A MECHANISTIC VIEW OF SPATIAL REASONING IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Spatial reasoning is an emerging capability of Vision Language Models (VLMs). As VLMs see broader deployment in real-world settings requiring reliability and interpretability, understanding spatial reasoning mechanisms is critical. Through a mechanistic view, we find evidence that spatial reasoning emerges from a coordinated division of labor in VLMs defined by vision encoder (VE) components that produce tokens consumed by large language model (LLM) components. Specifically, we show that spatial reasoning in VLMs depends strongly on positional encodings in the VE and on interactions between background tokens, while the LLM selectively integrates this information from background image tokens and foreground image tokens through a sparse set of functionally specialized attention heads. Despite our probing task's focus on queried objects, we find that background tokens also play a critical role by providing spatial context that is actively leveraged by the model during spatial reasoning. Based on these findings, we show that mechanistic interpretability can guide visual token compression. The performance on real-world datasets and diverse spatial tasks suggests that the discovered pathways extend beyond the original analysis setting, enabling more efficient visual token processing without sacrificing reasoning performance. Overall, our results improve understanding of the mechanisms underlying spatial reasoning in VLMs, and we demonstrate the resulting insights for inference efficiency.

1 INTRODUCTION

Spatial reasoning (Zheng et al., 2025) is a fundamental capability of VLMs, requiring identifying relevant entities, recovering their arrangement in a scene, and using that information to answer a question. Recent VLMs have shown progress in understanding spatial relations (Chen et al., 2024a), estimating spatial quantities (Chen et al., 2026), and performing compositional spatial reasoning (Lee et al., 2026). As VLMs become widely used, understanding how they reason about space becomes increasingly important for explaining and trusting their behaviors (Chen et al., 2025b). However, the internal mechanisms supporting spatial reasoning remain poorly understood.

This gap has motivated efforts to study the mechanisms underlying spatial relation understanding in VLMs. Li et al. (2026) show that object representations may encode relative spatial relations. Assouel et al. (2026) identify a spatial indexing mechanism for visual binding, through which models use an object's position to associate it with its attributes. Kang et al. (2026) find that information about an object's location can be isolated in the activation of the corresponding text token and that intervening on this information changes the model's spatial judgments. Together, these studies illuminate positional structure and spatial representations associated with target objects. However, what remains less understood is how visual information outside the queried objects contributes to judgments about their spatial relationship. Beyond this object-level focus, we therefore ask:

1. What role do spatial-context tokens play in spatial reasoning beyond target-object tokens?
2. How are positional signals constructed and propagated through visual tokens, and how does the language model utilize that information?

To answer these questions, we focus on how spatial relation grounding is performed by VLMs that combine a ViT encoder with an LLM. This setting allows us to distinguish target-object tokens from the remaining visual tokens, which we refer to as background tokens, and examine their contributions separately. We combine controlled CLEVR-Spatial and real-image What's Up examples,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

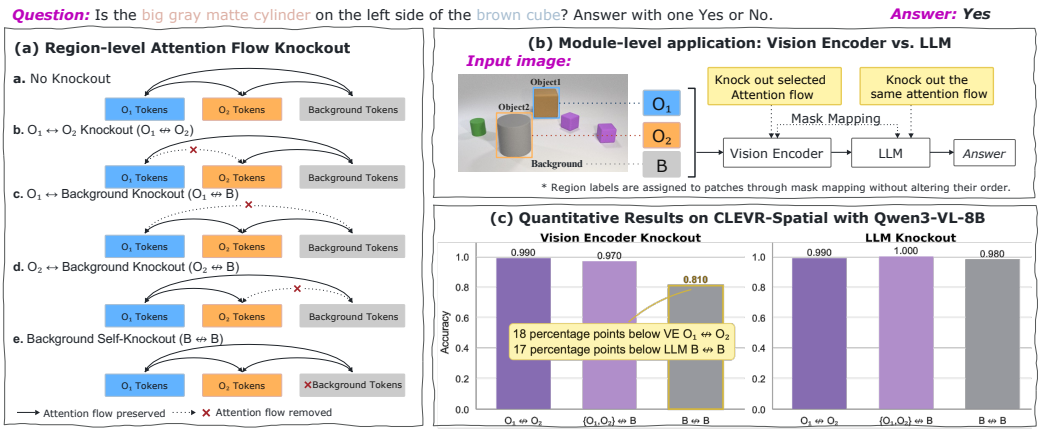


Figure 1: **Region-level attention-flow knockout for analyzing spatial reasoning in VLMs.** (a) We partition image tokens into three semantic groups: two target objects (O_1 and O_2), and the background B, then selectively remove attention flows between specific groups. (b) The same region masks are applied separately in the VE and the LLM, enabling direct comparison of how attention flow affects spatial reasoning per component. (c) On CLEVR-Spatial with Qwen3-VL-8B, $B \nleftrightarrow B$ in the VE causes larger degradation than $O_1 \nleftrightarrow O_2$ in the VE and $B \nleftrightarrow B$ in the LLM, indicating that interactions among background visual tokens have a pronounced effect on spatial reasoning.

perturb positional encodings and region-level visual attention flows, and identify sparse LLM head sets that preserve access to particular visual sources. We then test the practical value of these head sets through visual KV compression on a broader range of spatial benchmarks. We find that VLM spatial reasoning *arises from a division of labor across model ViT and LLM components, which together form a coordinated processing pipeline*. We organize our analysis into three key findings.

Finding 1 (Sec. 3): Spatial relationships depend on positional signals during visual encoding and on interactions among non-target visual tokens. Removing or reflecting positional encodings in the vision encoder substantially disrupts spatial judgments, with effects that vary across spatial axes and relations. Corresponding interventions on visual-to-visual positional encoding in the LLM have much smaller effects. Separately, blocking attention flows among background tokens generally causes greater degradation than blocking direct attention between the queried objects. This dependence occurs in both the vision encoder and the LLM, with its magnitude varying across datasets and relations. Together, these results support a role for position-aware visual encoding and contextual interactions in spatial relation grounding, showing that information outside the queried objects cannot be treated as uniformly dispensable.

Finding 2 (Sec. 4): Access to global visual context for spatial reasoning can be concentrated in sparse LLM head sets. We introduce a differentiable mask search with source-conditioned sufficiency objectives to identify small sets of heads whose access to background or target-object tokens preserves prediction performance. Across Qwen3-VL-8B and InternVL3-8B, these searches identify sparse head sets that retain accuracy close to the full model. These findings characterize how access to target and contextual evidence is distributed within the LLM and show that, although background tokens are important, broad visual access need not be retained uniformly across heads.

Finding 3 (Sec. 5): The identified background-specialized and target-object-specialized heads provide actionable guidance for visual KV compression. We preserve complete visual KV entries for the background heads identified on CLEVR-Spatial and retain a compact target visual-token subset for the remaining heads. This design maintains accuracy comparable to the uncompressed model while retaining only 13.9–16.8% of its visual KV entries and reducing decoding latency by 14.9–24.3%. The broader evaluation extends beyond relation grounding to tasks such as viewpoint selection and mental rotation, demonstrating the utility of the discovered access patterns.

Together, these findings clarify how visual context supports spatial reasoning across the vision encoder and the LLM. They reveal the importance of non-target visual information, show how access to this information can be concentrated in a sparse set of LLM heads, and demonstrate how these insights can guide more efficient VLM inference.

2 Background and Related Work

VLM Backbones, Positional Encoding and Attention Mechanism. A typical VLM backbone (e.g., Qwen-VL and InternVL3) combines a vision transformer (ViT) encoder (Dosovitskiy et al., 2021) with an autoregressive LLM. Judging a spatial relation requires identifying the relevant objects and determining where they lie relative to one another. In a VLM, this information must be recovered from visual tokens whose spatial arrangement is made available through positional encoding (PE). We distinguish these explicit positional signals from image-token order. Qwen2.5-VL (Bai et al., 2025) uses two-dimensional RoPE in its vision encoder and multimodal RoPE (MRoPE) in its LLM. Qwen3-VL (Team, 2025) additionally uses learned absolute position embeddings in its vision encoder, while its LLM uses interleaved MRoPE. InternVL3-8B (Zhu et al., 2025) uses learned absolute position embeddings in its vision encoder and one-dimensional RoPE in its LLM. These positional signals influence how attention aggregates information: absolute embeddings affect the representations from which queries, keys, and values are computed, while RoPE directly modifies query-key similarities through position-dependent rotations. Within each attention head, these similarities determine the weights used to aggregate value vectors from accessible tokens.

Spatial Reasoning. Despite recent progress, VLMs still struggle with complex spatial reasoning (Zhang et al., 2025a). Existing work improves this capability through task decomposition, region-level grounding, and depth information. Chen et al. (2024a) demonstrate chain-of-thought reasoning with an external LLM that decomposes complex spatial questions into simpler queries answered by the trained VLM and combines these answers to reach a final conclusion. Cheng et al. (2024) introduce region-level visual features and an optional depth input to support spatial reasoning about user-specified object regions. Chen et al. (2025a) introduce depth positional encoding to incorporate the 3D physical scale and depth information into visual features, improving spatial measurement and understanding. These approaches build on VLMs that combine a vision encoder with an LLM, enhancing spatial reasoning through positional information and architectural improvements. Their progress motivates a closer examination of the internal mechanisms that support spatial reasoning in this architectural family.

Mechanistic Interpretability for VLMs. Interpretability studies examine how VLMs process and integrate visual and linguistic information across the vision encoder and LLM using logit lens (Neo et al., 2025; Li et al., 2026), linear probing (Yu & Lee, 2025), and attention-flow analysis (Neo et al., 2025; Zhang et al., 2025b). Recent work has investigated mechanisms underlying spatial representation and visual binding, providing important clues for understanding the internal symbolic computational mechanisms. Assouel et al. (2026) identify content-independent spatial indices that bind visual objects to their attributes, along with attention heads that retrieve an object's index, select the target index, and retrieve the associated features. Kang et al. (2026) show that object locations are encoded as linear components in the activations of corresponding object words, and that interventions on these components systematically alter spatial predictions. Complementing these studies of object-related representations, Fan et al. (2026) categorize visual tokens as sink, dead, or alive based on their semantic content and computational behavior, showing that pruning sink and dead tokens can preserve performance. We extend these investigations by examining how positional signals and interactions among target-object and background tokens support spatial reasoning. Through interventions in the vision encoder and the LLM, we analyze the contributions of these visual sources and identify sparse attention-head sets whose access to them preserves task performance.

3 How Visual Position and Context Support Spatial Information

Spatial relation grounding refers to the capability of identifying and reasoning about spatial relationships (e.g., left/right, front/behind, distance, size) between visual entities based on their geometric arrangement in 2D/3D space and the observer's viewpoint. We investigate how such ability depends on explicit positional signals and contextual interactions among visual tokens, including interactions involving non-target tokens that may appear "uninformative" in isolation. We examine these dependencies in both the vision encoder and LLM by perturbing positional encodings and performing region-level attention-flow knockout for vision tokens.

Table 1: **Positional-encoding perturbation (a) and attention-flow knockout (b) in the VE vs. the LLM.** Accuracy (%) of Qwen3-VL-8B with interventions applied in the vision encoder (VE) or the LLM. O_1, O_2: the two target objects; B: background tokens. L–R / F–B / O–U: Left–Right / Front–Behind / On–Under. Blue intensity increases linearly with the accuracy drop relative to the corresponding full-model baseline; **bold**: drop of more than 20 percentage points. The What's Up splits used in (a) and (b) differ (Sec. 3.1), hence the different full-model accuracies.

Stage	(a) Positional-encoding perturbation						(b) Attention-flow knockout					
	Perturbation	CLEVR-Spatial		What's Up			Blocked flow	CLEVR-Spatial		What's Up		
		L–R	F–B	L–R	F–B	O–U		L–R	F–B	L–R	F–B	O–U
Full model	None	99.2	99.7	99.1	99.1	96.5	None	99.2	99.7	99.8	99.5	100.0
Chance	–	50.0	50.0	50.0	50.0	50.0	Chance	50.0	50.0	50.1	51.4	64.4
VE	Remove	**54.1**	**51.6**	**54.1**	**50.9**	**60.0**	$O_1 \leftrightarrow O_2$	98.4	99.5	100.0	99.5	85.6
	Reflect X	**14.6**	97.2	**5.7**	89.9	91.9	$\{O_1, O_2\} \leftrightarrow B$	93.3	98.6	100.0	98.6	100.0
	Reflect Y	92.0	**29.3**	90.5	**53.1**	**52.4**	$B \leftrightarrow B$	88.3	**69.2**	84.3	**66.4**	96.6
LLM	Remove	92.3	99.7	91.9	98.7	96.2	$O_1 \leftrightarrow O_2$	98.8	99.6	99.8	99.5	100.0
	Reflect X	97.5	99.8	98.5	99.1	96.2	$\{O_1, O_2\} \leftrightarrow B$	98.5	99.4	99.8	99.5	100.0
	Reflect Y	99.1	99.6	99.1	99.1	96.5	$B \leftrightarrow B$	93.1	94.9	**50.6**	**22.4**	**62.3**

3.1 Dataset and Evaluation Protocol

CLEVR-Spatial dataset. We construct CLEVR-Spatial, a spatial-relation grounding benchmark from CLEVR-style rendered scenes (Johnson et al., 2017). By randomizing scene configurations and questions within this controlled setting, we avoid confounding factors such as noisy backgrounds, shortcut cues, or dataset bias that may obscure the causal findings. CLEVR-Spatial contains 100 images, each with 4-8 objects varying in shape, size, color, and material, and 2,286 questions of two types. 1,182 Left–Right questions test horizontal relations from the viewer's perspective, whereas 1,104 Front-Behind questions test relative depth ordering. Given an image and a query referring to two objects A and B and a candidate spatial relation R, a VLM predicts whether R holds between A and B via binary response y. Formally, the model prediction is

$$y = \mathrm{VLM}(\mathrm{Image}, \mathrm{Query}(A, B, R)) \in \{\mathrm{Yes}, \mathrm{No}\}. \tag{1}$$

We report accuracy as the proportion of predicted Yes/No answers that match the ground truth. An example one-hop reasoning query for left-right spatial relationships is:

> $\mathrm{Query}(A, B, R) =$ Is the big brown rubber cube to the right of the large gray rubber cylinder from the viewer's perspective? Answer with one word: Yes or No.

where $A =$ the big brown rubber cube, $B =$ the large gray rubber cylinder, $R =$ right.

What's Up dataset. We additionally evaluate on the two-object subsets of What's Up (Kamath et al., 2023). For positional-encoding experiments in Sec. 3.2, we reformulate the original two-object spatial-relation grounding questions as two-choice relation questions, yielding 953 Left-Right, 228 Front-Behind, and 370 Above-Below questions. For attention-flow knockout in Sec. 3.3 that requires object masks, we use 767 questions with manually verified object masks for both targets: 407 Left-Right, 214 Front-Behind, and 146 On-Under. These questions ask the model to select one of two relations, for example: "Where is the wineglass in relation to the table? Answer with one word: on or under." We report accuracy as the proportion of predicted relations that match the ground truth.

3.2 How Do Position Encodings Affect Spatial Reasoning?

We first ask at which stage explicit position signals affect spatial-relation predictions. We apply two positional interventions, Remove and Reflect X/Y, to the vision encoder and LLM of Qwen3-VL-8B. In the vision encoder, Remove zeros the learned absolute position embeddings and computes visual-to-visual attention without Vision-RoPE. Reflect X/Y assigns each patch the absolute embedding and Vision-RoPE coordinate of its horizontally or vertically mirrored location, without moving the patch itself. In the LLM, the interventions modify only the MRoPE contribution to visual-to-visual attention: Remove omits that contribution, while Reflect X/Y uses mirrored image-token coordinates. The visual embeddings received from the vision encoder, text positions, and causal mask remain unchanged.

Table 2: **Accuracy under source-restricted access for the head sets visualized in Figure 2.** Head sets are identified independently for each relation, dataset, and model using the sufficiency objectives keep-bg and keep-tg. Bg and Tg report accuracy (%) when selected heads retain access only to background or target-object tokens, respectively, while non-selected heads are blocked from the evaluated source. Highlighted columns indicate the identification source. Full Acc. denotes accuracy without intervention. #Heads (%) reports the number and proportion of selected KV-head groups for Qwen3-VL-8B GQA, and individual attention heads for InternVL3-8B.

Setting	Condition	Full Acc.	Heads identified with keep-bg			Heads identified with keep-tg		
			Bg	Tg	#Heads (%)	Bg	Tg	#Heads (%)
Qwen3-VL-8B	Left–Right	100.00	**99.16**	97.48	33 (11.46)	97.48	**99.16**	32 (11.11)
relation-specific	Front–Behind	100.00	**100.00**	86.49	7 (2.43)	98.20	**100.00**	20 (6.94)
Qwen3-VL-8B	CLEVR-Spatial	98.28	**98.28**	98.28	35 (12.15)	98.28	**98.28**	32 (11.11)
dataset-specific	What's Up	85.83	**87.50**	74.17	34 (11.81)	71.67	**85.00**	30 (10.42)
InternVL3-8B	CLEVR-Spatial	100.00	**98.00**	82.00	123 (15.69)	58.00	**99.00**	82 (10.46)
dataset-specific	What's Up	82.00	**84.00**	77.00	113 (14.41)	44.00	**87.00**	95 (12.12)

Table 1(a) reports the resulting accuracies on CLEVR-Spatial and What's Up. Across both datasets, spatial-relation predictions are substantially more sensitive to the vision-encoder interventions than to the LLM interventions. On CLEVR-Spatial, removing vision-encoder positional signals reduces Left-Right accuracy from 99.2% to 54.1% and Front-Behind accuracy from 99.7% to 51.6%. Coordinate reflection reveals a directional pattern: horizontal reflection reduces Left-Right accuracy to 14.6% on CLEVR-Spatial and 5.7% on What's Up, whereas vertical reflection reduces Front–Behind accuracy to 29.3% and 53.1%, respectively. On What's Up, vertical reflection also reduces On-Under accuracy from 96.5% to 52.4%. Spatial reasoning seems to depend strongly on positional signals during visual encoding, with sensitivity that varies by spatial axis and relation.

Takeaway. The observed asymmetry suggests that spatial information is incorporated into visual representations with positional encoding. Understanding how the LLM uses these representations can therefore help identify which visual information is important to sustain spatial reasoning.

3.3 Which Visual Interactions Affect Spatial Reasoning and Where?

Another question is whether spatial relations in VLMs are computed through explicit pairwise interactions between the queried ("target") objects, or instead emerge from globally contextualized visual representations that integrate information across the full scene. We perform targeted attention flow knockout interventions that selectively disrupt information exchange among different groups of visual tokens within the Vision Encoder (Figure 1). We consider two categories of visual tokens: target-object tokens and background tokens. Target-object tokens correspond to image tokens containing the queried objects and provide semantic evidence like color, category, and shape. Background tokens correspond to the remaining image regions outside the objects. We design three complementary routing channels to remove attention between (Figure 1): (i) object-and-object tokens; (ii) object-and-background tokens; and (iii) background-and-background tokens.

Figure 1(c) summarizes the knockout results on the CLEVR-Spatial dataset, while Table 1(b) reports accuracy by dataset, relation, and intervention stage. Disrupting direct object-object attention causes only negligible performance degradation. In the Vision Encoder, removing $O_1 \leftrightarrow O_2$ interactions retains an accuracy of 0.990, while removing $\{O_1, O_2\} \leftrightarrow B$ interactions still preserves 0.970 accuracy. By contrast, suppressing background-background interactions ($B \leftrightarrow B$) causes accuracy to collapse to 0.810. Large-scale disruption of background-token communication causes severe degradation on What's Up, particularly in the LLM. Across the tested relations, blocking attention among non-target visual tokens generally has a larger effect on predictions than blocking direct attention between the queried objects. This dependence appears in both the vision encoder and the LLM, with its magnitude varying across datasets and relations.

Takeaway. Interactions with background tokens contribute substantially to spatial relation grounding and should not be treated as uniformly redundant in the vision encoder or LLM. We next examine whether LLM access to those background tokens is distributed across heads or concentrated in a subset (Sec. 4), guiding head-aware selective visual KV retention for efficient inference (Sec. 5).

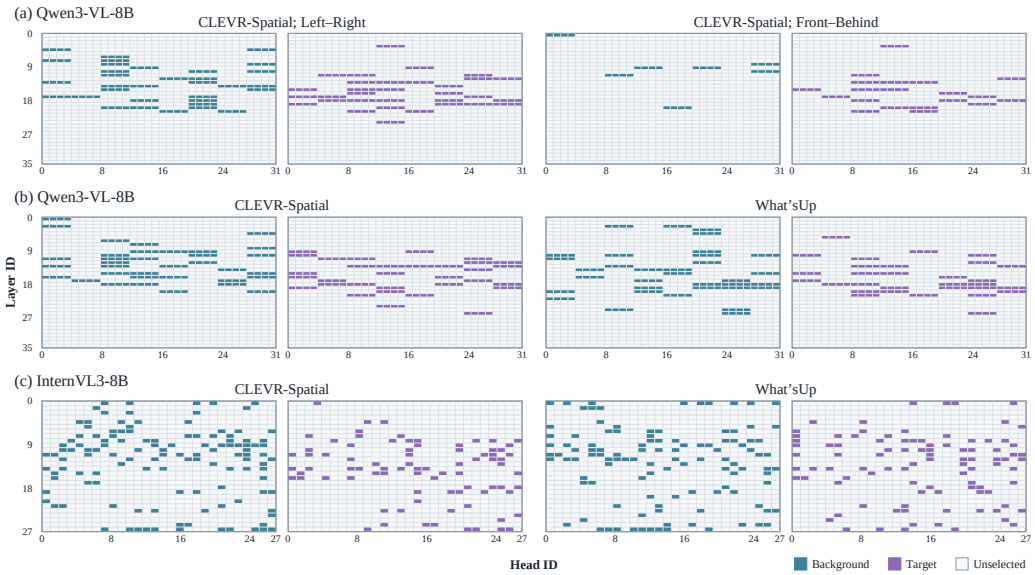


Figure 2: **Layer-head distributions of spatial heads across relations, datasets and models.** Blue and purple cells denote heads identified by keep-bg and keep-tg, respectively; gray cells indicate unselected heads. (a) Relation-specific masks for Left-Right and Front-Behind on CLEVR-Spatial with Qwen3-VL-8B. (b) Dataset-specific masks for CLEVR-Spatial and What's Up with Qwen3-VL-8B. (c) Corresponding masks for InternVL3-8B illustrate how searched spatial heads differ across model architectures. Rows and columns index transformer layers and attention heads, respectively. For Qwen3-VL-8B, each selected GQA group is expanded to its four associated query heads.

4 The LLM Reads Spatial Information with Specialized Heads

Spatial relation grounding in VLMs relies on both object-centric evidence from the queried objects and spatial context from the background (Sec. 3). We study how these two sources of tokens are differentially routed within the LLM, and propose a method to identify specialized attention heads responsible for integrating object-centric and contextual information during spatial reasoning.

4.1 Identifying Information-Source Attention Heads

We automatically identify attention heads associated with different visual information sources using a differentiable binary mask search strategy. We first define *spatial heads* as a functional group of heads associated with spatial reasoning, categorized based on their visual information sources, comprising *background heads* or *target-object heads*. We probe the LLM component of the VLM by selectively blocking information flow from different visual token sources, and analyze how such interventions affect the VLM's predictions. We first describe our masking strategy for blocking information flow. Then, to identify attention heads associated with the proposed semantic visual information sources, we formulate an intervention-based objective characterizing head functionality from the perspective of *sufficiency*. We also formulate a complementary *necessity* loss and provide further analyses in Appendix E.3.

Masking Strategy. Inspired by (Bhaskar et al., 2024; Niu et al., 2025), we introduce a binary mask m_j for each attention head h_j. When $m_j = 1$, the head remains active; when $m_j = 0$, its output is suppressed, blocking the information it extracts from specific visual tokens. Since discrete masks are non-differentiable, we adopt a Gumbel-sigmoid relaxation with a straight-through estimator. We sample $s_i = \sigma\left(\frac{l_i - \log \frac{\log U_1}{\log U_2}}{\tau}\right)$ and construct the mask as $m_i = \left[\mathbf{1}_{s_i > 0.5} - s_i\right]_{\text{detach}} + s_i$, where l_i is a learnable logit, τ is a temperature parameter, and $U_1, U_2 \sim \text{Uniform}(0, 1)$. This relaxation provides a differentiable approximation of binary head selection while preserving discrete behavior in the forward pass. After we get the selected attention heads, we suppress information flow from

background tokens or target tokens for these heads. Functionally, we suppress information by setting the embeddings of selected tokens to zero before computing attention for the selected heads. Zeroing removes the information carried by these tokens from the key and value representations, preventing the selected heads from using information from the semantic region.

Let $x \in \mathbb{R}^{m \times n}$ denote the input embeddings to the LLM, for m input tokens and n the embedding dimension. Let $\mathcal{S} \in \{\mathrm{bg}, \mathrm{tg}\}$ denote a token source (background or target-object). For a selected head subset $\mathcal{H}$, we denote by $x^{\mathcal{S}}_{\mathcal{H}}$ the modified input in which only heads in $\mathcal{H}$ are allowed to attend to source-$\mathcal{S}$ tokens, while all other heads are blocked from accessing them.

Sufficiency Loss. To identify a minimal subset of heads that can *preserve* performance using background or target information, we optimize the prediction confidence under the intervention:

$$\mathcal{L}_{\text{keep-bg}} = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}} \left[\log P\Big(y^* \mid x^{bg}_{\mathcal{H}}\Big) \right], \tag{2}$$

where $x^{bg}_{\mathcal{H}}$ denotes the modified input in which the unselected heads are prevented from attending to background tokens (*bg*) via embedding masking. The same formulation is applied to target object information (*tg*) by replacing the modified input $x^{bg}_{\mathcal{H}}$ with $x^{tg}_{\mathcal{H}}$.

Sparsity Penalty. To avoid selecting an excessive number of heads, we also utilize a sparsity regularization that encourages the mask to activate only a small subset of attention heads. The final objective is $\mathcal{L} = \mathcal{L}_{bg} + \lambda \cdot \frac{1}{|\mathcal{H}|} \sum_{i=1}^{|\mathcal{H}|} \sigma(l_i)$, where $\mathcal{H}$ denotes the number of full attention heads and λ controls the sparsity strength. This objective encourages the discovery of a minimal set of heads whose information flow from specific visual regions significantly affects the model's prediction.

4.2 Experiments and Results

We conduct spatial head identification experiments with Qwen3-VL-8B-Instruct and InternVL3-8B. For Qwen3-VL-8B-Instruct, the units of selection are its 288 key-value (KV) head groups under grouped-query attention (GQA), with each mask applied jointly to all query heads sharing the same KV head. For InternVL3-8B, masks operate on individual attention heads. We split CLEVR-Spatial into training (80%), validation (10%), and test (10%) sets. Head masks are optimized on the training set for 20 epochs using AdamW (Loshchilov & Hutter, 2019), with the checkpoint selected on the validation set based on prediction logit differences and the number of selected heads. The sparsity coefficient λ is selected from the range $\lambda \in [0.1, 12]$, and the learning rate is chosen from $[10^{-3}, 10^{-2}]$. We fix the temperature $\tau = 1.0$. For different head types and loss functions, we perform a grid search over λ and the learning rate on the validation set. All reported accuracies are evaluated on the test set. Using the sufficiency objectives (Sec. 4.1), we identify relation-specific head sets for Left-Right and Front-Behind, and dataset-level head sets for CLEVR-Spatial and What's Up. Figure 2 visualizes the selected heads across layers and head indices. Table 2 reports accuracy under source-restricted access: each identified head set is evaluated with access to background or target-object tokens, while all non-selected heads are blocked from the evaluated source.

Sparse head sets suffice to retain performance under restricted access to background or target-object tokens. Our sufficiency-based searches identify compact head sets for retaining access to each visual source. As shown in Table 2, the background-focused searches select only 2.43–15.69% of the available head units, while the target-focused searches select 6.94–12.12%. Under their respective identification conditions, all selected sets retain accuracy within 2 percentage points of, or above, the full-model baseline. For example, on Qwen3-VL-8B, preserving background access in just 7 of 288 KV-head groups (2.43%) maintains 100.00% accuracy on Front–Behind. At the dataset level, retaining background or target-object access in 35 (12.15%) or 32 (11.11%) KV-head groups, respectively, matches the full-model accuracy of 98.28% on CLEVR-Spatial. The same pattern extends to What's Up and InternVL3-8B, where compact head sets also retain near-baseline performance. These results show that direct access to each visual source can be restricted to a small subset of LLM heads while preserving spatial relation grounding accuracy under the tested interventions.

Sparse head access preserves performance only when useful visual information remains accessible. Holding the selected heads fixed, changing their accessible visual tokens can substantially reduce accuracy (Table 2). For InternVL3-8B, switching target-selected heads from target-object to

background access lowers accuracy from 99.00% to 58.00% on CLEVR-Spatial and from 87.00% to 44.00% on What's Up. Conversely, Qwen3-VL-8B's background-selected Front-Behind heads achieve 100.00% with background access but 86.49% with target-object access. This dependence varies across settings: Qwen3-VL-8B's dataset-level CLEVR-Spatial sets achieve 98.28% under either condition. Thus, sufficiency is a property of the evaluated head-token access configuration; a head set that preserves performance under one access condition need not do so under another.

Takeaway. The results suggest a functional decoupling in spatial reasoning: target heads encode local semantic evidence tied to object identity, while background heads support information from the spatial visual context. Further, these heads comprise a sparse subset of attention heads across the VLM, providing a basis for head-dependent visual KV retention during inference.

5 Spatial Heads for Efficient Spatial Reasoning

Section 4 identifies a sparse set of attention heads whose access to background tokens supports spatial reasoning. We translate this finding into a head-dependent visual-token budget that preserves full visual access for the identified heads and retains a compact set of foreground image tokens for the remaining heads. We evaluate this design on broader spatial tasks to test whether the head-level analysis provides useful guidance for efficient inference beyond the setting used for head discovery.

5.1 Spatial-Head-Guided Visual Token Compression

We implement the method on Qwen3-VL-8B, whose language model contains 36 layers with eight KV heads each. A fixed background-head mask obtained on CLEVR-Spatial selects 35 of the 288 layer–KV-head pairs (12.15%). These pairs retain all visual KV entries; each remaining pair retains the same subset of $K = 8$ visual tokens. Let $\mathcal{I}$ denote the visual-token indices, $\mathcal{P}$ all textual and special-token indices, and $\mathcal{F} \subseteq \mathcal{I}$ the selected foreground subset. The retained prompt indices are

$$\mathcal{R}_{l,h} = \begin{cases} \mathcal{P} \cup \mathcal{I}, & (l,h) \in \mathcal{H}_{\mathrm{bg}}, \\ \mathcal{P} \cup \mathcal{F}, & (l,h) \notin \mathcal{H}_{\mathrm{bg}}, \end{cases} \qquad |\mathcal{F}| = \min(K, |\mathcal{I}|).$$

All query heads sharing a KV head use the same retention policy. We use the same head mask and token budget across all datasets. We select the shared visual-token set $\mathcal{F}$ once using the value representations at layer 20. Following a similar scoring strategy to Sun et al. (2026), we score each visual token by combining its relevance to the post-image text tokens with its similarity to semantic anchors, defined as the top 1% of visual tokens by text relevance. These anchors form a priority pool. We retain its highest-scoring tokens up to the budget $K = 8$, filling any remaining slots from the global semantic ranking. During prefill, we preserve the full residual sequence through layer 20, the final layer containing a selected full-retention head. Once $\mathcal{F}$ is determined, we compact the cached visual KV entries of non-selected heads to $\mathcal{F}$, while preserving the complete visual KV sequence for selected heads. From layer 21 onward, we also reduce the residual sequence to $\mathcal{P} \cup \mathcal{F}$, where $\mathcal{P}$ contains all textual and special tokens. This design combines head-dependent KV retention with sequence-level pruning in subsequent layers.

5.2 Experiments and Results

We compare our method with Full KV, IVC-Prune (Sun et al., 2026), and Nüwa (Huang et al., 2026) on Qwen3-VL-8B. We evaluate all methods on the complete CLEVR-Spatial, What's Up, and SpinBench. CLEVR-Spatial contains 2,286 left–right and front–behind questions. What's Up contains 4,958 examples across all six subsets, covering spatial relations in single-object and pairwise settings. SpinBench (Yao et al., 2025) contains 2,599 examples spanning 51 task types and four visual domains, including viewpoint selection, mental rotation, relation transformation, and dynamic spatial reasoning. This coverage extends the evaluation beyond the spatial relation grounding used for head discovery. Table 3 reports accuracy, visual KV retention, and inference efficiency. We report prefill and decoding costs separately because our method compresses the cache after full-prompt prefill. This distinction separates reductions in retained visual KV from their realized effect on inference latency. We also compare all three methods on general visual reasoning benchmarks MMStar (Chen et al., 2024b) and MME (Fu et al., 2026) in Appendix F and Table 11.

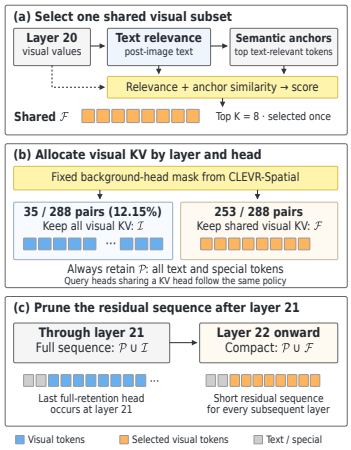


Figure 3: Spatial-head-guided KV retention with residual-sequence pruning. Background heads retain full visual KV; others retain $\mathcal{F}$.

Table 3: Task accuracy and inference efficiency of Qwen3-VL-8B under KV compression. Full KV is the uncompressed reference; among compression methods, **bold** and underline mark the best and second best. Tokens: average number of retained visual tokens per layer; arrows indicate preferred directions.

Dataset	Method	Acc. (%)↑	Latency E2E↓	Latency Prefill↓	Latency Decode↓	Throughput ↑	Tokens (avg.)↓
CLEVR-Spatial	Full KV	99.52	0.1494	0.0828	0.0666	23.60	150.00
	IVC-Prune	97.24	0.1437	0.0870	0.0568	27.95	75.00
	Nüwa	99.34	**0.1250**	**0.0720**	**0.0530**	**28.66**	71.92
	Ours	**99.52**	0.1429	0.0862	0.0567	28.09	**25.26**
What's Up	Full KV	73.40	0.1530	0.1056	0.0474	23.48	401.79
	IVC-Prune	73.34	**0.1456**	**0.1057**	0.0400	27.82	200.89
	Nüwa	72.65	0.1658	0.1267	**0.0391**	28.37	193.07
	Ours	**73.44**	0.1472	0.1081	**0.0391**	**28.52**	**55.86**
SpinBench	Full KV	61.95	0.2232	0.0912	0.1320	23.95	251.05
	IVC-Prune	**62.22**	0.2022	0.0934	0.1088	28.45	125.52
	Nüwa	61.25	0.1976	**0.0871**	0.1105	28.66	121.30
	Ours	**62.22**	**0.1953**	0.0954	**0.0999**	**28.69**	**37.54**

Performance on Spatial Reasoning Tasks. As shown in Table 3 and Table 11, our method preserves Full KV accuracy across five datasets while retaining only 13.9–16.8% of the visual KV entries. Compared with existing compression methods, it achieves comparable accuracy to IVC-Prune and consistently higher accuracy than Nüwa, using over 64% fewer visual KV entries than either baseline. The same head mask and token budget apply across all three datasets without dataset-specific tuning, extending from spatial-relation grounding to the viewpoint, rotation, and dynamic reasoning tasks covered by SpinBench. These results demonstrate the practical value of the identified background heads: preserving their access to complete visual context enables substantial compression in the remaining heads and maintains performance across a broader spatial evaluation.

Efficiency Analysis. Table 3 and Table 11 shows that the reduction in visual KV retention translates into faster decoding across all five datasets, reducing decoding latency by 14.9–24.3% relative to Full KV. These gains offset the modest increase in prefill latency, yielding end-to-end latency reductions of up to 12.5%. Our method achieves the lowest end-to-end latency on SpinBench, while IVC-Prune and Nüwa are faster on What's Up and CLEVR-Spatial, respectively. The proposed allocation delivers practical inference gains alongside substantial visual KV compression.

Takeaway. We find that spatial heads identified by path-level interventions provide actionable guidance for efficient inference. Our results show that preserving the small subset of heads responsible for global spatial structure while aggressively compressing redundant visual memory elsewhere retains spatial reasoning ability and general visual ability while reducing computational overhead.

6 Conclusion

In this paper, we investigate spatial reasoning in VLMs that combine a ViT-based vision encoder with an autoregressive LLM, focusing on a capability important for downstream tasks that require understanding object arrangements and spatial relations. In the studied VLMs, we find that spatial structure first arises inside the vision encoder, where attention across image patches composes positional cues into a spatially organized representation. Spatial relation grounding depends strongly on interactions among background tokens. The LLM then reads out this structure with two groups of specialized heads: background heads and target heads, showing that the importance of visual context does not require uniform access across heads. Building on the findings, we preserve full visual KV access for the identified background heads while compressing visual KV entries in the remaining heads. This design maintains accuracy comparable to the uncompressed model across spatial benchmarks while substantially reducing visual KV retention and decoding latency. Together, our results clarify how positional signals and visual context support spatial reasoning and demonstrate how mechanistic analysis can guide more efficient VLM inference.

AI USE STATEMENT

In this work, we used generative AI tools to assist with figure layout, writing, and modifying experimental code, including code for generating object masks for the What's Up dataset. All generated masks used in our experiments were manually reviewed. We also used generative AI tools to discuss experimental results. GPT-4o-mini was used in the evaluation of the efficiency experiments in Sec. 5. All AI-assisted code and outputs were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Rim Assouel, Declan Iain Campbell, Yoshua Bengio, and Taylor Whittington Webb. Visual symbolic mechanisms: Emergent symbol processing in vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3RQ863cRbx.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. URL https://api.semanticscholar.org/CorpusID:276449796.

Adithya Bhaskar, Alexander Wettig, Dan Friedman, and Danqi Chen. Finding transformer circuits with edge pruning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=8oSY3rA9jY.

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, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=evP9mxNNxJ.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a. URL https://openreview.net/forum?id=PRCizVyL1K.

Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. SD-VLM: Spatial measuring and understanding with depth-encoded vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=PRCizVyL1K.

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, 2025b. 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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Yingqi Fan, Junlong Tong, Anhao Zhao, and Xiaoyu Shen. What do visual tokens really encode? uncovering sparsity and redundancy in multimodal large language models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2026. URL https://openreview.net/forum?id=DgH9YCsqWm.

Yihong Huang, Fei Ma, Yihua Shao, Jingcai Guo, Zitong YU, Laizhong Cui, and Qi Tian. Nüwa: Mending the spatial integrity torn by VLM token pruning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=C9yclwdquU.

Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, XinQiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=6nZKT2rL0H.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9161–9175, 2023.

Raphi Kang, Hongqiao Chen, Georgia Gkioxari, and Pietro Perona. Linear mechanisms for spatiotemporal reasoning in vision language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2zXRGiorSu.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pp. 5583–5594. PMLR, 2021.

Youngwan Lee, Soojin Jang, Yoorhim Cho, Seunghwan Lee, Yong-Ju Lee, and Sung Ju Hwang. Multihopspatial: Multi-hop compositional spatial reasoning benchmark for vision-language model. In *European Conference on Computer Vision (ECCV)*, 2026.

Yueyan Li, Chenggong Zhao, Zeyuan Zang, Caixia Yuan, and Xiaojie Wang. Reading images like texts: Sequential image understanding in vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qYcZVezPW7.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chanJGoa7f.

Jingcheng Niu, Xingdi Yuan, Tong Wang, Hamidreza Saghir, and Amir H. Abdi. Llama see, llama do: A mechanistic perspective on contextual entrainment and distraction in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16218–16239, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.791. URL https://aclanthology.org/2025.acl-long.791/.

Zhichao Sun, Yidong Ma, Gang Liu, Nemo Chen, Xu Tang, Yao Hu, and Yongchao Xu. IVC-prune: Revealing the implicit visual coordinates in LVLMs for vision token pruning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=46LbXtFgBm.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Jianzhu Yao, Kevin Wang, Ryan Hsieh, Haisu Zhou, Tianqing Zou, Zerui Cheng, Zhangyang Wang, and Pramod Viswanath. SPIN-bench: How well do LLMs plan strategically and reason socially? In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=mgsS73kvOA.

Zhuoran Yu and Yong Jae Lee. How multimodal LLMs solve image tasks: A lens on visual grounding, task reasoning, and answer decoding. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=kjNJYWvfPA.

Wanyue Zhang, Yibin Huang, Yangbin Xu, JingJing Huang, Helu Zhi, Shuo Ren, Wang Xu, and Jiajun Zhang. Why do mllms struggle with spatial understanding? a systematic analysis from data to architecture, 2025a. URL https://arxiv.org/abs/2509.02359.

Zhi Zhang, Srishti Yadav, Fengze Han, and Ekaterina Shutova. Cross-modal information flow in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19781–19791, 2025b.

Xu Zheng, Zihao Dongfang, Lutao Jiang, Boyuan Zheng, Yulong Guo, Zhenquan Zhang, Giuliano Albanese, Runyi Yang, Mengjiao Ma, Zixin Zhang, Chenfei Liao, Dingcheng Zhen, Yuanhuiyi Lyu, Yuqian Fu, Bin Ren, Linfeng Zhang, Danda Pani Paudel, Nicu Sebe, Luc Van Gool, and Xuming Hu. Multimodal spatial reasoning in the large model era: A survey and benchmarks, 2025. URL https://arxiv.org/abs/2510.25760.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Appendix

A Limitations **13**

B Broader Impact **13**

C CLEVR-Spatial Construction Details **13**

D Additional Experiments for Sec. 3 **14**

D.1 Finding 1 Across Different VLMs 14

D.2 Finding 1 Across Different Datasets 15

D.3 Influence of token order 16

E Additional Experiments for Sec. 4 **16**

E.1 Finding 2 Across Different VLMs 16

E.2 Finding 2 Across Datasets 17

E.3 Additional Necessity-Based Spatial Head Analysis of Qwen and SpaceQwen . . . 17

F Additional Experiments for Sec. 5 **18**

A Limitations

Our analysis focuses on the Qwen family VLMs and InternVL3-8B, composed of a dedicated vision encoder and a separate autoregressive language decoder. Therefore, our findings may not directly generalize to other architectures such as ViLT-style models (Kim et al., 2021), where visual and language inputs are processed within a shared transformer stack without an explicit vision encoder. Nevertheless, the proposed mechanistic methods may extend to other architectures with different visual backbones. Establishing the extent to which our findings generalize across broader classes of VLM architectures remains an important direction for future work.

B Broader Impact

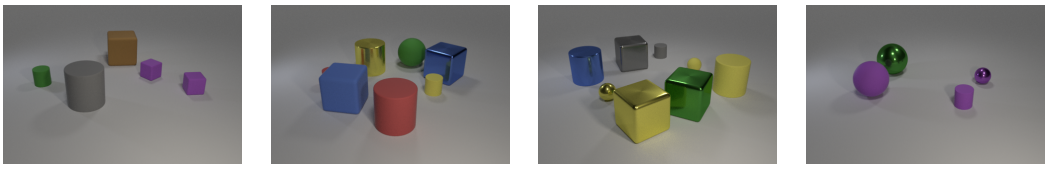
VLMs are increasingly used in settings where spatial understanding affects reliability and safety like robotics and autonomous navigation. This work aims to support more responsible deployment of such systems by providing a mechanistic account of how VLMs form and use spatial information. Our analysis also offers tools for designing more efficient inference methods that preserve critical visual evidence while reducing computation.

C CLEVR-Spatial Construction Details

Scene generation. Scenes are rendered at 480×320 pixels, with an average of 5.61 objects per scene. Object attributes are sampled from three shapes (cube, sphere, cylinder), eight colors (gray, red, blue, green, brown, purple, cyan, yellow), two materials (rubber, metal), and two sizes (small, large). Objects are placed on a ground plane using jittered grid locations or uniformly sampled coordinates. The renderer enforces a minimum inter-object clearance of 0.4 Blender units, a minimum projected separation of 0.6 units along the camera-relative left–right and front–behind axes, and at least 300 visible pixels per object. Camera and lighting positions are randomly perturbed. Annotations include object attributes, 3D and image-space coordinates, segmentation masks, bounding boxes, and pairwise spatial relations.

Relation annotation. For each ordered object pair, the relative 3D displacement is projected onto the camera-relative left, right, front, and behind direction vectors. A relation holds when the corresponding projection exceeds 0.1. Above and below relations are excluded from both annotations and question generation.

Question generation. Questions describe objects by size, color, material, and shape, using multiple templates and lexical synonyms. Each functional program uniquely identifies the reference object, retrieves objects satisfying the queried relation, and checks for a match to the target description. Each program therefore traverses exactly one spatial-relation edge. For every positive question, a paired negative is constructed by swapping the two objects while retaining the queried relation. The resulting subsets contain 591 positive and 591 negative Left–Right questions, and 552 positive and 552 negative Front–Behind questions.



(a) Left–Right.
Prompt: Is the big brown rubber cube to the right of the large gray rubber cylinder from the viewer's perspective?
Gold answer: `True`

(b) Left–Right.
Prompt: Is the big red rubber cylinder on the right side of the large blue shiny cube?
Gold answer: `False`

(c) Front–Behind.
Prompt: Is the big yellow shiny block positioned in front of the small gray matte cylinder?
Gold answer: `True`

(d) Front–Behind.
Prompt: Is the large green metal ball in front of the large purple matte sphere?
Gold answer: `False`

Figure 4: Representative CLEVR-Spatial examples for Left–Right and Front–Behind relation grounding.

D Additional Experiments for Sec. 3

D.1 Finding 1 Across Different VLMs

We evaluate the generality of our two mechanistic observations across VLMs on the CLEVR-Spatial Left–Right task, testing positional-encoding interventions and attention-flow knockout in more VLMs.

Positional-encoding interventions. As shown in Table 4, shuffling positional encodings in the vision encoder reduces accuracy across all models. Corresponding LLM interventions generally have smaller effects, with sensitivity varying by model and perturbation.

Table 4: Accuracy (%) under positional-encoding perturbations across four VLMs on CLEVR-Spatial Left–Right.

Module	PE perturbation	Qwen2.5-VL-3B	SpaceQwen2.5-VL-3B	Qwen3-VL-8B	InternVL3-8B
Full VLM	None	100.00	100.00	100.00	100.00
VE	Shuffle XY	49.87	49.19	56.40	33.12
VE	Shuffle X	49.96	93.12	87.12	21.45
VE	Shuffle Y	95.07	49.79	60.49	42.26
LLM	Shuffle XY	99.92	99.90	91.81	92.26
LLM	Shuffle X	100.00	100.00	90.10	95.30
LLM	Shuffle Y	99.92	100.00	99.83	93.53

Attention-flow knockout. Table 5 compares attention-flow knockout across five VLMs on CLEVR-Spatial Left–Right. Across all models, removing background–background interactions in

the VE yields the lowest accuracy among the tested flows, whereas removing direct object–object interactions leaves accuracy largely intact. The effect of object–background knockout varies across models: it is comparatively limited in the Qwen family but substantially larger in InternVL3-8B. Background–background knockout is also consistently more disruptive in the VE than in the LLM, although the magnitude of this difference varies across models. These results identify background-token interactions as a consistent source of sensitivity in spatial relation grounding, while highlighting model-dependent reliance on other attention flows.

Table 5: Accuracy (%) under attention-flow knockout across five VLMs on CLEVR-Spatial Left–Right. Columns specify the attention flow removed in the vision encoder (VE) or LLM. O_1 and O_2 denote the queried objects, and B denotes background tokens.

	Vision encoder			LLM		
Model	$O_1 \leftrightarrow O_2$	$\{O_1, O_2\} \leftrightarrow B$	$B \leftrightarrow B$	$O_1 \leftrightarrow O_2$	$\{O_1, O_2\} \leftrightarrow B$	$B \leftrightarrow B$
Qwen2.5-VL-3B	99.90	99.50	51.70	99.80	99.30	98.90
SpaceQwen2.5-VL-3B	100.00	99.40	51.10	100.00	99.60	97.50
Qwen3-VL-4B	100.00	99.40	51.10	99.40	99.10	94.30
Qwen3-VL-8B	98.40	93.30	88.30	98.80	98.50	93.10
InternVL3-8B	93.78	64.72	55.71	96.45	72.84	87.69

D.2 Finding 1 Across Different Datasets

To assess whether Finding 1 extends across visual distributions and question formats, we evaluate positional-encoding perturbations and attention-flow knockout on SpinBench-ORG and What's Up-COCO using Qwen2.5-VL-3B-Instruct.

Positional-encoding perturbations. Across all four spatial-relation subsets, Table 6 shows that each PE perturbation causes a larger accuracy drop in the vision encoder (VE) than in the LLM. This consistent asymmetry supports the role of position-aware visual encoding in spatial relation grounding across datasets.

Table 6: Accuracy (%) under positional-encoding perturbations for Qwen2.5-VL-3B-Instruct on SpinBench-ORG and What's Up-COCO. LR, FN, and AB denote Left–Right, Far–Near, and Above–Below, respectively.

		SpinBench-ORG		What's Up-COCO	
Module	PE perturbation	LR	FN	LR	AB
Full VLM	None	100.00	100.00	100.00	100.00
VE	Remove	61.97	62.60	63.14	5.17
VE	Shuffle XY	65.73	51.91	62.29	25.86
VE	Shuffle X	63.38	62.60	62.71	63.79
VE	Shuffle Y	74.65	55.73	69.92	32.76
LLM	Remove	99.06	96.95	100.00	91.38
LLM	Shuffle XY	99.53	96.95	98.73	87.93
LLM	Shuffle X	99.06	96.95	98.73	100.00
LLM	Shuffle Y	99.06	96.95	99.58	94.83

Attention-flow knockout. Table 7 reveals a consistent ordering of VE knockout effects on both datasets: background–background removal produces the largest accuracy drop, followed by object–background removal, while direct object–object removal has the smallest effect. Background–background knockout also reduces accuracy in the LLM, particularly on What's Up-COCO, but less severely than in the VE. Together, these results extend the sensitivity to VE background-token interactions across datasets, while showing that interactions among background tokens within the LLM also contribute to spatial predictions.

Table 7: Accuracy (%) under attention-flow knockout for Qwen2.5-VL-3B-Instruct on SpinBench-ORG and What's Up-COCO. Columns specify the attention flow removed in the vision encoder (VE) or LLM. O_1 and O_2 denote the queried objects, and B denotes background tokens.

		Vision encoder			LLM		
Dataset	Baseline	$O_1 \leftrightarrow O_2$	$\{O_1, O_2\} \leftrightarrow B$	$B \leftrightarrow B$	$O_1 \leftrightarrow O_2$	$\{O_1, O_2\} \leftrightarrow B$	$B \leftrightarrow B$
SpinBench-ORG	100.00	97.37	88.30	63.16	97.37	97.08	91.23
What's Up-COCO	100.00	98.97	92.10	57.04	97.23	93.13	71.48

D.3 Influence of token order

Table 8: Image token order perturbation results of Qwen2.5-VL-3B on the CLEVR-Spatial dataset. Token order is perturbed at the vision-encoder (VE) input or at the LLM input.

Module	Mode	Left-Right		Front-Behind	
		Accuracy	Score	Accuracy	Score
VLM		100.00	98.10	100.00	99.01
VE Input	Shuffle XY	52.93	45.59	48.37	43.02
	Shuffle X	62.87	58.09	90.76	68.92
	Shuffle Y	34.24	37.45	8.61	27.51
LLM Input	Shuffle XY	99.92	96.43	95.65	91.29
	Shuffle X	99.92	97.87	100.00	98.90
	Shuffle Y	99.49	95.46	95.02	87.15

We apply *token order perturbations* at multiple points along the visual pipeline, including the LLM input (image tokens) and the Vision Encoder input. We consider structured perturbations that progressively destroy 2D geometry, such as global shuffling and directional reversals. The results are shown in Table 8. When token order perturbations are applied at the LLM input stage, model performance remains nearly unchanged, even under severe disruptions such as global shuffling, directional reversal, and additional embedding noise. The LLM is largely insensitive to the ordering of image tokens. In contrast, applying the same perturbations at the Vision Encoder input leads to a dramatic performance collapse. Disrupting the spatial arrangement of patches reduces accuracy to near-random levels, with the most severe degradation observed under row-wise reversal (left–right flip), which directly interferes with the task-relevant spatial dimension.

Together, these results reveal a clear structural dependency: spatial information is tightly coupled to the geometry of visual representations, but largely decoupled from token ordering at the LLM stage. Our results suggest that the Vision Encoder constructs a geometry-aware embedding space in which spatial relations are already encoded, while the LLM reads out these representations with little reliance on image-token order.

E Additional Experiments for Sec. 4

E.1 Finding 2 Across Different VLMs

We extend the analysis to Qwen2.5-VL-3B and Qwen3-VL-4B on CLEVR-Spatial, evaluating Left–Right across all three models and Front–Behind additionally for Qwen2.5-VL-3B. Across all evaluated settings, independently learned background and target head sets achieve higher accuracy with access to their matched visual source than to the mismatched source, while remaining close to or matching full-model accuracy. The selected head sets vary in size across models: Qwen2.5-VL-3B exhibits the strongest sparsity, whereas InternVL3-8B retains a larger fraction of heads. These results support source-specific visual access across model families, with model-dependent degrees of sparsity.

Table 9: Source-specific head analysis across VLMs on CLEVR-Spatial. Accuracy is reported in percent. The specificity gap is the accuracy under the matched visual source minus that under the mismatched source.

Model	Task	Head Type	Full Acc.	Background Only Acc.	Target Only Acc.	Specificity Gap	#Heads (%)
Qwen2.5-VL-3B	Left–Right	Background	100.00	**100.00**	80.67	19.33	10 (1.95%)
	Left–Right	Target	100.00	88.24	**98.32**	10.08	38 (7.42%)
	Front–Behind	Background	100.00	**100.00**	81.08	18.92	11 (2.15%)
	Front–Behind	Target	100.00	82.88	**100.00**	17.12	24 (4.69%)
Qwen3-VL-4B	Left–Right	Background	98.32	**100.00**	84.03	15.97	41 (14.24%)
	Left–Right	Target	98.32	94.96	**97.48**	2.52	46 (15.97%)

E.2 Finding 2 Across Datasets

Across all three datasets, fewer than 3.5% of attention heads in Qwen2.5-VL-3B-Instruct recover most or all of the accuracy achieved by the full background reasoning path. These selected heads consistently achieve higher accuracy when accessing background tokens than when restricted to target tokens, supporting their source-specific role across datasets.

Table 10: Source-specific head analysis across CLEVR-Spatial, SpinBench-ORG, and What'sUp-COCO.

Dataset	Head Type	Full Acc.	Background Only Acc.	Target Only Acc.	Specificity Gap	#Heads (%)
CLEVR-Spatial	Background Heads	100.00	**100.00**	80.67	19.33	10 (1.74%)
CLEVR-Spatial	Target Heads	100.00	87.39	**98.32**	10.92	38 (6.60%)
SpinBench-ORG	Background Heads	88.37	**86.05**	65.12	20.93	9 (1.56%)
SpinBench-ORG	Target Heads	88.37	74.42	**93.02**	18.60	33 (5.73%)
What'sUp-COCO	Background Heads	70.45	**72.73**	50.00	22.73	20 (3.47%)
What'sUp-COCO	Target Heads	70.45	50.00	**72.73**	22.73	19 (3.30%)

E.3 Additional Necessity-Based Spatial Head Analysis of Qwen and SpaceQwen

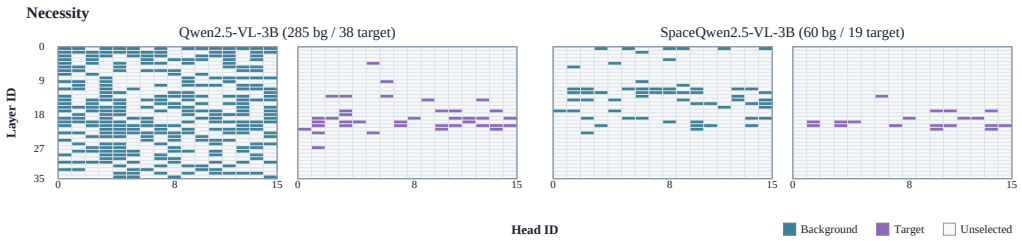


Figure 5: Necessity-based head identification results on the CLEVR-Spatial Left–Right task for Qwen2.5-VL-3B-Instruct and SpaceQwen2.5-VL-3B-Instruct. Background heads and target-object heads exhibit distinct spatial distributions across layers and attention heads.

Necessity Loss. To measure *reliance* on background information, we evaluate the prediction drop after removing access to background tokens:

$$\mathcal{L}_{\text{remove-bg}} = -\mathbb{E}_{(x,y^*)\sim\mathcal{D}}\left[\log P(y^* \mid x) - \log P(y^* \mid x^{tg}_{\mathcal{H}})\right]. \tag{3}$$

A smaller (more negative) $\mathcal{L}_{\text{remove-bg}}$ value indicates that the selected heads depend more strongly on background information, since masking their access causes a larger confidence reduction.

In addition to the main experiments presented in §4, we also report the results obtained using the necessity loss in Equation 3 on both Qwen2.5-VL-3B-Instruct and SpaceQwen2.5-VL-3B-Instruct (Jia et al., 2026) under the CLEVR-Spatial Left-Right task (Figure 5).

The two models exhibit markedly different reliance structures over background and target-object information. In Qwen2.5-VL-3B-Instruct, we could not find a sparse background-necessity set. Even after blocking background access for 285/576 heads (49.5%), the model still maintains 100.0% accuracy, although its prediction confidence drops substantially ($\Delta \log p = 11.34$). By contrast, target-object information is concentrated in a much smaller and more necessary subset: blocking target-object access for only 38/576 heads (6.6%) reduces accuracy to 54.6%.

SpaceQwen2.5-VL-3B-Instruct shows a qualitatively different pattern. Both sources become supported by compact but accuracy-critical head subsets: the background-necessity set contains only 60/576 heads (10.4%), and the target-object set contains only 19/576 heads (3.3%). Removing the corresponding source access from either subset sharply degrades accuracy, to 57.1% for background and 55.5% for target objects. This suggests that SpaceQwen reorganizes spatial reliance so that both background and target-object pathways become sparse and functionally necessary. Together, these results indicate that SpaceQwen compresses spatial reasoning into smaller and more decisive head circuits, preserving the core target-object pathway while substantially reshaping background-dependent processing.

F Additional Experiments for Sec. 5

We evaluate Full KV, our method, IVC-Prune, and Nüwa on 1,500 MMStar (Chen et al., 2024b) and 2,374 MME (Fu et al., 2026) questions using Qwen3-VL-8B-Instruct. All methods use seed 42, greedy decoding, and a maximum of 64 generated tokens. Our method retains the same CLEVR-derived mask of 35 layer–KV-head pairs and compact budget of $K = 8$ visual tokens, without benchmark-specific tuning.

Table 11: **General visual understanding and inference efficiency.** Accuracy (%) uses rule-based answer matching. Latency is measured in seconds per example and throughput in generated tokens per second. Avg. Output Tokens denotes mean generated response length.

Benchmark	**Method**	**Acc.** ↑	**E2E** ↓	**Prefill** ↓	**Decode** ↓	**Throughput** ↑	**Tokens** (Avg.)
MMStar	Full KV	52.40	.7945	.0954	.6991	23.40	17.38
	IVC-Prune	52.33	1.2694	.1869	1.0825	16.32	17.28
	Nüwa	**54.60**	.7762	**.0783**	.6979	**28.69**	21.02
	Ours	53.20	**.6586**	.0968	**.5618**	28.09	16.80
MME	Full KV	**89.68**	.5777	.4925	.0852	23.41	3.00
	IVC-Prune	89.64	.8986	.7483	.1503	18.60	3.02
	Nüwa	89.39	.6186	.5533	**.0653**	**28.65**	2.88
	Ours	89.43	**.5057**	**.4366**	.0691	28.61	2.99

Our method achieves the lowest end-to-end latency on both benchmarks, reducing it by 17.1% on MMStar and 12.5% on MME relative to Full KV. Accuracy increases by 0.80 percentage points on MMStar and decreases by 0.25 points on MME. Nüwa achieves the highest MMStar accuracy and the lowest MME decode latency, but its higher MME prefill latency offsets the decoding savings. IVC-Prune maintains similar accuracy to Full KV but is slower under its reference implementation. These results extend the evaluation to general visual tasks beyond spatial tasks, showing comparable general visual accuracy with lower end-to-end latency using the fixed spatially derived mask.