# THE DUAL MECHANISMS OF SPATIAL REASONING IN VISION–LANGUAGE MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

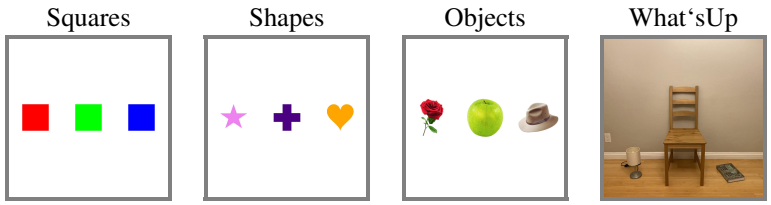
Many multimodal tasks, such as image captioning and visual question answering, require vision–language models (VLMs) to associate objects with their properties and spatial relations. Yet it remains unclear where and how such associations are computed within VLMs. In this work, we show that VLMs rely on two concurrent mechanisms to represent such associations. In the language model backbone, intermediate layers represent content-independent spatial relations on top of visual tokens corresponding to objects. However, this mechanism plays only a secondary role in shaping model predictions. Instead, the dominant source of spatial information originates in the vision encoder, whose representations encode the layout of objects and are directly exploited by the language model backbone. Notably, this spatial signal is distributed globally across visual tokens, extending beyond object regions into surrounding background areas. We show that enhancing these vision-derived spatial representations globally across all image tokens improves spatial reasoning performance on naturalistic images. Together, our results clarify how spatial association is computed within VLMs and highlight the central role of vision encoders in enabling spatial reasoning.

## 1 INTRODUCTION

Spatial reasoning, the ability to associate objects with their properties and their relations to other elements, is a fundamental component of multimodal understanding Cheng et al. (2024); Chen et al. (2021); Zheng et al. (2025); Chen et al. (2024b); Goetting et al. (2024); Wen et al. (2025). Tasks such as image captioning, visual question answering, and visual navigation require vision–language models (VLMs) to bind objects to spatial relations and reason about their relative arrangement within the scene. Although recent advances have made VLMs highly capable across many tasks, spatial reasoning remains a significant challenge Campbell et al. (2024); Chen et al. (2025); Wen et al. (2024); Kamath et al. (2023b). In this work, we investigate the internal mechanisms underlying spatial reasoning in VLMs and show how understanding these mechanisms enables targeted interventions to improve reasoning performance.

In language models (LMs), recent studies have shown that associating entities with their attributes (often referred to as variable binding) is done by forming symbolic representations for each entity in context. This representation encodes content-independent information about the order of entities in context, and allows the model to distinguish between entities and retrieve entity-specific information when needed Dai et al. (2024); Prakash et al. (2024; 2025). Follow-up studies suggest that VLMs similarly rely on ordering information when reasoning about objects in images Assouel et al. (2025); Kang et al. (2026). However, while these findings establish the use of ordering-based representations during multimodal reasoning, they do not reveal their origin. In particular, it remains unclear whether such representations are constructed within the LM backbone, inherited from the vision encoder, or emerge from interactions between the two. Identifying the source of these representations is essential for understanding spatial reasoning failures in VLMs and for developing principled methods to diagnose and correct them.

We show that the ordering-based representations underlying spatial reasoning in VLMs arise from two concurrent sources: The vision encoder represents the global layout of objects in the image, encoding ordering information that is directly projected into the embedding space of the LM backbone. Then, the LM backbone can further augment these representations by forming ordering information over




*"The <property> of the <object type> to the <direction> of <reference object> is _____"*

Figure 1: **Experimental Settings and Spatial Reasoning Task Definition:** We study four distinct image datasets to analyze the internal mechanisms of spatial reasoning in VLMs – three synthetic settings (Squares, Shapes, Objects) and one naturalistic setting (What'sUp). The model is queried to identify a specific property (e.g., color) of a target object based on its spatial relation (e.g., "to the left" or "to the right") relative to a central reference object.

object-associated visual tokens. While the vision encoder provides the primary source of ordering information, the LM-side mechanism plays a secondary role, augmenting spatial ordering when that information in the vision embeddings is degraded or completely removed.

Notably, the ordering information provided by the vision encoder is distributed globally across visual tokens, extending beyond object regions into surrounding background areas. This is in contrast to the LM backbone, where information is formed locally over object-related tokens, similar to the process found in language processing Prakash et al. (2025); Assouel et al. (2025).

We establish these results primarily through a series of controlled interchange intervention experiments Vig et al. (2020); Meng et al. (2022); Geiger et al. (2022) using carefully constructed counterfactual samples. We first examine three synthetic datasets, which allow us to isolate and analyze spatial mechanisms under controlled conditions (Fig. 1). We also show our finding generalizes to the What'sUp dataset Kamath et al. (2023b), a real-world benchmark designed to assess the spatial reasoning capabilities of state-of-the-art models. We validate our findings across two transformer-based VLMs, `Qwen2-VL-7B-Instruct` and `Gemma-3-4b-it`, showing that both rely on similar mechanisms to solve spatial reasoning tasks.

Finally, leveraging this mechanistic understanding, we introduce a simple global intervention on vision embeddings that amplifies the ordering information across all image tokens. On the What'sUp dataset, this intervention corrects more than $50\%$ of previously incorrect predictions in `Gemma-3-4b-it` and more than $30\%$ of incorrect predictions in `Qwen2-VL-7B-Instruct`.

Overall, our results elucidate the mechanisms underlying spatial association in VLMs, connect binding mechanisms identified in LMs to multimodal reasoning, and demonstrate the central role of vision encoders in supporting spatial reasoning in VLMs.

## 2 Related Works

**Benchmarking Spatial Reasoning** Many benchmarks have been proposed to evaluate spatial reasoning in VLMs, including synthetic datasets designed to test compositional relations and naturalistic benchmarks derived from real images and human annotations (Kamath et al., 2023a; Chen et al., 2024a; Gholami et al., 2025; Ma et al., 2025). These benchmarks have revealed that spatial reasoning remains challenging for current VLMs, particularly as relational complexity increases or when reasoning must generalize beyond object-centric cues.

**Understanding Inner Working of VLMs** A growing body of work has investigated the internal workings of VLMs, examining cross-modal attention patterns, token alignment, representational geometry, and information flow between vision encoders and LM backbones (Nikankin et al., 2025; Qin et al., 2025; Assouel et al., 2025; Jiang et al., 2025; Neo et al., 2025; Jiang et al., 2024; Kaduri et al., 2025; Rott Shaham et al., 2024; Cohen et al., 2026). These studies have provided valuable insights into how visual and textual information is integrated and how multimodal representations emerge. Our work builds on this line of research by focusing specifically on the mechanisms underlying spatial association and reasoning.

**Symbolic Representation in LMs & VLMs** Several studies have shown that LMs implement variable binding by forming content-independent ordering identifiers, over important tokens (Gur-Arieh et al., 2025; Prakash et al., 2025; Dai et al., 2024; Prakash et al., 2024; Feng & Steinhardt, 2023; Davies et al., 2023). These mechanisms underlying variable binding are fundamental to many in-context reasoning tasks and have been presented as evidence of symbol-like processing in neural networks. Recent work has extended this line of inquiry to VLMs, investigating similar symbolic representations Kang et al. (2026); Assouel et al. (2025); Saravanan et al. (2025); Hasani et al. (2025). Our findings relate to this body of work while revealing an important distinction in the multimodal setting. We show that although VLMs form symbolic representations that encode ordering information within the LM backbone, they only act as a supporting mechanism. Instead, symbolic representations originating from the vision encoder play a dominant role.

# 3 Experimental Setup

## 3.1 Data

We study a spatial reasoning task in which the model must identify a property (e.g., color or shape) of an object located at a specific spatial relation to a reference object. For example, for the *Squares* image shown in Fig. 1, the query may be "*The color of the square to the left of the green square is*", for which the correct answer is "*red*". For clarity, we present experimental setups and results for the *Squares* setting in the main text, and report results for the remaining settings in the appendix.

**Synthetic Settings** We consider three synthetic settings—*Squares*, *Shapes*, and *Objects*. In all settings, three equal-sized items are placed at fixed, equal distances and arranged either horizontally or vertically. The settings differ only in visual content: *Squares* uses colored squares, *Shapes* uses colored geometric shapes, and *Objects* uses real-world object images. All images are generated programmatically from predefined collections of colors, shapes, and objects (see full list at App. A), enabling precise control over attributes and spatial configuration. For each image, we query the VLM to predict an attribute (color, shape, or object name) of a target entity (square, shape, or object) based on its relative position to a reference entity. For our intervention experiments (see Sec. 5), we sample 50 pairs of clean-counterfactual images from each setting.

**Naturalistic Setting** In addition to the synthetic settings, we study real images from the What'sUp dataset Kamath et al. (2023b). Specifically, we use the control subset of the dataset, which consists of images containing pairs of objects: a central reference object (e.g., a chair) and a secondary object positioned adjacent to it on either side. To adapt this dataset to our intervention experiments (Sec. 5), which require images containing three objects, we construct composite scenes by merging pairs of What'sUp images into a single image with a shared reference object and two surrounding objects, resulting in 1074 images. We note that this control subset supports only horizontal spatial relations, limiting this setting to horizontal spatial reasoning.

## 3.2 Models

We study the mechanisms underlying spatial reasoning in two transformer-based VLMs: *Qwen2-VL-7B-Instruct*, *Gemma-3-4b-it*. These models differ in scale and training data, allowing us to assess whether the identified mechanisms generalize across diverse VLM designs. The behavioral performance of the models on the spatial reasoning task is summarized in Table 1. As shown, both models achieve near-perfect performance on the synthetic settings and strong performance on the What'sUp dataset, enabling reliable causal analysis.

Table 1: Model's accuracy on spatial reasoning tasks.

| | Squares | Shapes | Objects | What'sUp |
|---|---|---|---|---|
| Qwen2-VL-7B-Instruct | 1.00 | 1.00 | 1.00 | 0.89 |
| Gemma-3-4b-it | 0.98 | 0.99 | 0.99 | 0.91 |

## 3.3 Methods

**Representation Probing** Probing methods are widely used to analyze what information is encoded in the internal representations of neural networks Belinkov (2022). In this framework, a probe is typically a lightweight classifier trained to predict a target property $y$ from an internal representation $\mathbf{h} \in \mathbb{R}^d$ extracted from the model. In its simplest form, a linear probe takes the form $\hat{y} = \mathbf{W}\mathbf{h} + \mathbf{b}$,

where $\mathbf{W}$ and $\mathbf{b}$ are learned parameters. Successful probing that generalizes to test cases indicates that the target information is linearly decodable from $\mathbf{h}$, suggesting that it is present in the model's representation. However, probe performance alone does not establish that the probed representation plays a causal role in the model's behavior Hewitt & Manning (2019). As a result, probing is best interpreted as a diagnostic tool for localizing candidate representations, and is often combined with causal interventions to distinguish correlational encoding from functional relevance.

**Causal Mediation Analysis** Interchange intervention is a technique for testing causal relationships between a model's internal representations and its behavior Vig et al. (2020); Meng et al. (2022). Let $f(\cdot)$ denote the model, and let $\mathbf{h}_\ell(x)$ denote the internal representation at component or layer $\ell$ produced during a forward pass on input $x$. Given an *original* input $x$ and a corresponding *counterfactual* input $x'$, an interchange intervention replaces $\mathbf{h}_\ell(x)$ with $\mathbf{h}_\ell(x')$ while keeping the remainder of the computation unchanged. This yields an intervened output $\hat{y}_{\text{int}} = f\big(x;\, \mathbf{h}_\ell \leftarrow \mathbf{h}_\ell(x')\big)$, a procedure commonly referred to as *activation patching*.

If the intervened output $\hat{y}_{\text{int}}$ matches the target outcome associated with the counterfactual input, this provides evidence that the intervened representation $\mathbf{h}_\ell$ plays a causal role in the model's computation. We quantify this effect using *interchange intervention accuracy* (IIA), defined as the fraction of examples for which the intervened output matches the counterfactual target Geiger et al. (2022). By carefully constructing original–counterfactual input pairs, interchange interventions can be used not only to identify which components are causally involved in a task, but also to characterize the specific information or transformations those components mediate.

# 4 Preliminaries

## 4.1 Ordering Information in LMs

Variable binding is the process of associating attributes with their respective entities. For instance, given a red, green, and blue square, the process of assigning the color feature to each square involves variable binding. Owing to its fundamental role in reasoning, recent works have investigated the underlying mechanisms that enable it in LMs (Prakash et al., 2025; Dai et al., 2024). A common finding among these works is the reliance of LMs on content-independent ordering representations to bind entities to their attributes. These representations encode the order of entities independently of their content and align this ordering with a corresponding order over attributes in the context, enabling the model to retrieve entity-specific information when queried. For instance, given a prompt such as "*Box A contains an apple, box B contains a banana, and box C contains cherries. What is the color of the fruit in box B?*", the model uses ordering information corresponding to the second entity (*Box B*) to retrieve the attributes of the corresponding second fruit (*banana*). Prior analyses suggest that such ordering information is formed in intermediate layers on top of entity-related token positions and is subsequently transferred to the final token position when answering a query, where it is used to retrieve the relevant property (e.g., the color "*yellow*" of the banana). In this work, we find that VLMs similarly form ordering information over visual tokens corresponding to objects in an image and use this information to reason about the spatial locations of objects and their attributes. However, as we show in the following sections, this LM-style ordering mechanism plays only a secondary role in multimodal spatial reasoning.

## 4.2 Ordering Information in VLMs

In VLMs, spatial variable binding is a fundamental capability central to many spatial reasoning tasks. Recently, (Assouel et al., 2025) demonstrates that when VLMs are queried about objects within images, they retrieve symbolic representations that encode their ordering information. While this work establishes the existence of such ordering representations for visual variable binding tasks, it does not characterize the source of it. Specifically, the work mainly shows that for variable binding, the model uses this symbolic representation at the last token position, and that it comes from visual tokens; however it remains unclear *where and how* VLMs create ordering information. In particular, it remains unclear whether ordering representations are constructed within the LM backbone, inherited directly from the vision encoder, or emerge through interactions between the two components. In this work, we show that the source of this is two-fold: the vision encoder encodes positional information

reflecting the spatial layout of objects, while the LM backbone further enhances this information. Understanding the origin of these representations enables a simple and principled intervention that improves spatial reasoning performance.

# 5 EXPERIMENTS

## 5.1 VLMS USE ORDERING INFORMATION FOR SPATIAL REASONING

We begin by establishing the presence of ordering representations that the model uses at the final token position, across a range of tasks, as described in Sec. 3.1, by conducting interchange intervention experiments at the final token position. Specifically, we separately patch the residual stream vector of each layer at the last token position with that from a counterfactual sample (see Fig. 2) featuring different square colors in the image and a different directional query.

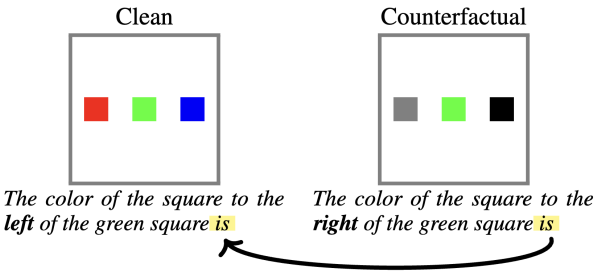



Figure 2: **Experimental Setup for Last-Token Position Patching.** We perform an interchange intervention experiment to test whether ordering information is used for spatial reasoning. The clean example contains an image with **Red**–**Green**–**Blue** squares, and the prompt queries the square to the left of the green square (answer: **Red**). The counterfactual example contains an image with **Gray**–**Green**–**Black** squares and queries the square to the right of the green square (answer: **Black**). We patch the residual stream vector at the final token ("is") from the counterfactual run into the clean run and examine how the model's output changes. If ordering information is transferred, the model should output the color of the object in the clean image with the same ordinal position as the correct object in the counterfactual image (**Blue**); if attribute information is transferred, the output should instead reflect the counterfactual color (**Black**).

If the model uses the ordinal position of the queried square to retrieve its color, we expect this ordering information to be transferred from the counterfactual run to the clean run. Concretely, in the counterfactual run corresponding to Fig. 2, if the model encodes a representation indicating that the third object is the correct one, patching this representation into the clean run should cause the model to select the object with the same ordinal position in the clean image. As a result, the output should change from the color of the originally selected object (the leftmost square; **Red**) to the color of the square in the clean image corresponding to the third position (**Blue**), which matches the correct position in the counterfactual run.

In contrast, if the model directly transfers attribute information rather than ordering information, patching the final-token representation from the counterfactual run should cause the output to reflect the color of the queried square in the counterfactual image itself (e.g., **Black** in Fig. 2), independent of the object selected in the clean image. Observing this behavior would indicate that color information (rather than ordinal position) is being transferred by the intervention. We quantify the effect of this intervention using interchange intervention accuracy (IIA; Section 3.3).

Figure 3 shows that the final output initially corresponds to the color of the correct square of the clean input between layers **0-19**, indicating that the intervention on these layers has no effect. It then switches to selecting the color of the square in the clean image corresponding to the ordinal position of the correct square in the counterfactual image **20-22**, indicating that ordering information is being transferred. At layers **23–27**, the model switches to the color of the counterfactual correct square, indicating that correct color attribute information is being transferred directly.

This pattern suggests that the model first computes an ordering representation of the correct square and then, in later layers, retrieves its corresponding color information. For clarity, we present results for the *Squares* setting for the Qwen model. Consistent behavior is observed across all settings (App. B.1). Results are averaged over 50 clean–counterfactual pairs per setting. For convenience, the plots use the color scheme from the example in Fig. 2; however, the specific colors, shapes, and object categories vary across data points, as well as the queried directions.

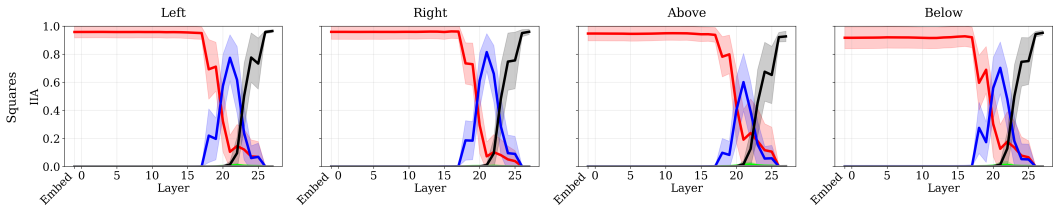



Figure 3: **Last Token Position Patching Results**: The final output of the clean run remains unchanged up to layer 19. From layer 20 through layer 22, the model produces the expected prediction, suggesting that these layers encode the order of the target square. Beyond layer 22, the model instead predicts the counterfactual output, indicating that layers after 22 encode the color of the target square.

### 5.2 Is Vision Encoder the Source of Ordering Representations?

While the previous subsection established the presence of ordering representations, it did not fully characterize them. In particular, a key open question concerns where these ordering representations are generated within VLMs: are they produced in the LM backbone, the vision encoder, or through interactions between the two components? Identifying the source of ordering representations is important not only for understanding how they arise, but also for enabling targeted mechanistic interventions to improve the spatial reasoning capabilities of VLMs.

#### 5.2.1 Probing Visual Embeddings

We begin by asking whether ordering representations originate in the vision encoder. To address this question, we train linear probes on visual token embeddings, extracted immediately after projection from the vision encoder, to predict the ordinal position of each object in the image. Successful probe performance indicates that ordering information is linearly decodable from the visual embedding.

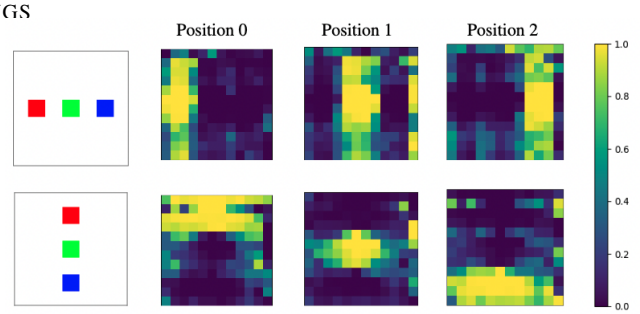



Figure 4: **Position information in visual embeddings extends beyond object regions.** We train linear probes on top of object tokens of the visual embedding to predict their order in the image. At test time, these probes generalize to background tokens, producing strips-like pattern aligned with object position. This suggests that background tokens encode positional information about nearby objects. Please see App. B.5 for more complex arrangements.

For each image setting described in Section 3.1, we train two linear probes, one for horizontal and one for vertical configurations, to classify embeddings as one of three spatial positions. Training data is constructed using token embeddings extracted from 90 images; the embeddings that correspond to the three objects in an image serve as positive examples for their respective positions. In the *Squares*, *Shapes*, and *Objects* datasets, identifying the visual tokens associated with each object is straightforward due to their fixed spatial layouts. In contrast, objects in the *What'sUp* dataset vary in size and are not aligned to fixed positions. To handle this, we preprocess *What'sUp* images by extracting bounding boxes for each object and use their spatial coordinates to identify the corresponding visual tokens (see Fig. 33).

The resulting classifiers achieve nearly perfect accuracy on a test set of 30 images. In addition to classifying object-token representations, we also apply the probes to all other tokens in the image to assess whether ordering information representations are present in them.

We find that the information about the object ordering is not confined to tokens corresponding to the objects themselves, but is instead distributed across multiple background tokens, as shown in Fig. 4. Notably, probe predictions exhibit a strip-like spatial pattern aligned with object position, indicating that positional information is distributed coherently across background tokens rather than localized to object regions. This behavior contrasts with prior work on LMs (Prakash et al., 2024; 2025; Dai et al., 2024), which shows that such representations are typically localized to a single token or a small set of adjacent tokens. In the following subsections, we demonstrate that the information contained in

these strips is causally relevant for producing the correct final output. In Sec. 5.4, we further leverage this observation by intervening along probe directions to improve model performance.

#### 5.2.2 Causal Intervention on Vision Tokens

Although the probing results demonstrate the presence of ordering information in visual representations, they do not by themselves establish a causal relationship with the model's final output. We therefore perform interchange intervention experiments to test whether systematically manipulating these representations leads to corresponding changes in the model's predictions.

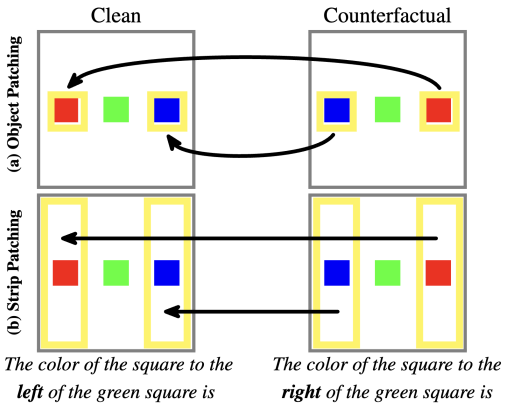



Figure 5: **Experimental design for causal intervention on visual token embeddings**: We perform two interchange interventions on the visual token embeddings: (a) patching only the tokens corresponding to the squares, and (b) patching the tokens corresponding to both the squares and their corresponding background strips as observed in probing experiments. In both cases, visual tokens corresponding to the left and right squares (and, in (b), their associated strips) are symmetrically swapped (i.e. lef↔right) from counterfactual to clean runs. If the patched visual tokens encode ordering information, the model's final output in the clean run should switch to the square on the opposite side of the queried reference object, since the patched representations now carry the ordering information corresponding to the queried direction.

As illustrated in Fig. 5, we intervene on visual token embeddings by swapping the left and right strips of the clean image with those from a counterfactual image. Specifically, we replace the embeddings of the left strip in the clean image with the right-strip embeddings from the counterfactual image, and vice versa. The intervention is designed such that both patched regions contain squares of the same color, ensuring that color values are preserved and that only ordering information is altered. In addition to intervening on the visual token embeddings, we also separately patch the residual stream vectors of each subsequent layer of the LM backbone.

If ordering information encoded by the vision encoder is causally relevant, this intervention should cause the model's final output on the clean image to switch from the left square (red) to the right square (blue), reflecting the swapped ordering information. As shown in Fig. 7, we observe the expected change in the final output when either the visual token embeddings or the residual stream vectors up to layer 24 are patched. In contrast, patching only the square-localized visual tokens, without the surrounding strip, does not reliably induce this change (Fig. 6). This result indicates that ordering information encoded by the vision encoder is not localized to object tokens alone, but is instead distributed across strip-aligned visual tokens.

We observe similar behavior in other settings, including Shapes, Objects, and What'sUp, which also generalizes to `Gemma-3-4b-it`, as described in App. B.3. All results are averaged over 50 samples. Moreover, the fact that the original correct square logit overtakes the intervened one around layer 23 indicates that by that layer, all the required information, including both ordering and color value information, is transferred to the last token, as discussed in Sec. 4.2.

### 5.3 LM Backbone Enhances Ordering Information

In the previous sections, we showed that visual token embeddings produced by the vision encoder and projector module already encode object ordering information. We now ask what role, if any, the LM backbone plays in spatial reasoning. In particular, does the LM backbone merely consume ordering information provided by the vision components, or can it generate such information when needed?

#### 5.3.1 Removing Ordering Information from Vision Embeddings

To isolate the contribution of the LM backbone, we first ablate ordering information from the visual token embeddings using the interventions illustrated in Fig. 8. Specifically, we apply two

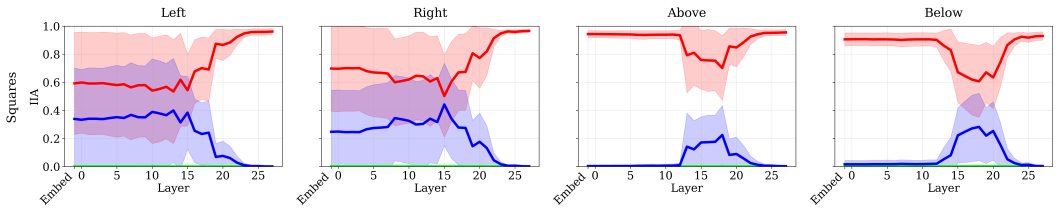



Figure 6: **Object patching**. Interchange interventions applied only to visual tokens corresponding to the squares. Patching square tokens at any layer does not reliably change the model's final output relative to the clean run, indicating that square-localized tokens alone do not carry sufficient ordering information to determine the model's prediction.

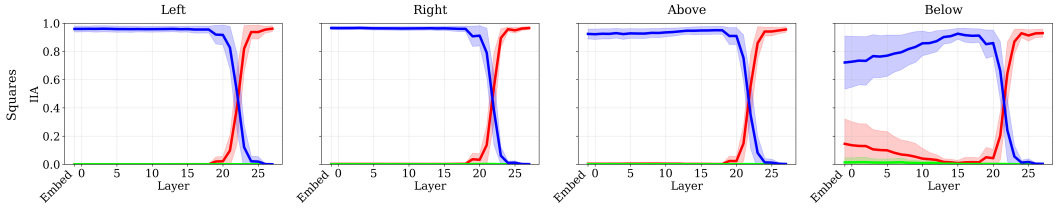



Figure 7: **Strip patching**. Interchange interventions applied to both square tokens and their background strips immediately switch the model's output to the incorrect color in the clean image, consistent with the transferred ordering information. This demonstrates that ordering information is distributed across strip-aligned visual tokens and is causally relevant for spatial reasoning.

modifications: (1) we replace the embedding of each square with an embedding of the same-colored square taken from an image in which that square appears alone in the middle of the image, thereby preserving color information while removing relative ordering; and (2) we replace all background-token embeddings with those from an empty image, eliminating ordering information encoded in background regions. Together, these interventions effectively remove ordering information originating from the vision encoder, while preserving object colors. We do not apply this intervention to the What'sUp dataset, as constructing a precise intervention is challenging for natural images.

As shown in Table 6, removing ordering information from the vision embeddings leads to a substantial drop in performance across datasets, confirming the central role of vision-derived ordering information. However, performance remains above chance ($33.3\%$), suggesting that the LM backbone can compensate for the missing ordering signals.

### 5.3.2 Do LMs Backup Ordering Information?

To test whether the LM backbone generates its own ordering information, we perform an interchange intervention experiment analogous to that described in Section 5.2.2, but under the ablated-vision setting described above (e.g. the LM gets images with no ordering information from the vision encoder). In contrast to earlier experiments, here we patch only the square-associated visual tokens, rather than the entire strips.

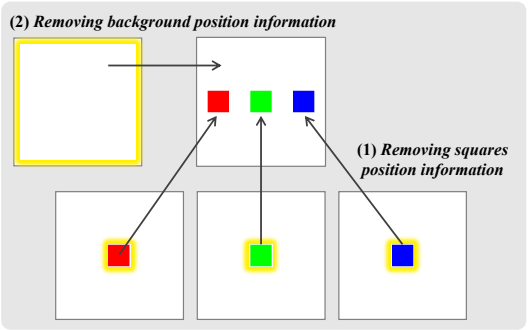



Figure 8: **Ordering information is ablated from vision embeddings** by replacing square embeddings with same-colored isolated middle squares and background embeddings with those from an empty image, preserving color while removing position cues.

If the LM backbone generates ordering information independently, then patching square tokens from a counterfactual run should transfer this order information and induce a change in the final output. If not, the model's prediction should remain unchanged. The results in Fig. 9 show a flip in model prediction according to order information at layers 13-17. This indicates that, when vision-derived ordering information is removed, the LM backbone forms its own ordering

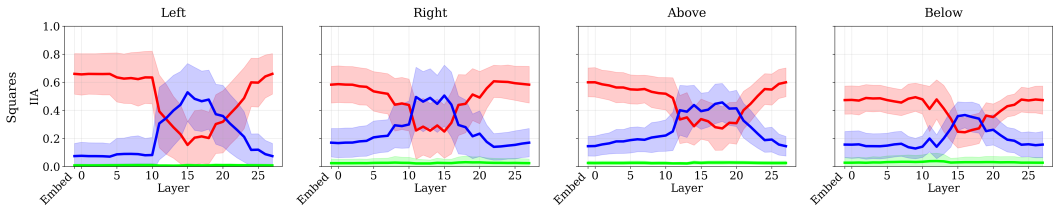



Figure 9: **Patching results after removing ordering information from vision encoder.** The model's output switches only in intermediate layers within the range of 11-20, indicating that the LM backbone generates ordering information even when vision-derived ordering representations are absent.

representations in these middle layers, and that patching these representations switches the prediction according to the patched order, causally affecting the final prediction.

Together, these results indicate that VLMs rely on two sources of ordering information: a primary signal originating from the vision encoder and projector module, and a secondary signal generated within the LM backbone. While the vision-derived signal dominates when present, the LM backbone can partially reconstruct ordering information and act as a backup mechanism. Supporting evidence for this interpretation is provided in Figs. 7, 18, 17, where we show that vertically arranged objects strengthen ordering signals in the same intermediate LM layers, indicating the LM backbone enhances the ordering information provided by the vision embedding in these cases.

### 5.4 Correcting Incorrect Predictions

We leverage the mechanistic insights from the previous sections to improve spatial reasoning performance on the What'sUp dataset. Our analysis shows that ordering information is essential for spatial reasoning and that the vision encoder and projector module provide the primary source of this information. We therefore hypothesize that amplifying vision-derived ordering representations can improve a VLM's spatial reasoning ability. To test this hypothesis, we enhance the ordering information in the vision embeddings by amplifying probe directions identified for the What'sUp dataset in Section 5.2. These probe directions are trained to capture dimensions encoding ordering information; amplifying them should therefore strengthen the ordering signal produced by the vision encoder and projector module.

Table 2: Amplifying ordering representation enhances spatial reasoning.

| | Gemma | | Qwen | |
|---|---|---|---|---|
| **Intervention** | **Acc.** | **% Corr. Fail.** | **Acc.** | **% Corr. Fail.** |
| None | 0.9 | – | 0.89 | – |
| Random amp. | 0.92 | 15.8% | 0.9 | 8.8% |
| Ordering amp. | **0.95** | **50.7%** | **0.93** | **32.4%** |

Formally, for each visual token $t$, we modify its embedding: $\text{emb}_t \leftarrow \text{emb}_t + \alpha \cdot \text{probe}_i$, where $\alpha \in [1, 15]$ is an amplification coefficient and $\text{probe}_i$ is the probe direction corresponding to object $i$ in the image. We apply this intervention to all visual token embeddings, across all probe directions. As a baseline, we perform the same intervention using randomly sampled directions, allowing us to isolate the effect of amplifying ordering-specific representations. Table 2 shows that amplifying visual ordering representations corrects more than $50\%$ of previously incorrect predictions for `Gemma-3-4b-it` and more than $30\%$ for `Qwen2-VL-7B-Instruct`, outperforming random-direction amplification. This intervention improves overall accuracy by up to $5\%$, without any fine-tuning or access to ground-truth labels. Together, these results demonstrate that directly strengthening vision-derived ordering information can improve spatial reasoning in VLMs.

## 6 Conclusions

We show that spatial reasoning in VLMs relies on ordering representations arising from two complementary sources: a primary signal encoded by the vision encoder and a secondary mechanism formed within the LM backbone. By isolating, intervening on, and amplifying these representations, we demonstrate a simple intervention that corrects spatial reasoning failures. Together, our results highlight how mechanistic understanding can be translated into insights that can lead to improved model performance.

## IMPACT STATEMENT

This work underscores the critical role of the vision encoder in multimodal systems, challenging the prevailing focus on the language backbone. By demonstrating that the dominant symbolic representations for spatial reasoning originate in the visual components, we highlight that future advances in VLMs must prioritize the training of robust vision encoders to ensure accurate spatial layout encoding. Furthermore, we establish the practical utility of mechanistic interpretability by showing that understanding these internal mechanisms allows for targeted interventions, specifically, amplifying vision-derived spatial signals, that significantly improve performance on state-of-the-art benchmarks without the need for additional training or fine-tuning.

Methodologically, our discovery that spatial information is diffused across multiple visual tokens, extending even into background regions, signals a necessary paradigm shift for interpretability research. Standard analysis techniques that focus on single tokens are insufficient for capturing these distributed representations. As the field moves toward reasoning models that generate increasingly long sequences of tokens, developing interpretability methods capable of analyzing information distributed across multiple tokens will be essential for diagnosing and enhancing model behaviors.

## REFERENCES


Rim Assouel, Declan Campbell, Yoshua Bengio, and Taylor Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. *arXiv preprint arXiv:2506.15871*, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. *Advances in Neural Information Processing Systems*, 37:113436–113460, 2024.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14455–14465, June 2024a.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14455–14465, 2024b.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523, 2021.

Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025. URL https://arxiv.org/abs/2503.01773.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024.

Ido Cohen, Daniela Gottesman, Mor Geva, and Raja Giryes. Performance gap in entity knowledge extraction across modalities in vision language models, 2026. URL https://arxiv.org/abs/2412.14133.

Qin Dai, Benjamin Heinzerling, and Kentaro Inui. Representational analysis of binding in language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17468–17493, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.967. URL https://aclanthology.org/2024.emnlp-main.967/.

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. Discovering variable binding circuitry with desiderata. *arXiv preprint arXiv:2307.03637*, 2023.

Jiahai Feng and Jacob Steinhardt. How do language models bind entities in context? *arXiv preprint arXiv:2310.17191*, 2023.

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. Inducing causal structure for interpretable neural networks. In *International Conference on Machine Learning*, pp. 7324–7338. PMLR, 2022.

Mohsen Gholami, Ahmad Rezaei, Zhou Weimin, Sitong Mao, Shunbo Zhou, Yong Zhang, and Mohammad Akbari. Spatial reasoning with vision-language models in ego-centric multi-view scenes. *arXiv preprint arXiv:2509.06266*, 2025.

Dylan Goetting, Himanshu Gaurav Singh, and Antonio Loquercio. End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. *arXiv preprint arXiv:2411.05755*, 2024.

Yoav Gur-Arieh, Mor Geva, and Atticus Geiger. Mixing mechanisms: How language models retrieve bound entities in-context. *arXiv preprint arXiv:2510.06182*, 2025.

Hosein Hasani, Amirmohammad Izadi, Fatemeh Askari, Mobin Bagherian, Sadegh Mohammadian, Mohammad Izadi, and Mahdieh Soleymani Baghshah. Uncovering grounding ids: How external cues shape multimodal binding. *arXiv preprint arXiv:2509.24072*, 2025.

John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Nick Jiang, Anish Kachinthaya, Suzie Petryk, and Yossi Gandelsman. Interpreting and editing vision-language representations to mitigate hallucinations, 2024. URL https://arxiv.org/abs/2410.02762.

Nick Jiang, Amil Dravid, Alexei Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. *arXiv preprint arXiv:2506.08010*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, June 2025.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's" up" with vision-language models? investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023a.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning, 2023b. URL https://arxiv.org/abs/2310.19785.

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. *arXiv preprint arXiv:2601.12626*, 2026.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6924–6934, 2025.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36, 2022. arXiv:2202.05262.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models, 2025. URL https://arxiv.org/abs/2410.07149.

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, and Yonatan Belinkov. Same task, different circuits: Disentangling modality-specific mechanisms in vlms. *arXiv preprint arXiv:2506.09047*, 2025.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. *arXiv preprint arXiv:2402.14811*, 2024.

Nikhil Prakash, Natalie Shapira, Arnab Sen Sharma, Christoph Riedl, Yonatan Belinkov, Tamar Rott Shaham, David Bau, and Atticus Geiger. Language models use lookbacks to track beliefs, 2025. URL https://arxiv.org/abs/2505.14685.

Yulu Qin, Dheeraj Varghese, Adam Dahlgren Lindström, Lucia Donatelli, Kanishka Misra, and Najoung Kim. Vision-and-language training helps deploy taxonomic knowledge but does not fundamentally alter it. *arXiv preprint arXiv:2507.13328*, 2025.

Tamar Rott Shaham, Sarah Schwettmann, Franklin Wang, Achyuta Rajaram, Evan Hernandez, Jacob Andreas, and Antonio Torralba. A multimodal automated interpretability agent. In *Forty-first International Conference on Machine Learning*, 2024.

Darshana Saravanan, Makarand Tapaswi, and Vineet Gandhi. Investigating mechanisms for in-context vision language binding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4852–4856, 2025.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. *Advances in neural information processing systems*, 33:12388–12401, 2020.

Chuan Wen, Dinesh Jayaraman, and Yang Gao. Can transformers capture spatial relations between objects?, 2024. URL https://arxiv.org/abs/2403.00729.

Congcong Wen, Yisiyuan Huang, Hao Huang, Yanjia Huang, Shuaihang Yuan, Yu Hao, Hui Lin, Yu-Shen Liu, and Yi Fang. Zero-shot object navigation with vision-language models reasoning. In *International Conference on Pattern Recognition*, pp. 389–404. Springer, 2025.

Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, et al. Multimodal spatial reasoning in the large model era: A survey and benchmarks. *arXiv preprint arXiv:2510.25760*, 2025.

# Appendix

We bring here additional results and experimental details.

## A Data

### A.1 Models' Accuracy

| | Squares | | | | Shapes | | | | Objects | | | | What'sUp | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | R | A | B | L | R | A | B | L | R | A | B | L | R |
| Qwen2-VL-7B-Instruct | 1.00 | 1.00 | 1.00 | 0.99 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 | 0.92 | 0.86 |
| Gemma-3-4b-it | 1.00 | 1.00 | 0.93 | 0.97 | 1.00 | 1.00 | 0.97 | 0.97 | 1.00 | 1.00 | 0.95 | 1.00 | 0.91 | 0.90 |

Table 3: Full behavioral performance, according to queried directins (Left,Right,Above,bellow)

| | Squares | | | | Shapes | | | | Objects | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | L | R | A | B | L | R | A | B | L | R | A | B |
| Qwen2-VL-7B-Instruct | 0.85 | 0.60 | 0.43 | 0.50 | 0.76 | 0.65 | 0.53 | 0.52 | 0.47 | 0.43 | 0.31 | 0.50 |
| Gemma-3-4b-it | 0.49 | 0.47 | 0.62 | 0.59 | 0.53 | 0.47 | 0.70 | 0.58 | 0.44 | 0.56 | 0.46 | 0.76 |

Table 4: Behavioral performance with visual embedding intervention, according to queried directins (Left,Right,Above,bellow)

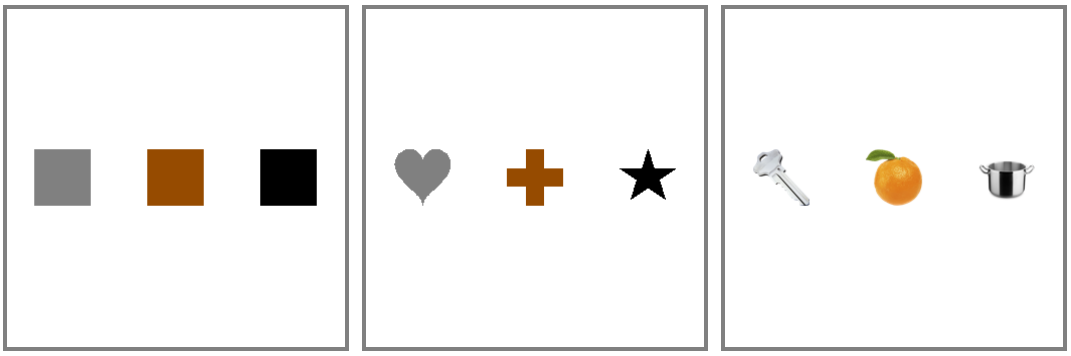
### A.2 Image Settings

Each of our settings consists of images constructed from the following possible objects:

**Squares:**
`red`, `green`, `blue`, `black`, `brown`, `gray`

**Shapes:**
`red square`, `green circle`, `blue triangle`, `black star`, `brown cross`, `gray heart`

**Objects:**
`orange`, `apple`, `pot`, `bell`, `hat`, `rose`, `key`, `bomb`



(a) Square setting (b) Shapes setting (c) Objects setting

Figure 10: Example images of the Square, Shapes, and Objects settings.

For each model, we report the following parameters in Table 5.

- **# Image tokens**: total number of image tokens in an image, as a square grid of `num tokens` × `num tokens` tokens.

| Model | # Image tokens | # Pixels / Token | # Tokens / Object | Object spacing | Strip width |
|---|---|---|---|---|---|
| Qwen2-VL-7B-Instruct | $12 \times 12$ | $28 \times 28$ | $2 \times 2$ | 2 | 4 |
| Gemma-3-4b-it | $16 \times 16$ | $56 \times 56$ | $3 \times 3$ | 3 | 3 |

Table 5: Image tokenization and spatial layout parameters for each model.

- **# Pixels / Token**: number of image pixels per token, as `num pixels` × `num pixels` pixels per token.
- **# Tokens / Object**: the number of tokens each object in the image occupies.
- **Object spacing**: the horizontal/vertical distance, in tokens, between adjacent objects in an image.
- **Strip width**: the strip width, in tokens, used for interventions such as those in Fig. 7.

### A.3 Prompts

Generation prompt tokens are added to the end of each prompt to signal the beginning of the assistant response.

**System prompt (all settings):**
`You are a helpful assistant. You respond in one token.`

**User prompt templates:**
Each `<direction>` $\in$ {`left`, `right`, `above`, `below`} is mapped to a preposition:

{left ↦ ``to the left of'', right ↦ ``to the right of'',
above ↦ ``above'', below ↦ ``below''}

**Squares:**
`The color of the square <preposition> the <middle color> square is`
**Shapes:**
`The color of the shape <preposition> the <middle color> is`
**Objects:**
`The object <preposition> the <middle object> is`
**What's Up (Objects):**
`The object on the floor <preposition> the <middle object> is`

## B Additional results

### B.1 Last Token Position Patching

In addition to the results described in Section 5.1, we report the results of the last-token interchange intervention experiment on other settings (Shapes, Objects, and What'sUP) and models (`Gemma-3-4b-it`). Results are shown in Fig. 11 and Fig. 12.

### B.2 Removing Ordering-Information from Visual Embeddings Impact on Behavioral Performance

We report the behavioral performance of the VLMs after removing the ordering information encoded by the vision encoder. We observe a significant drop, demonstrating its vital role in the spatial reasoning capabilities of VLMs.

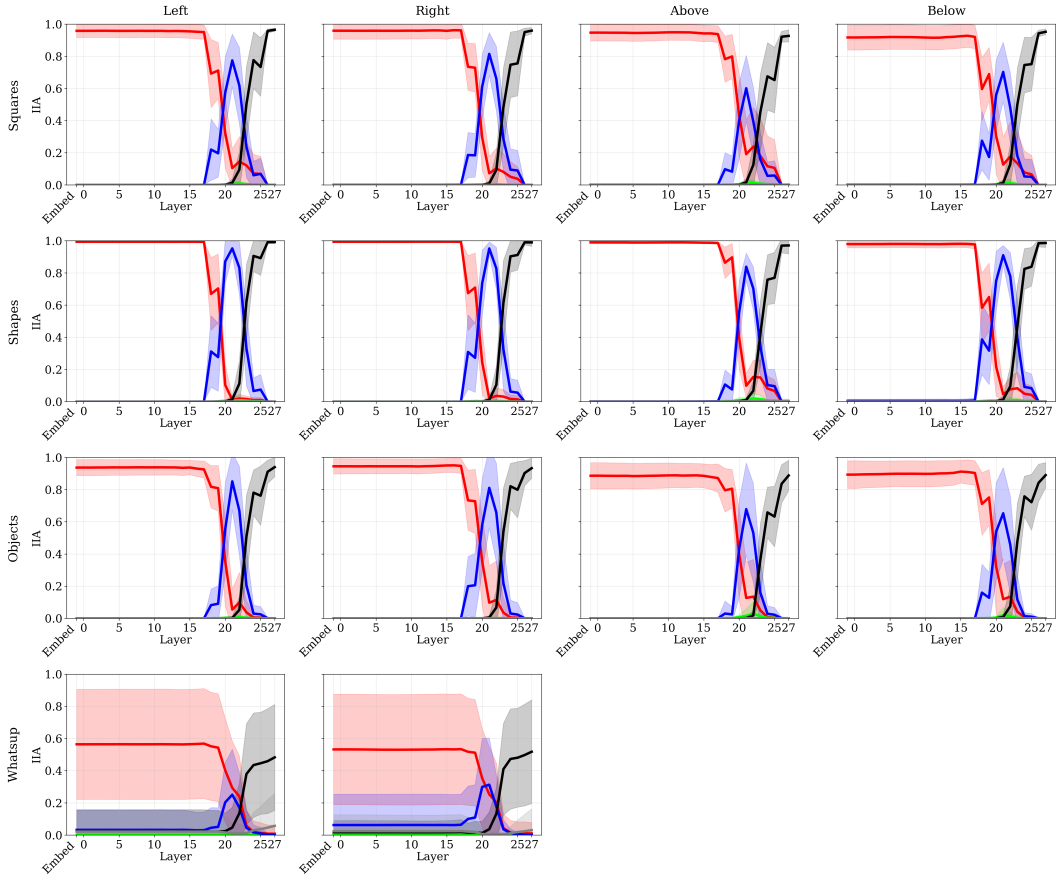



Figure 11: Final token interchange intervention experiments with the `Qwen2-VL-7B-Instruct` model.

Table 6: Behavioral performance after removing spatial information from vision embeddings.

| | Squares | Shapes | Objects |
|---|---|---|---|
| Qwen2-VL-7B-Instruct | 0.60 | 0.62 | 0.43 |
| Gemma-3-4b-it | 0.64 | 0.82 | 0.77 |

### B.3 Vision Token Patching

In this subsection, we present the results of the interchange intervention experiment described in Section 5.2.2. We report results for interventions that patch both the square and strip visual tokens for each model. The findings are shown in Fig. 14, Fig. 15, Fig. 13, and Fig. 16. In both cases, the results suggest that ordering information is not sufficiently encoded in the square visual token embeddings.

In Fig. 18 and Fig. 17, we present additional patching results after removing vision encoder-generated ordering information from visual embeddings using the procedure from Section 5.3.1. Output switching occurs in intermediate layers, corroborating our finding that the LM backbone produces its own ordering information.

Results are omitted when sufficient clean/counterfactual input pairs cannot be generated.

### B.4 Probing

One of the main findings of this work is that ordering information from the vision encoder is not concentrated in a small set of square tokens; instead, it is distributed across neighboring background tokens as well. In this subsection, we present probing results across multiple settings and models,

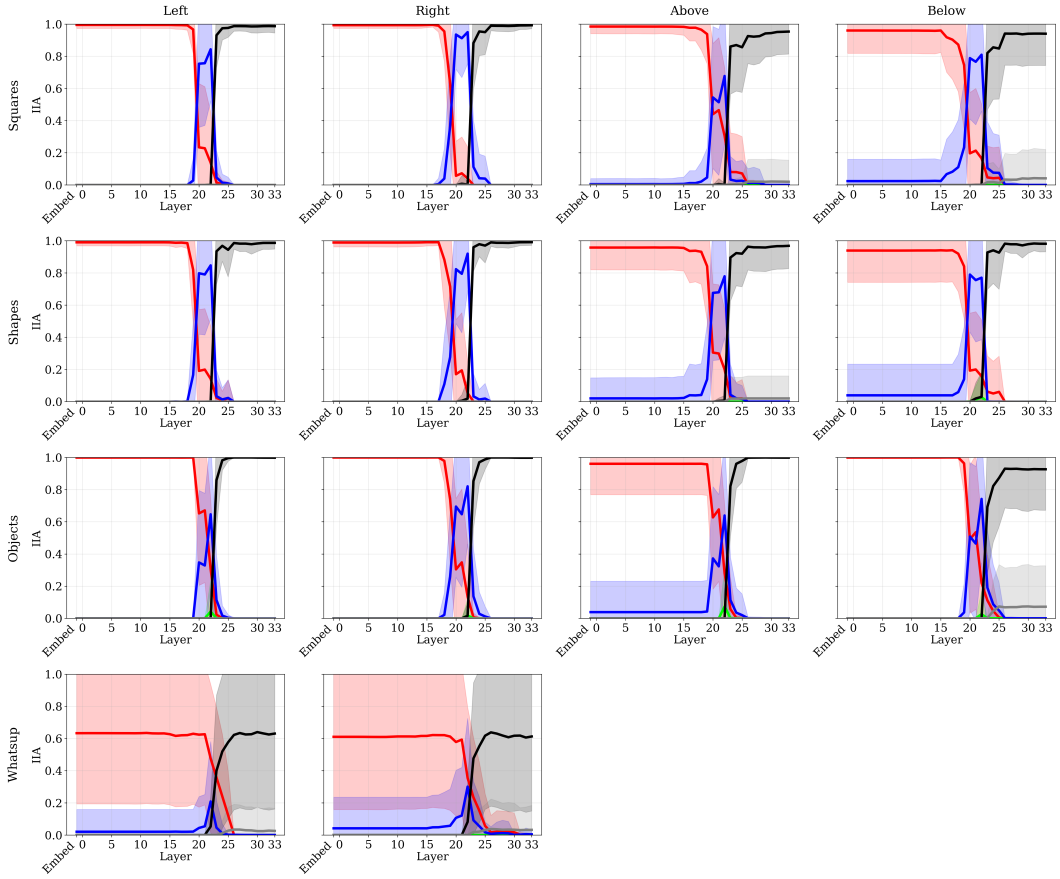



Figure 12: Final token interchange intervention experiments with the `Gemma-3-4b-it` model.

demonstrating the ubiquity of this phenomenon of diffused ordering information. Results are demonstrated in Fig. 19. Fig. 20, Fig. 21, Fig. 22, Fig. 23, and Fig. 24. We also show similar results with `Gemma-3-4b-it` in Fig. 25, Fig. 26, Fig. 27, Fig. 28, Fig. 29, and Fig. 30.

### B.5 Probing on grids

We extend the probing results from subsection B.4 to grid configurations consisting of both horizontal and vertical relationships between objects, with findings presented in Fig. 31 and Fig. 32. Probes are trained on $2 \times 2$ token objects for Qwen2-VL-7B-Instruct and $3 \times 3$ token objects for Gemma3-4b-it, with results exhibiting diffusion of positional information to surrounding tokens.

### B.6 Probing for What'sUp

Finally, we also show that the ordering information is diffused across background tokens in real-world images from What'sUp. Training data is extracted from the embeddings within the bounding boxes outlined in Fig. 33. Results are consistent with previous findings, as illustrated in Fig. 35.

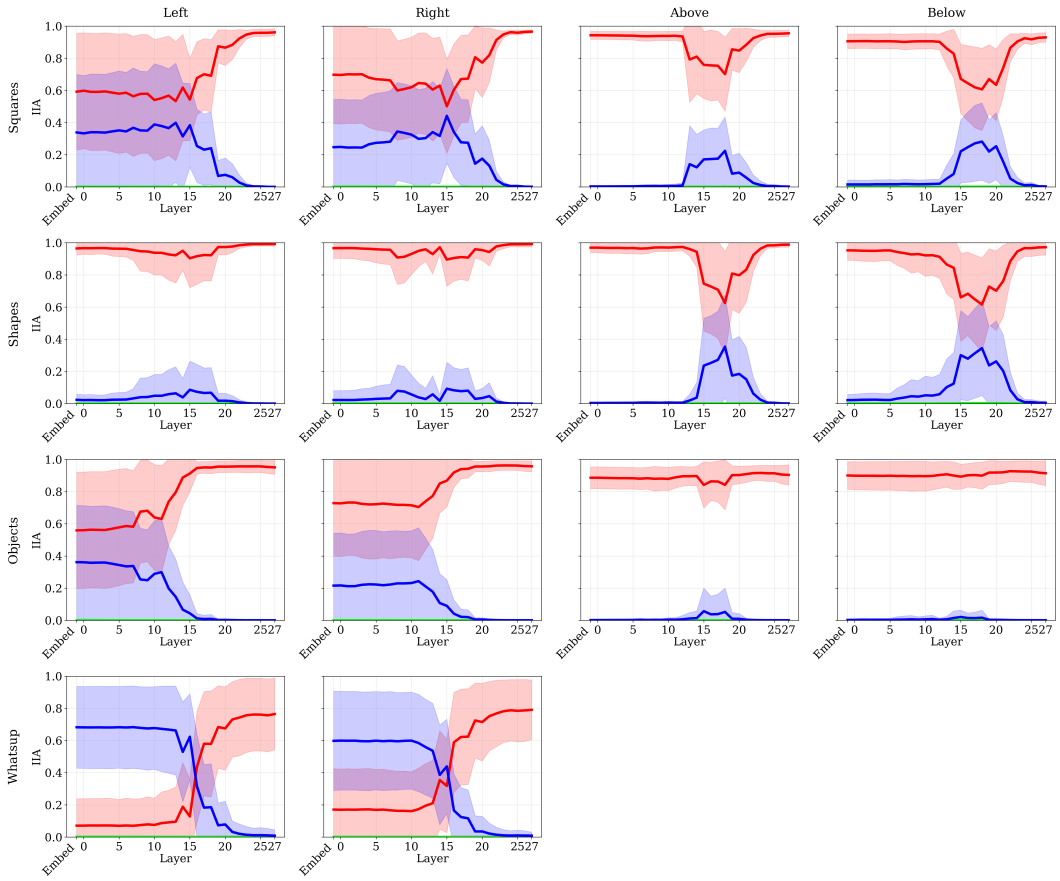



Figure 13: Visual token embeddings of square tokens interchange intervention experiment on `Qwen2-VL-7B-Instruct` model.

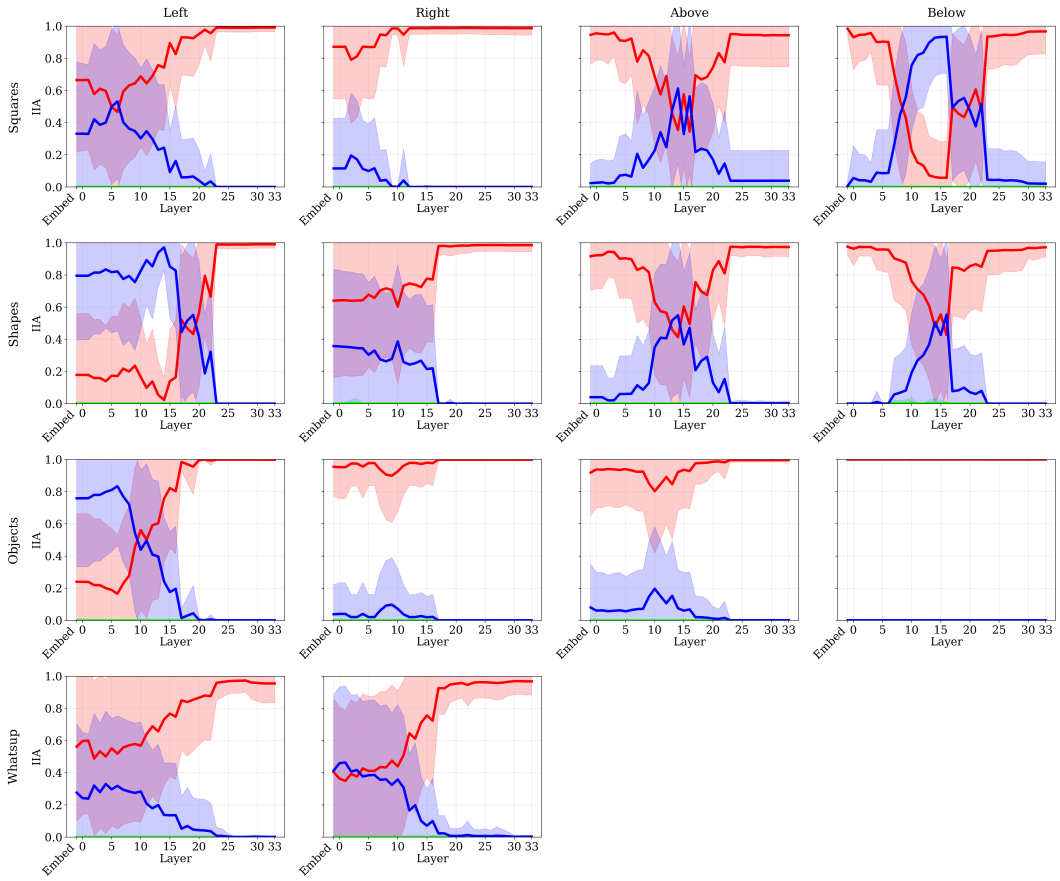



Figure 14: Visual token embeddings of square tokens interchange intervention experiment on `Gemma-3-4b-it` model.

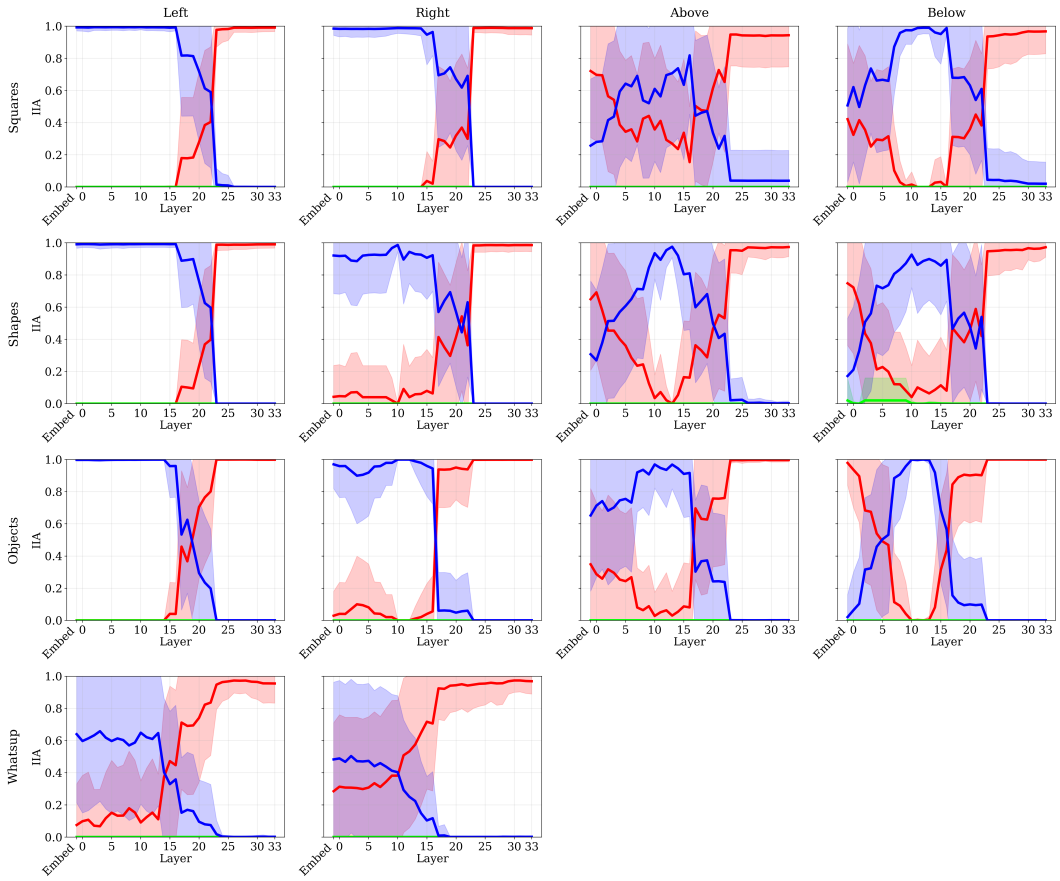



Figure 15: Visual token embeddings of strip tokens interchange intervention experiment on `Gemma-3-4b-it` model.

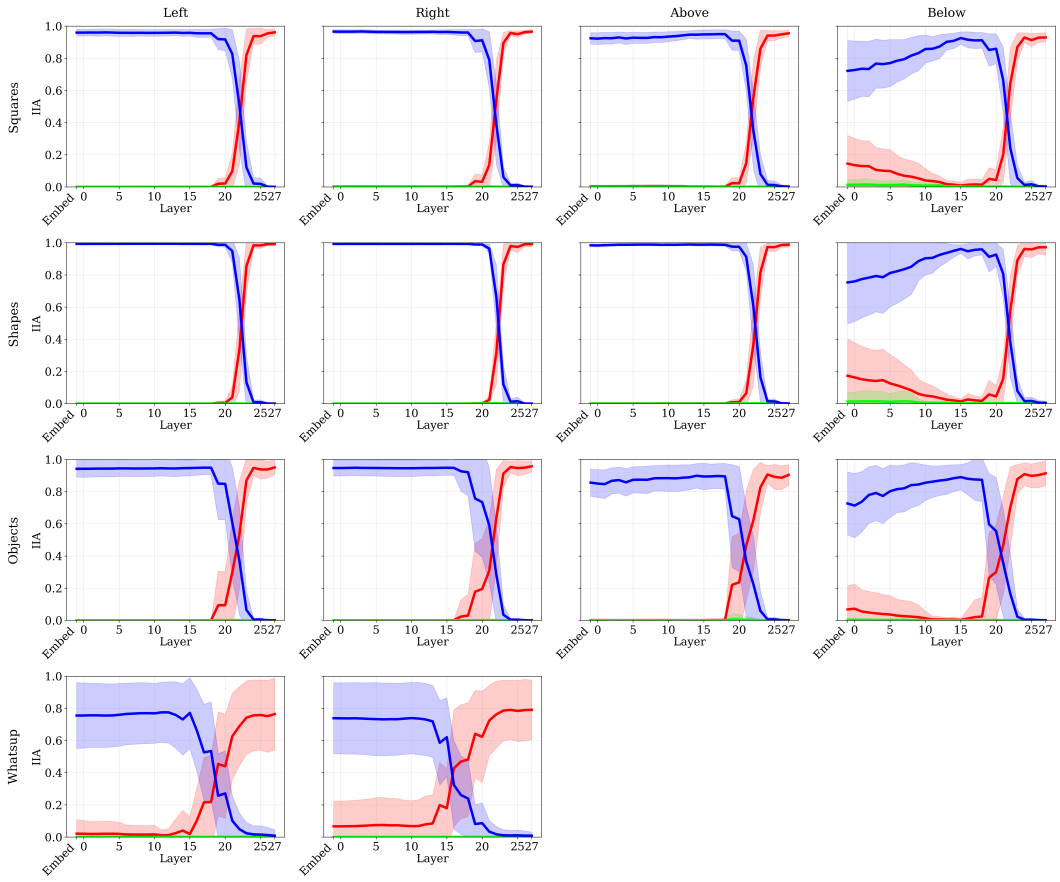



Figure 16: Visual token embeddings of strip tokens interchange intervention experiment on `Qwen2-VL-7B-Instruct` model.

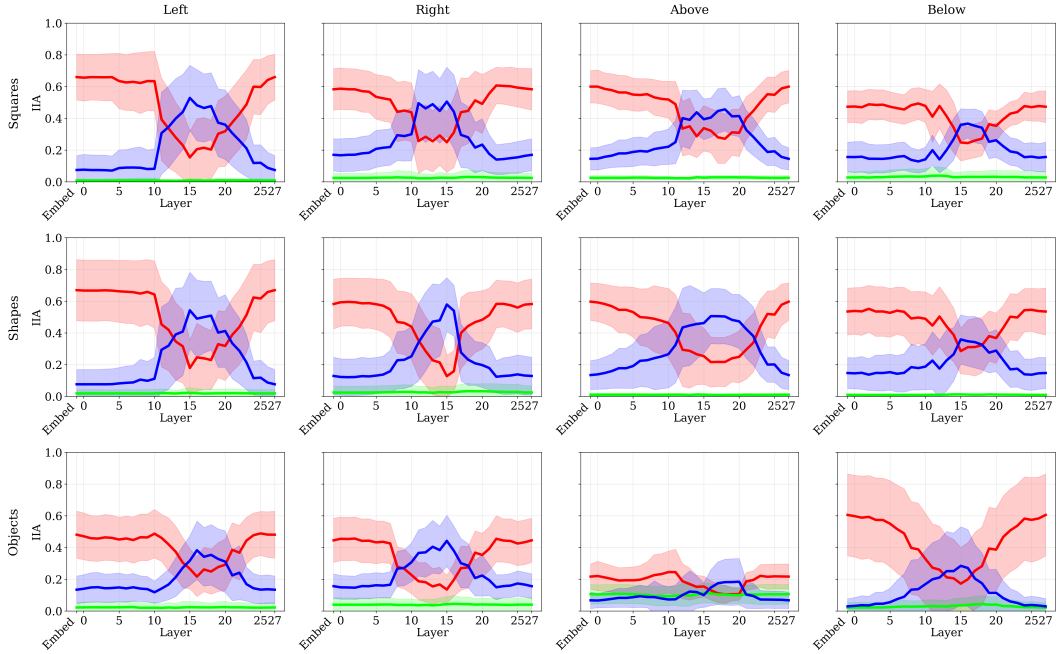



Figure 17: Visual embedding intervention, Qwen

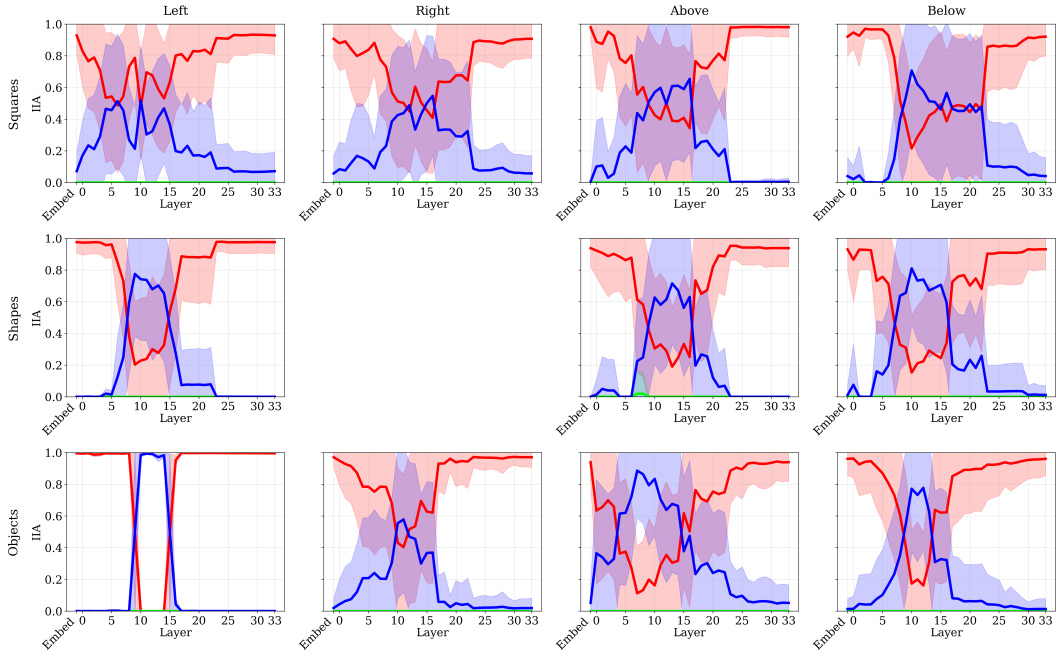



Figure 18: Visual embedding intervention, Gemma

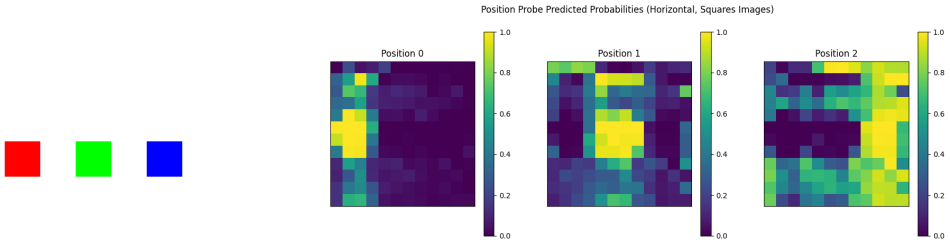



Figure 19: Horizontal position probe for square images, Qwen

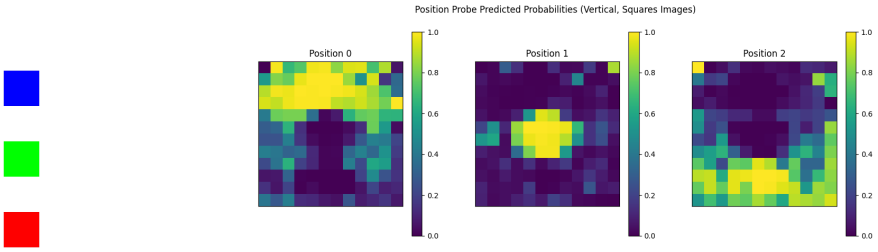



Figure 20: Vertical position probe for square images, Qwen

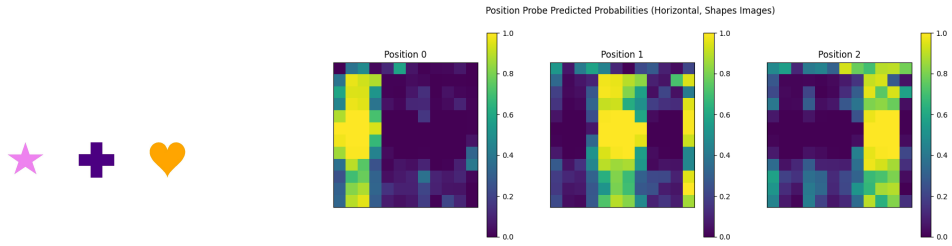



Figure 21: Horizontal position probe for shape images, Qwen

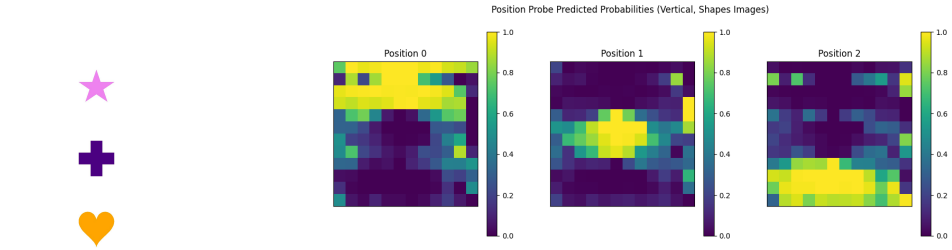



Figure 22: Vertical position probe for shape images, Qwen

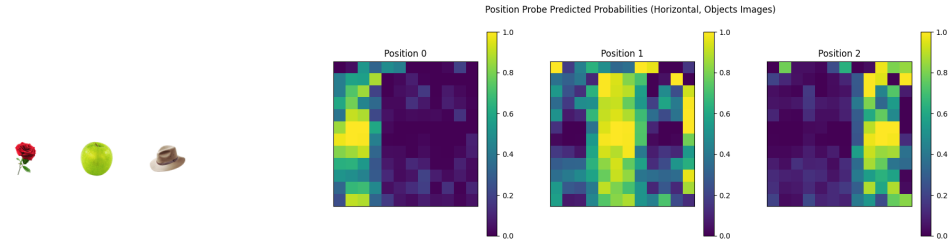



Figure 23: Horizontal position probe for object images, Qwen

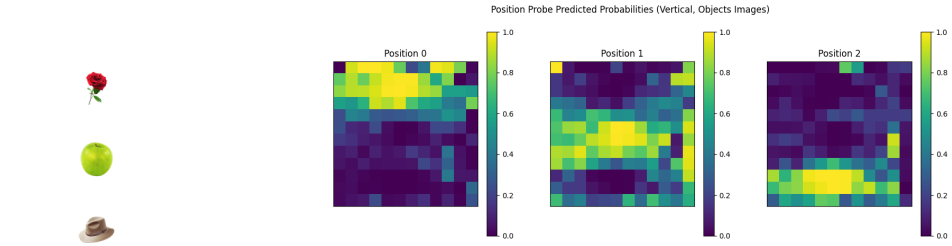



Figure 24: Vertical position probe for shape images, Qwen

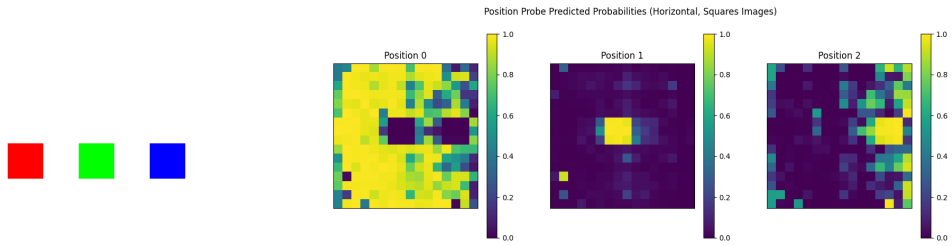



Figure 25: Horizontal position probe for square images, Gemma

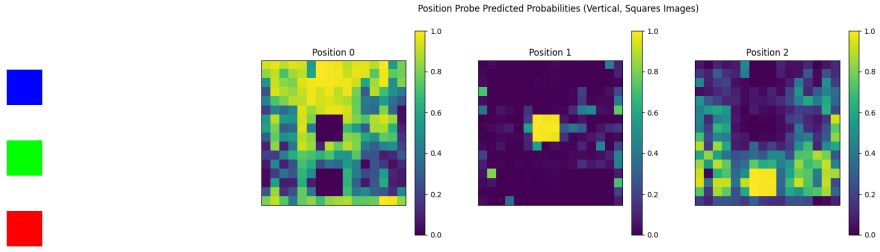



Figure 26: Vertical position probe for square images, Gemma

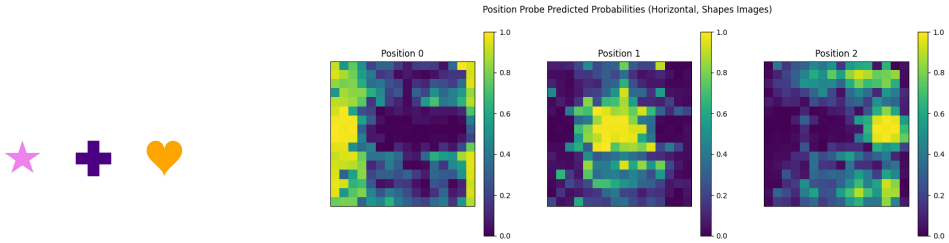



Figure 27: Horizontal position probe for shape images, Gemma

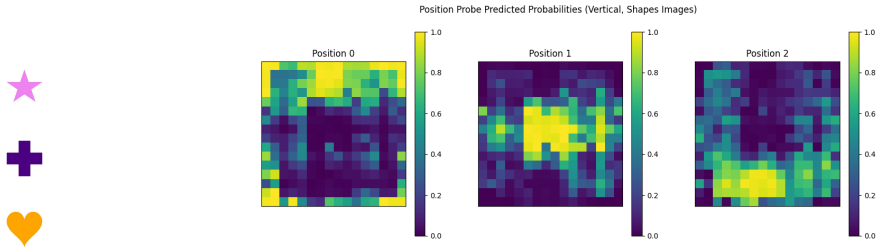



Figure 28: Vertical position probe for shape images, Gemma

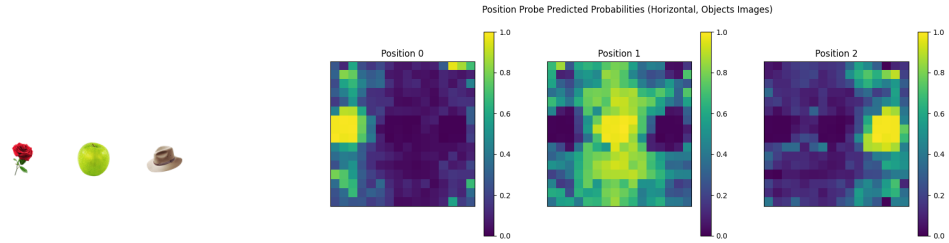



Figure 29: Horizontal position probe for object images, Gemma

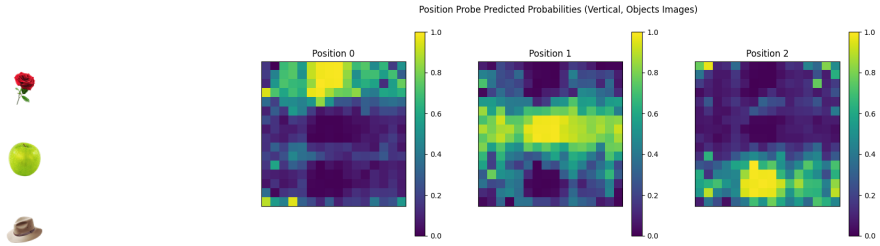



Figure 30: Vertical position probe for object images, Gemma

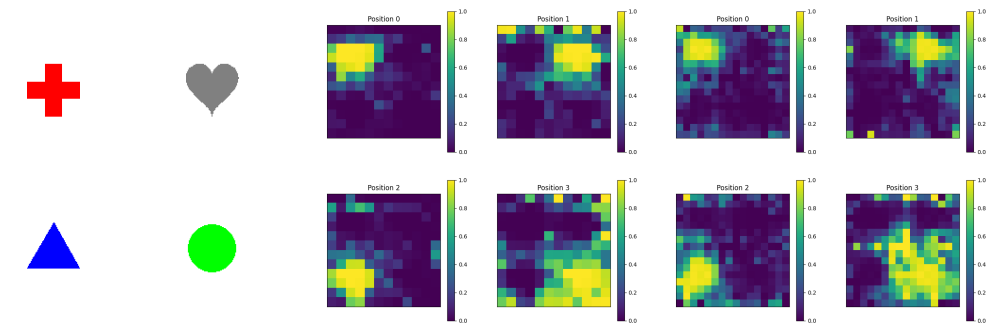



Figure 31: Qwen and Gemma probing results on 2x2 shapes grid.

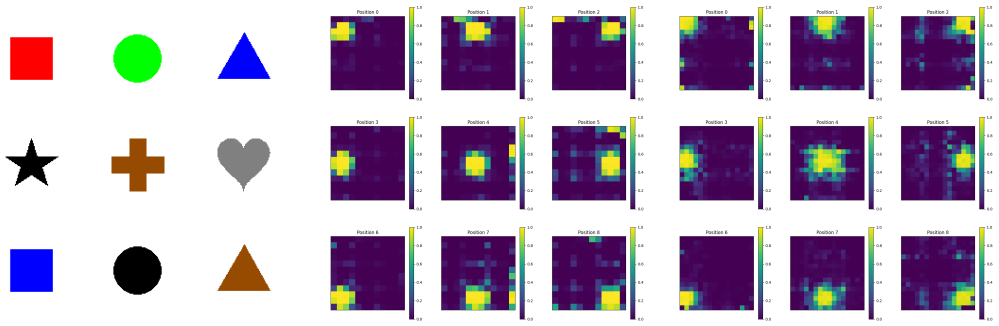


Figure 32: Qwen and Gemma probing results on 3x3 shapes grid.

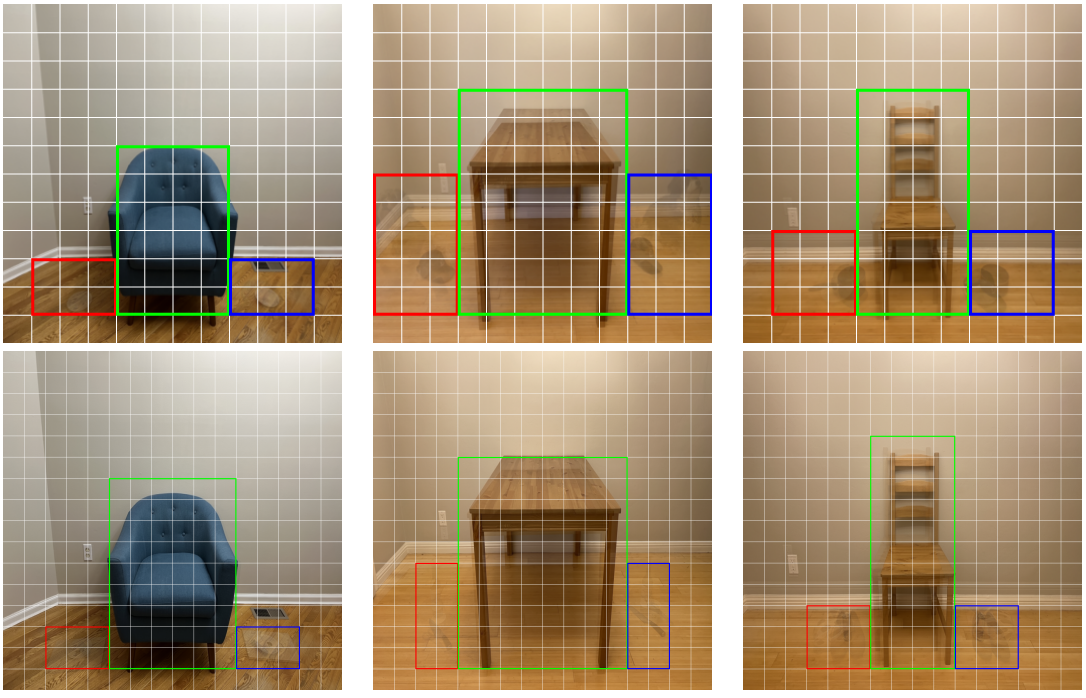


Figure 33: Preprocessing for What'sUp images: constructing bounding boxes around each object.


Figure 34: Qwen What'sUp probe

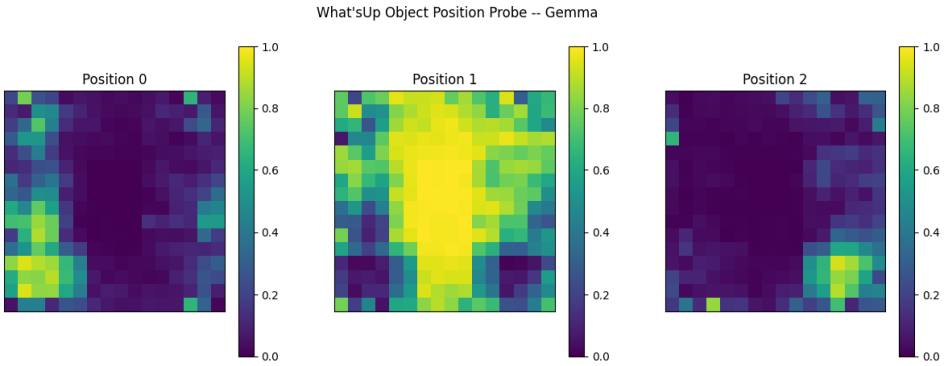



Figure 35: Gemma What'sUp probe