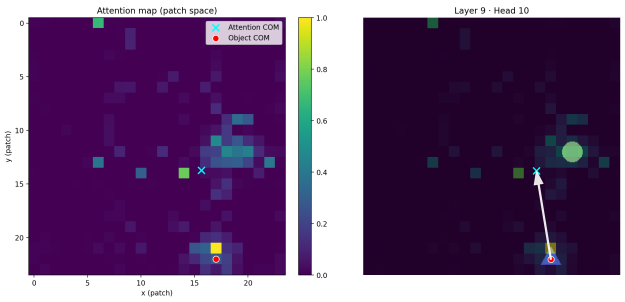
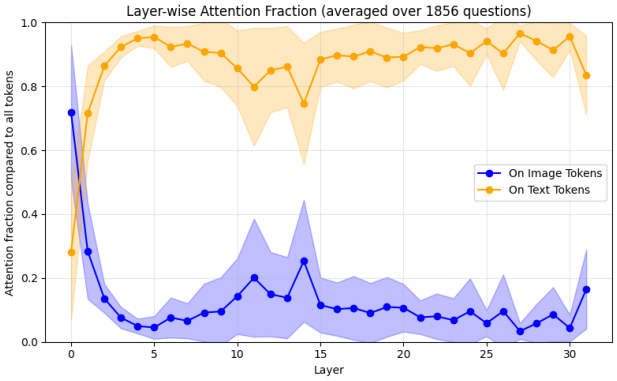
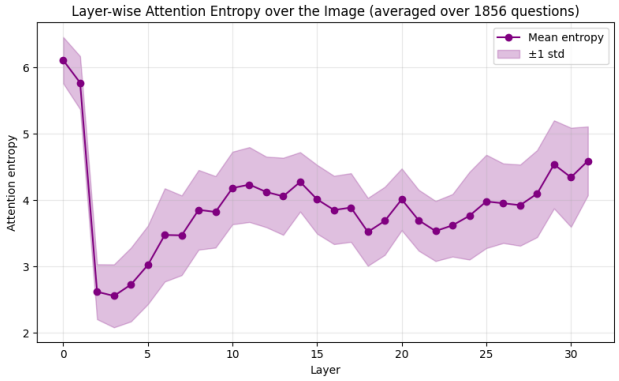



*Figure 8.* Illustration of the center-of-mass (CoM) computation, showing the vector from the object CoM to the attention map CoM.




(a) Layer-wise distribution of attention between text and image tokens, averaged over all heads per layer. Relational Phrase tokens were used as the source for attention.




(b) Layer-wise attention entropy for image tokens, using the relational phrase in the question as the source token. Higher values indicate more evenly distributed attention across image regions.

*Figure 9.* Visual attention patterns across layers in `LLaVA-1.5-7B`.

### E.2. Text-Image Attention Allocation

Figure 9a the attention to image tokens peaks in the early layers (0–1), re-emerges in layers 10 to 15, and rises again in the final layer, consistent with Kaduri et al. (2025). Attention is largely focused on text tokens, despite ∼10× more image tokens.

### E.3. Attention Entropy over Image Tokens

To characterize how visual attention is distributed spatially, we compute the entropy of the attention distribution over image tokens. Early layers exhibit high entropy, indicating diffuse attention spread broadly across the image (Figure 9b). In contrast, intermediate layers show reduced entropy, reflecting more spatially concentrated attention patterns. However, many heads focus on irrelevant shapes or sink regions, which is consistent with visual attention sinks (Kang et al., 2025).

### E.4. Dynamic Heads

Figure 10 shows that early-layer heads tend to be more static, whereas heads after layer 2 exhibit increasingly dynamic behaviour, changing based on the input image. Specifically, we see the most dynamic heads in layer 14, aligning with our other findings.

### E.5. Attention Masking

We implement a masking strategy via the `attention_mask` parameter in `LLaVA-1.5`. This intervention is applied consistently across all heads and layers to restrict attention based on the spatial relationship described in the query.

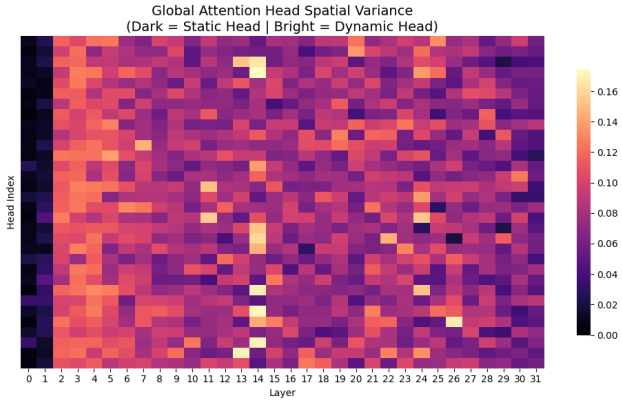
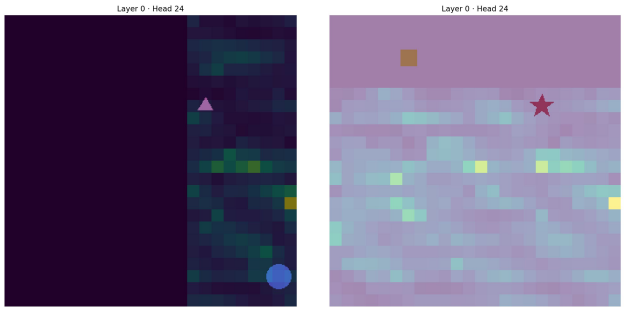



*Figure 10.* Spatial variance of attention center of mass (CoM) from the relational phrase token to image tokens, across heads and layers. Low values indicate static heads with consistent CoM across inputs, while higher values indicate dynamic heads whose attention shifts depending on the input image.




(a) Question: *Is the blue circle right of the pink triangle? Background: black*

(b) Question: *Is the yellow square below the red star? Background: white*

*Figure 11.* Visualization of the spatial attention masking strategy. The mask is applied to the region opposite the queried relation relative to the subject (e.g., masking the left side for a "*right of*" query). Arbitrary layers and heads are shown, the restriction is consistent across all layers and heads.

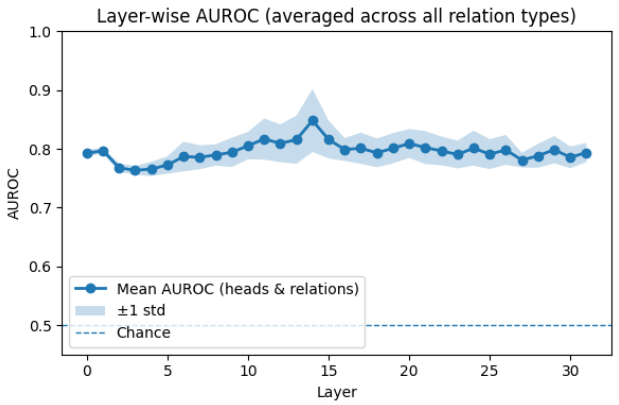



*Figure 12.* AUROC values for each attention head across layers, quantifying the separability between correct and incorrect predictions. Higher values indicate a stronger alignment between the attention center of mass and model correctness

*Table 6.* Accuracy and True Negative Rate (TNR) comparison between unmasked and always-masked conditions. All values are reported as percentages. TNR is defined as $\mathrm{TN}/(\mathrm{TN}+\mathrm{FP})$. Accuracy gains under masking are primarily driven by increased true negative predictions, as true positive rates remain consistently high across levels.

| **Level** | **Acc. (%)** | | **TNR (%)** | |
|---|---|---|---|---|
| | *Unmasked* | *Masked* | *Unmasked* | *Masked* |
| Level 0 | 51.3 | 58.8 ↑ +7.5 | 10.0 | 23.8 ↑ +13.8 |
| Level 1 | 51.3 | 63.8 ↑ +12.5 | 10.0 | 35.0 ↑ +25.0 |
| Level 2 | 54.0 | 60.5 ↑ +6.5 | 8.8 | 22.3 ↑ +13.5 |
| Level 3 | 51.9 | 53.2 ↑ +1.3 | 3.9 | 7.0 ↑ +3.1 |
| Level 4 | 50.6 | 52.6 ↑ +2.0 | 1.2 | 5.5 ↑ +4.3 |

### E.6. Object Grounding Signals

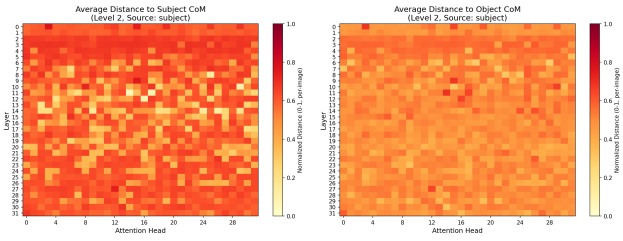



*Figure 13.* Average CoM distance heatmaps when the source token is the subject. Many heads show lower distance to the subject CoM, indicating stronger and more consistent attention toward the referenced subject compared to the object.

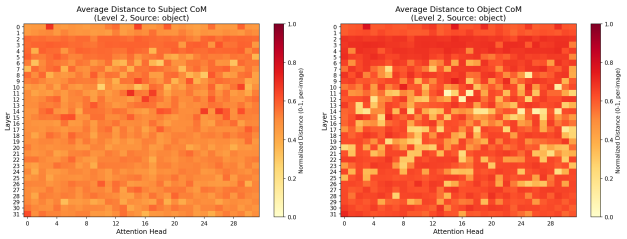



*Figure 14.* Average CoM distance heatmaps when the source token is the object. Heads now exhibit lower distance to the object CoM, showing that attention shifts toward the referenced object once it is mentioned in the prompt.

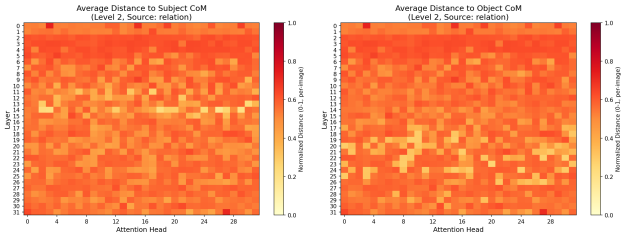



*Figure 15.* Average CoM distance heatmaps when the source token is the relation term. The pattern shows a weak cross-layer shift in which a subset of heads moves attention from the subject toward the object.

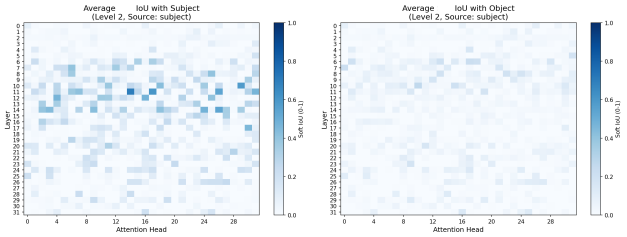



*Figure 16.* Average IoU heatmaps when the source token is the subject. IoU remains low overall but shows a weak increase for a small set of heads, confirming the mild subject-aligned grounding observed in the CoM distance maps.

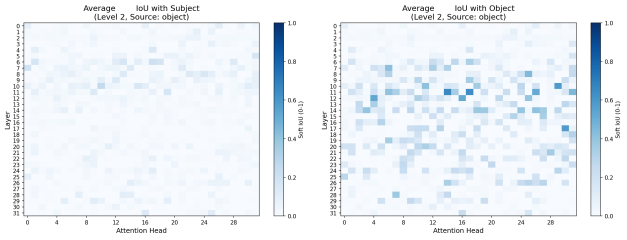



*Figure 17.* Average IoU heatmaps when the source token is the object. A slightly stronger but still weak increase in IoU appears for several heads, consistent with the object-aligned grounding seen in the corresponding CoM distance plots.