ATTENTION IN SPACE: FUNCTIONAL ROLES OF VLM HEADS FOR SPATIAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite remarkable advances in large Vision-Language Models (VLMs), spatial reasoning remains a persistent challenge. In this work, we investigate how VLMs represent and perform spatial reasoning through a mechanistic interpretability lens. We introduce CogVSR, a structured diagnostic framework that decomposes complex spatial reasoning tasks into compositional, step-by-step subquestions designed to simulate human-like reasoning, each aligned with specific cognitive functions such as spatial perception. This enables fine-grained analysis of how spatial reasoning is organized within model internals. Building on CogVSR, we develop a probing framework to identify and characterize attention heads specialized for these functions. Our analysis across diverse VLM families reveals that these functional heads are universally sparse, vary in number and distribution across functions. Interestingly, spatially specialized heads are fewer than those for other cognitive functions, highlighting their scarcity. We propose methods to better disentangle spatial-related functions to improve spatial understanding, which is consistent with the activation of latent spatial heads. Intervention experiments further demonstrate their critical role in spatial reasoning: removing functional heads leads to performance degradation, while emphasizing them enhances accuracy. This study provides new interpretability driven insights into how VLMs attend to space and paves the way for enhancing complex spatial reasoning in multimodal models. CogVSR and code will be publicly available upon acceptance.

1 INTRODUCTION

Spatial reasoning Yang et al. (2025b); Zhang et al. (2025b); Yin et al. (2025), the ability to understand and reason about spatial relationships between objects, is a fundamental aspect of human cognition. It underlies our capacity to interpret scenes, describe object relations, and perform actions in complex environments. Despite remarkable progress in multimodal understanding, Vision-Language Models (VLMs) (Zhu et al., 2023; Liu et al., 2023; Lu et al., 2024) still struggle with spatial reasoning, even in simple orientation problems such as "Is the dog facing the horse?"

Consider how humans can readily answer the question shown in Fig. 1 by integrating perception and reasoning processes distributed across multiple brain systems (Barsalou, 2014). As a simplified neuroscience-inspired analogy, visual perception has often been associated with occipital and occipitotemporal processing pathways involved in scene understanding and object recognition (Kravitz et al., 2013), while spatial reasoning has been linked to broader parietal and frontal systems supporting relational processing and decision-making (Krawczyk, 2012; Qi et al., 2025). Although contemporary neuroscience emphasizes distributed and interactive processing rather than strictly localized functions, this analogy provides an intuitive motivation for studying whether functional components may also emerge within VLMs.

To understand the "space" in VLMs, we need to examine their internal mechanisms, particularly the attention heads, which form the core of multimodal representation and alignment. Recent studies (Kang et al., 2025; Bi et al., 2025) have identified sparse attention heads with special functional roles for visual grounding. However, the extent to which such specialization supports complex spatial reasoning that requires coordination across multiple functions remains largely unexplored. Bridging this gap is essential not only for interpretability but also for improving models' spatial reasoning capabilities in a cognitively informed manner.

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.

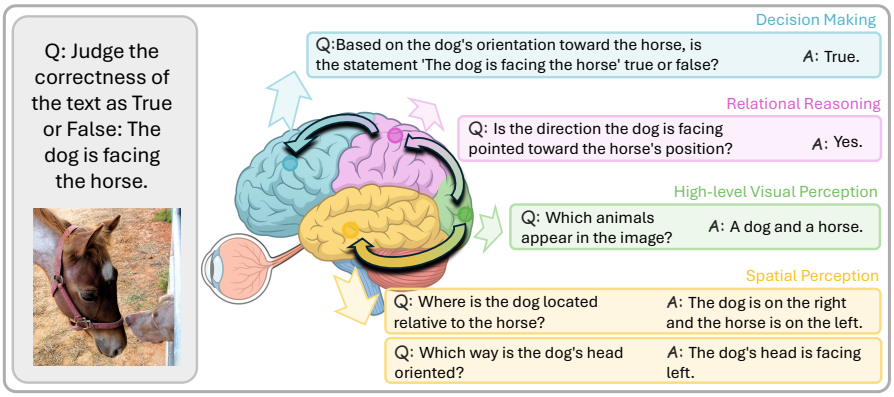


Figure 1: To solve a complex question, the human brain engages multiple regions to perform cognitive functions necessary for generating a response. We explore whether there are specific attention heads in VLMs that play functional roles in producing answers.

To facilitate this, we introduce CogVSR, a benchmark designed to decompose spatial reasoning into human cognition-inspired, interpretable components. CogVSR breaks complex spatial reasoning tasks into structured step-by-step subquestions, each designed to target specific functional processes such as high-level visual perception, spatial perception, and relational reasoning, enabling fine-grained analysis of how spatial reasoning is represented within VLMs.

Building upon CogVSR, we propose a probing framework that systematically identifies and characterizes attention heads associated with distinct cognitive functions. Through analysis on three major VLM families, including Intern (Zhu et al., 2025), Qwen (Yang et al., 2025a), and Llama (Lin et al., 2023) with different model scales, our results reveal the existence of cognitive heads, attention heads associated with specific cognitive functions, that consistently exhibit **universal**, **sparse**, and **intrinsic** properties across architectures. We validate the functional importance of these heads by showing that their removal degrades performance on related tasks. Further analysis shows that spatial heads, including those for spatial perception and relational reasoning, are relatively scarce and often exhibit weaker significance compared to functions such as Information Extraction and Understanding. This **limited prevalence of spatially specialized functions** underscores the constrained spatial reasoning capacity in current VLMs. To address this, we propose a Spatial Head Activation (SHA) approach, which is associated with the activation of latent spatial heads and yields over 10% accuracy improvement on InternVL3-2B across both spatial perception and relational reasoning tasks. Furthermore, we validate the importance of the identified cognitive heads on downstream spatial reasoning tasks and show that positively intervening in these functional heads can enhance the model's spatial reasoning capabilities. In summary, our contributions are as follows:

- We introduce CogVSR, a cognitively grounded benchmark for decomposing spatial reasoning into interpretable sub-processes.
- We develop an interpretability framework for uncovering the functional roles of attention heads in spatial reasoning.
- We analyze the properties of cognitive heads and develop methods to activate latent spatial heads, along with intervention strategies to enhance spatial understanding.

2 Related Works

Spatial reasoning, the ability to interpret and reason about object relationships in space, remains a core challenge for VLMs. Despite strong performance in image classification Pratt et al. (2023), captioning Alaluf et al. (2024), and object detection Kuo et al. (2022), recent studies Liu et al. (2023); Chen et al. (2025); Yang et al. (2025b) show that VLMs still struggle to capture geometric and positional relations, often failing even simple tasks such as identifying whether one object is under, above, or behind another. To address these issues, several studies have explored specialized prompting strategies Li et al. (2024); Zhang et al. (2025a); Chen et al. (2023), post-training approaches such as supervised fine-tuning Li et al. (2025); Ranasinghe et al. (2024) and reinforcement learning (RL)

Table 1: An example from the **CogVSR** dataset, illustrating a main question, its final answer, and a breakdown into subquestions with corresponding answers and multi-label cognitive function annotations.

Main / Image	Subquestion	Answer	Cognitive Function Label
 **Main Question:** Judge whether the text is true or false: "The dog is facing the horse." **Main Answer:** True.	1. Which animals appear in the image?	A dog and a horse.	High-level Visual Perception
	2. Where is the dog located relative to the horse?	The dog is on the right and the horse is on the left.	**Spatial Perception, High-level Visual Perception**
	3. Which way is the dog's head oriented?	The dog's head is facing left.	**Spatial Perception, High-level Visual Perception**
	4. Is the direction the dog is facing pointed toward the horse's position?	Yes.	Relational Reasoning
	5. Based on the dog's orientation toward the horse, is the statement "The dog is facing the horse" true or false?	True.	Decision Making

Feng et al., as well as architectural modifications to VLMs Ma et al. (2025b); He et al. (2025); Cheng et al. (2024); Zhang et al. (2024), mainly by enhancing input representations or redesigning spatial reasoning modules. Recent studies Qi et al. (2025); Wen et al. (2024); Chen et al. (2025); Zhang et al. (2025c) have also analyzed why MLLMs struggle with spatial reasoning, with broader discussions in the survey by Zheng et al. (2025). While these methods have improved performance on certain spatial benchmarks and identified issues such as attention bias and insufficient geometric information, they primarily focus on simple spatial relation tasks and analyze attention at the token level. In contrast, our work provides a deeper explanation of VLMs' behavior in complex, multi-step spatial reasoning and investigates the functional roles of internal components, i.e., attention heads, through a cognition-driven perspective. Additional related work on the functional roles of attention heads is provided in Appendix A.

3 PRELIMINARIES

Large VLMs, such as Qwen (Yang et al., 2025a), typically consist of three components: a visual encoder, a language model, and a multimodal projector that connects them. The visual encoder extracts spatially structured visual embeddings from an image, which are then mapped into the token space of the language model through the projector. The language model, often a transformer with L stacked layers, contains a Multi-Head Attention (MHA) module and a feedforward network in each layer. For layer l, given input $X \in \mathbb{R}^{n\times d}$ (where n is the number of tokens and d the embedding dimension), MHA computes:

$$\text{MHA}^{(l)}(X) = \text{Concat}\big(N_h^{(l,1)}, \ldots, N_h^{(l,H)}\big)W_o, \tag{1}$$

where H is the number of heads, and $N_h^{(l,h)} = \text{Softmax}(\frac{QK^\top}{\sqrt{d_h}} + M)V$. Here $Q = XW_q^{(h)}$, $K = XW_k^{(h)}$, $V = XW_v^{(h)}$, $W_q^{(h)}, W_k^{(h)}, W_v^{(h)} \in \mathbb{R}^{d\times d_h}$ are learnable projection matrices for head h, and M denotes the attention mask matrix. Spatial reasoning in VLMs is supported by attention heads that integrate visual features, geometric relationships, and textual cues for inference. We analyze the functional specialization of these heads to understand their roles in spatial reasoning.

4 COGVSR

In this section, we present Cognitive Vision Spatial Reasoning (CogVSR), a cognitively inspired dataset for spatial reasoning. CogVSR contains 1,142 main questions and 3,759 subquestions. Each example comprises a main question–answer pair, its corresponding subquestions and subanswers, and annotations specifying the cognitive function associated with each subquestion (an example is provided in Table 1).

4.1 COGNITIVE FUNCTIONS

To systematically capture the cognitive processes involved in complex spatial reasoning tasks, we consider eight functions, inspired by established frameworks in cognitive science (Anderson, 2014; Diamond, 2013; Cotabish et al., 2024).

The functions directly related to space include:

- **Spatial Perception**: Comprehending the spatial arrangement, positions, and geometric relationships among objects within a scene.
- **Relational Reasoning**: Understanding and comparing relationships among entities, with an emphasis on structural patterns.

Other functions related to spatial reasoning include:

- **Low-level Visual Perception**: Recognizing basic visual features such as color, shape, size.
- **High-level Visual Perception**: Integrating visual information to recognize objects, patterns, and scene structure at a semantic level.
- **Language Information Extraction and Understanding**: locating and understanding relevant information from an external source or prior context.
- **Knowledge Recall**: Accessing long-term knowledge including language and vision.
- **Math Reasoning**: Performing counting, arithmetic, comparison, and logic-based operations.
- **Decision-Making**: Selecting the best outcome among alternatives based on reasoning.

This categorization reflects a natural progression from basic information processing to complex cognitive integration. Both the human brain and VLMs encompass a wide range of functional modules. Our focus in this work is specifically on spatial reasoning-related cognitive functions. By identifying and organizing these eight core functions, we can more clearly examine how VLMs handle different types of thinking steps, in a way that is both systematic and easy to interpret. The detail descriptions of each function and more examples in CogVSR can be found in the Appendix B.

4.2 DATA COLLECTIONS

Based on our categorization of reasoning functions, we sampled 1600 diverse questions from existing commonly used visual reasoning benchmarks, selecting 400 examples each from VSR Liu et al. (2022), SpatialEval Wang et al. (2024a), 3DSRbench Ma et al. (2025a), Spatial457 Wang et al. (2025) dataset. These datasets cover a broad spectrum of spatial reasoning tasks, ranging from simple position and orientation judgments to more complex challenges such as occlusion, counting, path planning, and maze navigation. Using the chain-of-thought (CoT) paradigm, we prompted GPT-o4-mini (OpenAI, 2025) to decompose each main question (with its final answer) into subquestions, each targeting a subanswer and a specific function. The prompt encourages structured, step-by-step reasoning, ensuring that each subquestion is clear, answerable, and sequentially dependent. This process yields a set of subquestion–answer–function (subQAF) triples for each QA pair: $\mathrm{subQAFs} = \{(q_i, a_i, f_i)\}_{i=1}^{k}$, where each contains a subquestion q_i, its concise subanswer a_i, and the corresponding function label f_i. The prompt for generating CoT, subquestions, and subanswers are list in Appendix C.

4.3 TOWARDS HUMAN-ALIGNED SPATIAL REASONING ANNOTATION

Recent advances enable leveraging large pre-trained models for dataset construction, thanks to their reasoning capabilities and ability to generate high-quality annotations at scale (Wang et al., 2024b). While our dataset is initially generated using a large VLM to reduce manual effort, the entire annotation pipeline is explicitly designed to align with human spatial reasoning processes. To this end, we employ a rigorous two-stage human verification procedure to ensure data quality and mitigate hallucination.

In the first stage, three expert annotators independently evaluate whether each subquestion is logically structured and the overall reasoning chain is consistent with natural human reasoning, as well as whether each assigned cognitive function is accurate. QA pairs with incoherent, unnatural, or inconsistent decompositions are filtered out. By removing approximately 28.7% of human non-aligned QA pairs, the second stage relabels the cognitive functions of remaining subquestions where

annotator disagreement occurs, ensuring closer alignment with human-interpretable reasoning steps. Notably, although GPT-o4-mini assigns a single primary function to each subquestion, a single subquestion may require the simultaneous engagement of spatial perception (e.g., orientation) and high-level visual perception (e.g., object recognition), as illustrated by subquestion 2 in Table 1. To better align with human reasoning processes in solving such questions, we adopt a multi-label annotation scheme. Simultaneously, subanswers are reviewed and, if necessary, corrected by human annotators (details in Appendix D).

This multi-stage human-in-the-loop verification process ensures that each retained subQAF triplet reflects a coherent and interpretable reasoning step that is grounded in approximations of human cognitive processes, rather than purely synthetic model-generated logic. After this refinement, our final dataset comprises 1,142 main QA pairs and 3,759 validated subQAF triplets. Appendix E includes detailed dataset statistics.

5 Attention Heads Related to Spatial Reasoning

Using the CogVSR dataset, we adopt a probing-based framework (Alain & Bengio, 2016; Belinkov, 2022; Tenney et al., 2019) to identify which attention heads in VLMs are associated with specific functions in the reasoning process. Specifically, for each functionally annotated subquestion, we extract head activations (see Subsection 5.1), train a multi-label classifier and compute importance scores to identify contributing heads (see Subsection 5.2).

5.1 Head Feature Extraction

Given a large VLM $\mathcal{M}$, we generate an answer $a_i^{\mathcal{M}}$ for each subquestion q_i derived from a main question Q_i. To support coherent multi-step reasoning, we include preceding subquestions and their correct answers as contextual input, emulating the incremental reasoning process observed in human cognition. During inference, each input token is first mapped into an embedding and then propagated through the transformer's layers. At each layer, attention and feedforward operations update the residual stream, which is ultimately decoded into token predictions. For each generated token i, we extract attention head outputs $X_i = \{x_l^m(i) \mid l = 1, \ldots, L,\ m = 1, \ldots, M\}$ across all layers, where x_l^m denotes the value vector from the m-th head in layer l, with M the number of heads per layer and L the total number of layers.

Let N_t denote the number of tokens in the generated answer $a_i^{\mathcal{M}}$. To isolate semantically informative content relevant to reasoning, we select the top-k most important tokens, determined by prompting Gemini2.5-Flash (Comanici et al., 2025), yielding an index set $\mathcal{I}_k$ with $|\mathcal{I}_k| = k$. For each index $j \in \mathcal{I}_k$, we extract the corresponding attention head activations from X_j, and compute the averaged activation feature for the m-th head in layer l as $\bar{x}_l^m = \frac{1}{k}\sum_{j\in\mathcal{I}_k} x_l^m(j)$. Given prior findings suggesting that cognitive functions may vary by layer depth (Zheng et al.), we incorporate layer-wise information by computing the average activation $\bar{x}_l = \frac{1}{M}\sum_{m=1}^{M} \bar{x}_l^m$ for each layer. We then augment each head-level vector with its corresponding layer summary, resulting in enriched features $\bar{x}_l^{m'} = [\bar{x}_l^m; \bar{x}_l]$. The top-$k$ selection prompt and example appear in Appendices C and G, respectively. An ablation study in Appendix G examines the effects of token selection (first, last, or all tokens) and layer information.

5.2 Function Probing

For a dataset with N subQAF triplets, we construct the probing dataset using the corresponding augmented activations.

$$\mathcal{D}_{\text{probe}} = \left\{(\bar{x}_l^{m'},\ c)_i\right\}_{i=1}^{N}, l \in \{1, \ldots, L\},\ m \in \{1, \ldots, M\} \tag{2}$$

For classification based on CogVSR, the training set includes 909 main questions with 2,969 subQAF triplets, while the testing set has 233 main questions with 790 triplet. Notably, we use the attention-head output values only when the VLM $\mathcal{M}$ generates the correct answer for the subquestions, as incorrect responses may be associated with noisy or non-informative head activations. Each attention

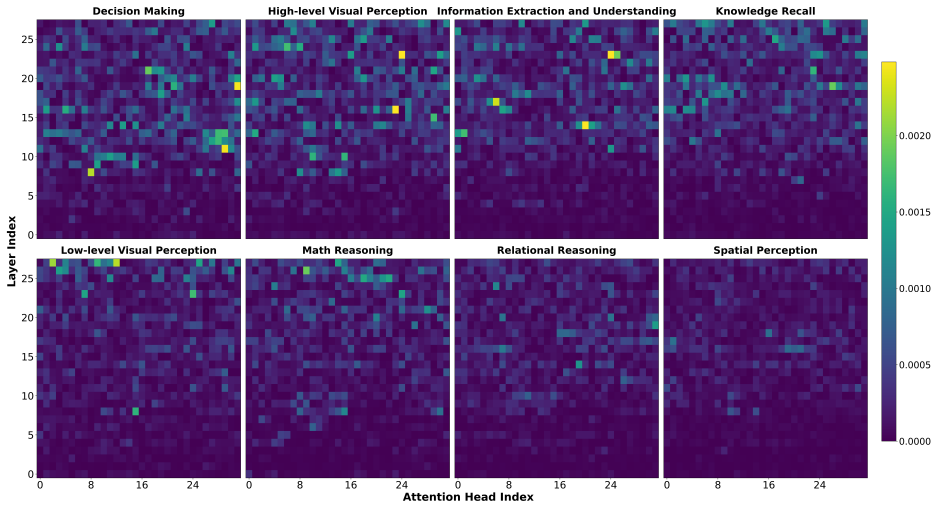


Figure 2: The existence of cognitive heads in Qwen2.5-VL-7B responsible for eight distinct functions in complex reasoning tasks. The x-axis represents the head index, while the y-axis indicates the layer index. The values denote head importance scores.

head feature is first passed through a trainable linear projection for dimensionality reduction, after which all projected features are concatenated into a single vector and fed into a two-layer MLP for multi-label classification over cognitive functions (training details are provided in Appendix H). To interpret the contribution of individual heads to each function, we use a gradient-based attribution method. Specifically, for each function class c, we compute the contribution of each head feature via the gradient×activation technique: $I_j^{(c)} = \mathbb{E}_{(\bar{x},c)\sim\mathcal{D}_{\text{probe}}}\left[\frac{\partial \hat{y}_c}{\partial \bar{x}_j} \cdot \bar{x}_j\right]$, where $\bar{x}_j$ is the j-th head input feature, and $\hat{y}_c$ is the classifier's predicted logit for class c. This yields an importance score for each attention head with respect to each cognitive function. We aggregate the scores into a matrix $\mathbf{I} \in \mathbb{R}^{C\times(L\cdot M)}$, where each row corresponds to a function class and each column to a specific head in a specific layer.

We hypothesize that attention heads with higher importance scores contribute more significantly to each cognitive function. By ranking heads according to their importance, we can identify which heads and layers are specialized for specific functions. Subsequent targeted interventions on these heads validate the effectiveness of this approach. We compare MLP probing with logistic regression in Appendix H to demonstrate the advantages of the proposed method.

6 Experiments

We conduct a series of experiments on three VLM families across various model scales, including Intern (Zhu et al., 2025) (InternVL3-2B and InternVL3-8B), Qwen (Yang et al., 2025a) (Qwen2.5-VL-7B and Qwen2.5-VL-3B), and Llama3.2-Vision (Liu et al., 2025) (Llama-3.2-11B-Vision and Llama-3.2-90B-Vision). We analyze the common properties of functional heads (Subsection 6.1) and validate their contributions (Subsection 6.2). We further investigate the limited prevalence of spatial heads (Subsection 6.3), which likely contributes to the failure of VLMs in spatial reasoning. We study input-induced activation patterns of latent spatial heads, leading to improved performance. Additionally, we assess the causal impact of cognitive heads on downstream reasoning tasks (Subsection 6.5); our results show that modulating spatially related heads can improve spatial reasoning.

6.1 Properties of Cognitive Heads

Sparsity, Universality, and Intrinsic Organization: Figure 2 shows the heatmap of head importance scores across eight functions in Qwen2.5-VL-7B on the CogVSR test set, revealing a sparse distribution. In total, fewer than 9% of all heads have importance scores above 0.001 across the eight functions (about 3% for high-level visual perception and decision making, and less than 1% for the others), suggesting that only a compact subset of heads meaningfully contribute to task performance.

Table 2: Intervention results (%) of cognitive heads vs. random heads across 8 functions: **Spatial** Perception, **Relational** Reasoning, **Low-level** Visual Perception, **High-level** Visual Perception, Language **Info**rmation Extraction and Understanding, Knowledge **Recall**, **Math** Reasoning, and **Decision** Making.

Model	Inter_Head	Spatial Functions				Other Functions											
		Spatial		Relational		Low-Level		High-Level		Info		Recall		Math		Decision	
		llm	acc	llm	acc	llm	acc	llm	acc	llm	acc	llm	acc	llm	acc	llm	acc
Intern2B	random	61.37	56.82	63.25	58.13	65.48	60.91	69.72	64.35	71.19	66.04	69.87	64.11	59.44	54.27	64.03	58.76
	cognitive	**37.92**	**34.37**	**31.56**	**26.88**	**13.43**	**18.21**	**34.27**	**28.65**	**39.82**	**33.17**	**37.11**	**31.49**	**23.68**	**18.02**	**21.54**	**16.43**
Intern8B	random	72.84	68.12	74.33	69.57	76.05	71.44	81.26	76.09	83.17	78.03	81.92	76.48	73.66	68.40	77.41	72.05
	cognitive	**46.39**	**42.07**	**52.84**	**48.51**	**36.12**	**31.76**	**59.73**	**53.28**	**65.04**	**58.69**	**62.91**	**56.33**	**49.82**	**44.15**	**46.37**	**40.88**
Llama11B	random	74.95	70.44	76.68	72.11	78.37	73.90	83.29	78.16	85.06	79.87	83.64	78.33	75.92	70.83	79.18	74.06
	cognitive	**44.73**	**40.58**	**50.41**	**56.33**	**34.86**	**30.17**	**57.24**	**51.63**	**62.77**	**56.94**	**60.58**	**54.72**	**47.39**	**41.92**	**43.85**	**38.14**
Llama90B	random	86.21	82.07	88.03	83.92	89.77	85.63	93.18	89.02	94.56	90.41	93.02	88.93	87.49	83.36	91.07	86.95
	cognitive	**72.44**	**68.26**	**76.71**	**72.39**	**60.03**	**55.89**	**74.82**	**69.44**	**79.15**	**73.86**	**77.03**	**71.74**	**69.42**	**64.13**	**66.85**	**61.27**
Qwen3B	random	69.12	64.88	71.03	66.74	73.49	69.05	77.85	73.12	79.64	75.02	78.03	73.41	69.77	65.23	73.54	69.01
	cognitive	**35.63**	**42.94**	**31.28**	**47.46**	**14.92**	**10.08**	**38.47**	**32.80**	**44.35**	**38.06**	**41.72**	**35.94**	**27.58**	**22.16**	**25.03**	**19.77**
Qwen7B	random	80.04	75.92	82.11	77.86	84.39	80.21	88.27	84.10	90.02	85.73	88.39	84.03	82.76	78.44	86.51	82.19
	cognitive	**57.69**	**53.42**	**63.85**	**59.57**	**47.92**	**43.18**	**68.31**	**63.04**	**73.77**	**68.55**	**71.23**	**65.98**	**62.84**	**57.63**	**59.41**	**54.22**

These results demonstrate that VLMs rely on highly specialized, localized components for distinct cognitive abilities. Moreover, this sparse functional organization is consistent across architectures and scales: heatmaps for five additional models (in Appendix F) confirm its universality. Within the same model family (e.g., Qwen2.5-VL-7B vs. Qwen2.5-VL-3B), we observe similar distributions, suggesting that such specialization is intrinsic to VLMs.

6.2 Functional Contributions of Cognitive Heads

We define the heads with high importance scores for each function as cognitive heads. Specifically, we identify cognitive heads by selecting those before the elbow point of each function's descending importance curve (see Appendix K), and find notable variation in head counts across functions.

After identifying the cognitive heads associated with each function, we examine their functional roles by evaluating the model's behavior on the CogVSR test set under targeted interventions. We perform head ablation by scaling the output of a specific attention head with a small factor ϵ (e.g., 0.001), effectively suppressing its contribution: $x_i^{\text{mask}} = \epsilon \cdot \text{Softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i$. Specifically, we compare model performance when masking identified cognitive heads versus masking an equal number of randomly-selected heads. To quantify the impact, we employ both an LLM-based judge (Gemini2.5-flash-lite Comanici et al. (2025)) and an integrated accuracy metric. An output is considered unaffected if its BLEU score (Papineni et al., 2002) exceeds 0.8, or if either the ROUGE score (Chin-Yew, 2004) or the semantic similarity score surpasses 0.6. This provides a comprehensive evaluation of performance degradation.

Table 3: Effect of spatial input augmentation on spatial and relational functions. "Original" uses the raw image; "+BBox" and "+Mask" add bounding-box and instance-mask cues, respectively.

Model	Input	Spatial		Relational	
		llm	acc	llm	acc
Intern2B	Original	61.37	56.82	63.25	58.13
	+BBox	65.84	58.94	64.95	62.39
	+Mask	59.70	59.80	66.82	56.47
	+Color	64.92	60.35	65.74	61.28
	+Shape	68.45	64.91	68.83	65.72
	+Color+Shape	69.36	66.27	69.21	66.18
	+BBox+Mask	**71.82**	**68.64**	**70.12**	**67.04**
Llama90B	Original	86.21	82.07	88.03	83.92
	+BBox	90.07	86.93	89.84	85.75
	+Mask	87.57	83.38	91.22	85.11
	+Color	88.96	84.85	89.71	85.64
	+Shape	91.48	87.12	91.76	88.34
	+Color+Shape	92.02	87.35	92.08	88.66
	+BBox+Mask	**92.90**	**87.77**	**92.41**	**89.27**

As shown in Table 2, masking cognitive heads leads to a substantial decline in performance, whereas masking an equal number of random heads results in only minor degradation across all VLMs. In some cases, masking the identified cognitive heads reduces accuracy to below 20%, indicating that the model can almost no longer perform the corresponding function without them. We further investigate the LLM performance under different numbers of masked heads in Appendix J. Masking functional heads consistently causes greater degradation than masking the same number of random heads, with the gap peaking near the elbow points, supporting their sparsity and functional importance.

6.3 Non-significant of Spatially-Related Heads

By examining spatially related attention heads, we find that, within the generally sparse set of functional heads in VLMs, spatial and relational reasoning heads are relatively fewer than those associated with other functions, with only a small subset showing high importance scores. These heads appear to encode information such as object positions, orientations, geometric layouts, and object-to-object relations, and play an important role in spatial reasoning tasks. Their limited presence may be associated with weaker performance on spatial reasoning, suggesting a possible imbalance in how spatial computation is represented in current VLMs.

6.4 Input-Induced Activation for Spatial Reasoning

How can we encourage VLMs focusing on spatial? We propose a Spatial Head Activation (SHA) approach that disentangles high-level visual perception (e.g., object recognition) from spatial perception. Concretely, we use Gemini2.5-Flash Balestri (2025) to detect objects and obtain bounding boxes, and then apply masking to the detected regions to provide additional object priors while reducing the model's reliance on high-level visual cues (see Fig. 3). Prompt details are provided in Appendix I.

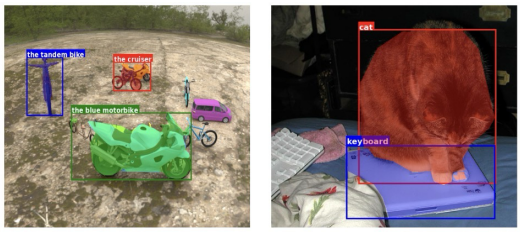


Figure 3: Objects with applied bounding boxes and masks as input.

Table 3 shows that InternVL3-2B and Llama3.2-90B-Vision achieve about 10% and 5% accuracy improvements with "+BBox+Mask", respectively, on both spatial perception and relational reasoning tasks when applying SHA. At the mechanistic level, we further analyze head activation patterns. As illustrated in Fig. 4, after applying SHA, Qwen2.5-VL-3B and Llama3.2-90B-Vision exhibit an increased number of spatial and relational reasoning heads with higher importance scores.

We further explore different input induction strategies to encourage the model to more effectively engage spatially oriented capabilities. Specifically, we highlight object instances using color cues ("+Color") and replace semantic objects with simple geometric shapes ("+Shape"). We observe consistent improvements in spatial perception and relational reasoning tasks, as well as increased activation of spatially relevant heads. The Pearson correlation analysis in Appendix L further indicates that SHA approach can stimulate stronger activation of intrinsic spatial functional heads, while preserving stable activation patterns across different perturbations.

Table 4: Negative intervention on visual spatial reasoning benchmarks. Scores are LLM-Judge accuracy (%). "before" uses the original model; "after" applies intervention.

Dataset	Inter_Head	Intern8B	Qwen7B	Llama11B
VSR	before	64.35	62.47	65.82
	after	**24.13**	**21.76**	**17.94**
Spatial457	before	41.23	38.71	32.18
	after	**14.97**	**12.43**	**9.86**
SpatialEval	before	47.92	45.33	39.61
	after	**20.08**	**18.57**	**15.19**
3DSRBench	before	36.58	34.16	28.87
	after	**12.39**	**10.84**	**8.72**

Overall, structured input induction improves spatial reasoning while consistently engaging spatially relevant attention heads. SHA provides explicit localization cues, potentially reducing the demand on high-level visual perception. Moreover, different induction strategies produce highly consistent head-importance patterns (Appendix L), suggesting that they engage a shared internal spatial mechanism. While these observations are correlational, the positive intervention experiments in Section 6.5 provide causal evidence: directly intervening on the identified heads improves spatial reasoning even without modifying the input. Together, these findings suggest that the identified heads capture latent spatial representations that can be selectively engaged to support spatial reasoning.

6.5 Influence of Functional Heads on Downstream Tasks

We further investigate how functional heads influence downstream tasks through both negative (masking out function heads) and positive interventions (shifting heads toward specific functions).

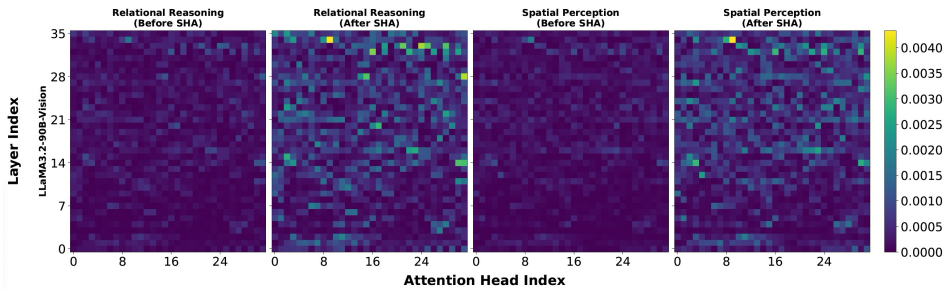


Figure 4: Comparison of spatial perception and relational reasoning heads before and after spatial head activation across various models.

Table 5: Positive intervention on spatial reasoning benchmarks (200 samples per dataset). Scores are LLM-Judge accuracy (%). For out-of-domain data, we positively intervene on spatial perception and relational reasoning heads.

Model	Inter_Head	In Domain								Out of Domain			
		Spatial	Relational	Low	High	Info	Recall	Math	Decision	VSR	Spatial457	SpatialEval	3DSRBench
Qwen7B	before	80.04	82.11	**84.39**	88.27	90.02	88.39	82.76	86.51	62.47	38.71	45.33	34.16
	after	**81.89**	**83.76**	84.10	**90.63**	**92.41**	**90.92**	**82.90**	**87.74**	**63.72**	**39.15**	**45.81**	**34.95**
Llama11B	before	74.95	76.68	78.37	83.29	85.06	83.64	75.92	79.18	65.82	32.18	39.61	28.87
	after	**75.62**	**77.40**	**79.10**	**83.75**	**86.20**	**83.10**	**75.40**	**79.65**	**66.91**	**32.40**	**39.90**	**29.40**
Intern8B	before	72.84	74.33	76.05	81.26	83.17	81.92	**73.66**	77.41	64.35	41.23	47.92	36.58
	after	**74.55**	**74.90**	**77.80**	**81.90**	**83.50**	**82.60**	73.20	**79.89**	**65.28**	**41.85**	**48.45**	**37.20**

Negative Intervention: We randomly sample 200 question for four benchmarks, VSR Liu et al. (2022), Spatial457 Wang et al. (2025), SpatialEval Wang et al. (2024a), and 3DSRBench Ma et al. (2025a), both spatial reasoning tasks. We perform negative intervention by masking spatial perception heads of VLMs, effectively suppressing their activations. As shown in Table 4, masking these key cognitive heads leads to a significant performance drop across different models and datasets.

Positive Intervention: We calculate the activation directions of different functions using the CogVSR dataset. For each function, the activation direction of a head at layer l and index h is computed as: $\text{dir}_l^h = \mathbb{E}_{i\in\mathcal{D}_{\text{correct}}}\left[x_l^h(i)\right] - \mathbb{E}_{i\in\mathcal{D}_{\text{incorrect}}}\left[x_l^h(i)\right]$, where $x_l^h(i)$ denotes the activation of head at layer l and index h, and $\mathcal{D}_{\text{correct}}$ and $\mathcal{D}_{\text{incorrect}}$ represent the sets of samples answered correctly and incorrectly, respectively. Then we estimate the standard deviation of activations (Li et al., 2023) along the cognitive function direction to be σ_l^h, and shift original head activation as $x_l^h(i) \leftarrow x_l^h(i) + \alpha\sigma_l^h \,\text{dir}_l^h$, where α is a parameter. Results in Table 5 show that enhancing the activation of functional heads along their corresponding functional directions improves performance on the related tasks. For example, positive intervention on spatial perception heads in Qwen2.5-VL-7B increased accuracy on the corresponding CogVSR spatial perception subquestions from 80.04% to 81.89%. Similarly, enhancing other functional heads can also boost performance on downstream tasks. Here, we set $\alpha = 0.1$ for all datasets, the hyperparameter analysis is provided in Appendix M.

7 Conclusion

In this work, we investigate how VLMs perform spatial reasoning through attention head specialization. We introduce CogVSR, a cognitively grounded benchmark that decomposes spatial reasoning into interpretable functions. Across multiple VLM families, we identify specialized and localized attention heads associated with spatial reasoning. Targeted activation of latent spatial heads improves spatial understanding without retraining, and intervention experiments further support their functional importance: ablating these heads degrades performance, while emphasizing them improves accuracy. Our findings shed light on the internal organization of VLMs and emphasize the need for more interpretable and cognitively inspired VLMs. Additionally, our eight-function taxonomy and human-verified decompositions may not fully capture human reasoning. Probing only correct responses may introduce selection bias, and head-level analysis does not establish exclusive functional implementation or capture broader interactions with MLPs. We primarily examine cognitive capacity under structured prompting, leaving its utilization during unconstrained reasoning for future work.

AI USE STATEMENT

We used o4-mini (OpenAI, 2025) to assist in constructing CogVSR by decomposing each main question, together with its final answer, into subquestions, each paired with a subanswer and a specific cognitive function. We also used LLM to polish the manuscript and assist with finding related work. The authors take responsibility for the final dataset and manuscript, including all AI-assisted content.

REPRODUCIBILITY STATEMENT

The main text and appendix describe the dataset construction process, proposed method, and experimental setup to facilitate reproduction of our results. Upon acceptance, we will release the CogVSR dataset and source code, including the configurations and scripts needed to reproduce our experiments.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Yuval Alaluf, Elad Richardson, Sergey Tulyakov, Kfir Aberman, and Daniel Cohen-Or. Myvlm: Personalizing vlms for user-specific queries. In *European Conference on Computer Vision*, pp. 73–91. Springer, 2024.

John R Anderson. *Rules of the mind*. Psychology Press, 2014.

Roberto Balestri. Gender and content bias in large language models: a case study on google gemini 2.0 flash experimental. *Frontiers in Artificial Intelligence*, 8:1558696, 2025.

Lawrence W Barsalou. *Cognitive psychology: An overview for cognitive scientists*. Psychology Press, 2014.

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

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Bingjie Wang, and Chenliang Xu. Unveiling visual perception in language models: An attention head analysis approach. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4135–4144, 2025.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *European Conference on Computer Vision*, pp. 565–580. Springer, 2020.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

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. *arXiv preprint arXiv:2503.01773*, 2025.

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.

Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Alicia Cotabish, Debbie Dailey, Jason Trumble, and Rachelle Miller. Spatial reasoning excellence: A synergy of vantassel-baska's integrated curriculum model and talent development. *Education Sciences*, 14(7):716, 2024.

Adele Diamond. Executive functions. *Annual review of psychology*, 64(1):135–168, 2013.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms, 2025. *URL https://arxiv. org/abs/2503.21776*.

Peiqi He, Zhenhao Zhang, Yixiang Zhang, Xiongjun Zhao, and Shaoliang Peng. Spatial-ormllm: Improve spatial relation understanding in the operating room with multimodal large language model. *arXiv preprint arXiv:2508.08199*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9339–9350, 2025.

Dwight J Kravitz, Kadharbatcha S Saleem, Chris I Baker, Leslie G Ungerleider, and Mortimer Mishkin. The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in cognitive sciences*, 17(1):26–49, 2013.

Daniel C Krawczyk. The cognition and neuroscience of relational reasoning. *Brain research*, 1428: 13–23, 2012.

Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022.

Chengzu Li, Caiqi Zhang, Han Zhou, Nigel Collier, Anna Korhonen, and Ivan Vulić. Topviewrs: Vision-language models as top-view spatial reasoners. *arXiv preprint arXiv:2406.02537*, 2024.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8592–8603, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 5265–5275, 2020.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2023.

Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. Visual spatial reasoning. *ArXiv*, abs/2205.00363, 2022.

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

Zhen Liu, Qixu Wang, Hao Lu, and Yifang Wang. Feasibility and usability practice on local hosting open source large language models (llms) including llama 3.2 vision 90b in multi-functional agentic artificial intelligence (ai) system to drive service for design in the latest affordable small personal computer (pc) system. In *International Conference on Human-Computer Interaction*, pp. 261–279. Springer, 2025.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

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, 2025a.

Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 17249–17260, 2025b.

Xueqi Ma, Jun Wang, Yanbei Jiang, Sarah Monazam Erfani, Tongliang Liu, and James Bailey. Cognitive mirrors: Exploring the diverse functional roles of attention heads in llm reasoning. *arXiv preprint arXiv:2512.10978*, 2025c.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

OpenAI. Openai o3 and o4-mini system card. Technical report, OpenAI, 4 2025. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15691–15701, 2023.

Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025.

Kanchana Ranasinghe, Satya Narayan Shukla, Omid Poursaeed, Michael S Ryoo, and Tsung-Yu Lin. Learning to localize objects improves spatial reasoning in visual-llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12977–12987, 2024.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. *arXiv preprint arXiv:1905.05950*, 2019.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024a.

Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Spatial457: A diagnostic benchmark for 6d spatial reasoning of large mutimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24669–24679, 2025.

Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. Human-llm collaborative annotation through effective verification of LLM labels. In Florian 'Floyd' Mueller, Penny Kyburz, Julie R. Williamson, Corina Sas, Max L. Wilson, Phoebe O. Toups Dugas, and Irina Shklovski (eds.), *Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI 2024, Honolulu, HI, USA, May 11-16, 2024*, pp. 303:1–303:21. ACM, 2024b. doi: 10.1145/3613904.3641960. URL https://doi.org/10.1145/3613904.3641960.

Chuan Wen, Dinesh Jayaraman, and Yang Gao. Can transformers capture spatial relations between objects? *arXiv preprint arXiv:2403.00729*, 2024.

Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality, 2024. *URL https://arxiv. org/abs/2404*, 15574.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, et al. Ferret-v2: An improved baseline for referring and grounding with large language models. *arXiv preprint arXiv:2404.07973*, 2024.

Haoyu Zhang, Meng Liu, Zaijing Li, Haokun Wen, Weili Guan, Yaowei Wang, and Liqiang Nie. Spatial understanding from videos: Structured prompts meet simulation data. *arXiv preprint arXiv:2506.03642*, 2025a.

Jiarui Zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. *arXiv preprint arXiv:2502.17422*, 2025b.

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. *arXiv preprint arXiv:2509.02359*, 2025c.

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.

Z Zheng, Y Wang, Y Huang, S Song, M Yang, B Tang, F Xiong, and Z Li. Attention heads of large language models: A survey. arxiv 2024. *arXiv preprint arXiv:2409.03752*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. *arXiv preprint arXiv:2410.13708*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

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.

A More Related Works

Functional Roles of Attention Heads. The transformer architecture, which underpins most modern VLMs, relies on multi-head attention to process and integrate information. A growing body of interpretability research has revealed that attention heads in LLMs exhibit functional specialization, supporting capabilities such as induction, truthfulness, information retrieval, and safety alignment (Olsson et al., 2022; Li et al., 2023; Wu et al.; Zhou et al., 2024; Zheng et al.). In the multimodal domain, recent studies (Li et al., 2020) have begun to uncover the internal mechanisms of VLMs. Evidence shows that certain sparse attention heads play distinct roles in visual grounding, aligning textual tokens with image regions without additional fine-tuning (Kang et al., 2025; Bi et al., 2025). Similarly, probing analyses of multimodal pre-trained models demonstrate that specific subsets of attention heads encode cross-modal interactions and semantic alignment between vision and language modalities (Cao et al., 2020). While these works highlight the presence of specialized heads in VLMs, they primarily focus on single perception-oriented tasks such as grounding and alignment. In contrast, we investigate functionally specialized heads in the context of complex spatial reasoning.

B CogVSR functioin details

Spatial Perception: The capacity to understand the spatial arrangement of objects, including their positions, orientations, distances, and geometric relationships. This function supports tasks

involving mental rotation, spatial navigation, and the interpretation of relative location (e.g., left/right, inside/outside).

Relational Reasoning: The capacity to analyze and compare relationships between objects, entities, or concepts (e.g., larger vs. smaller, cause vs. effect). It supports analogical thinking, comparison, and evaluating structural relations rather than isolated features.

Low-level Visual Perception: The ability to detect basic visual features such as edges, colors, brightness, motion, and textures, without assigning semantic meaning.

High-level Visual Perception: The ability to recognize and interpret visual content by identifying objects, scenes and semantic meaning.

Information Extraction and Understanding: The ability to identify, retrieve, and semantically interpret relevant information from visual or multimodal inputs.

Knowledge Recall: The ability to retrieve stored facts, concepts, or learned information from long-term memory or pretrained knowledge. It includes recalling definitions, factual associations, and world knowledge relevant to the current task.

Math Reasoning: The ability to process quantitative information, including counting, measuring, estimating magnitude, and performing symbolic or arithmetic operations. It requires integrating numerical rules with logical inference.

Decision-Making: The process of selecting the most appropriate option or answer based on prior reasoning, evaluation of evidence, or predefined objectives.

A Concrete Case in CogVSR Dataset



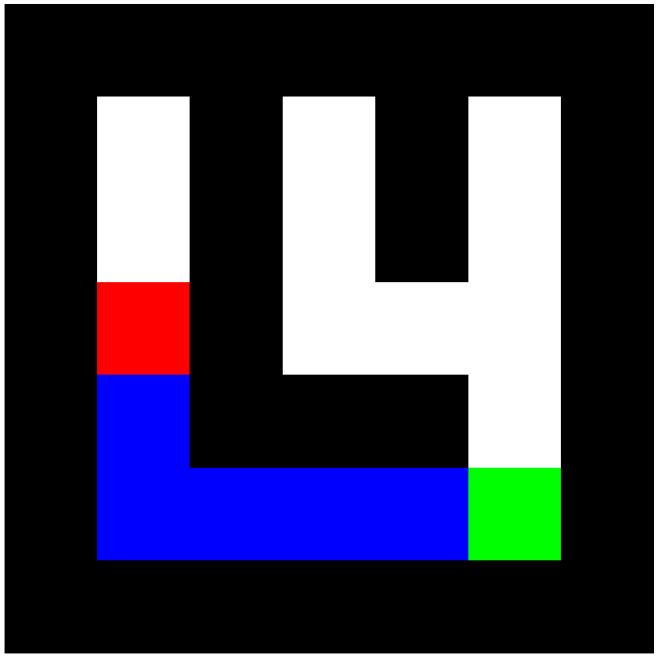
Main Question:
How many planes are left of the blue thing that is in front of the small blue object?
Ground Truth Answer: 1

Generated Reasoning Steps:

1. **Subquestion:** Which object in the image is the small blue object?
   Answer: The small blue bus
   [SKILL: HIGH-LEVEL VISUAL PERCEPTION]
2. **Subquestion:** Which blue thing is located directly in front of that small blue bus?
   Answer: The big blue car
   [SKILL: SPATIAL PERCEPTION]
3. **Subquestion:** How many planes are located to the left of the big blue car?
   Answer: 1
   [SKILL: MATH REASONING]

Figure 5: A detailed case in CogVSR dataset.

A Concrete Case in CogVSR Dataset



Main Question:
The figure represents a maze where black blocks are walls, white blocks are navigable paths, the green block denotes the starting point (S), and the red block marks the exit (E). The objective is to navigate from S to E by following the path marked in blue, moving only in the four cardinal directions (up, down, left, right). A right turn is defined as a 90° clockwise change in direction, and a left turn as a 90° anticlockwise change. Based on the given blue path, how many right turns occur from S to E?

Available options: A. 3 B. 7 C. 4 D. 1.

Ground Truth Answer: D. 1

Generated Reasoning Steps:

1. **Subquestion:** What movement directions does the blue path follow, in sequence, from the starting block (S) to the exit block (E)?
   Answer: West, then North
   [SKILL: SPATIAL PERCEPTION]

2. **Subquestion:** How many times does the path change its movement direction along the blue path from S to E?
   Answer: 1
   [SKILL: RELATIONAL REASONING]

3. **Subquestion:** Changing direction from West to North corresponds to a 90° clockwise rotation; is this a right turn or a left turn?
   Answer: Right turn
   [SKILL: RELATIONAL REASONING]

4. **Subquestion:** How many right turns does the blue path include from S to E?
   Answer: 1
   [SKILL: MATH REASONING]

Figure 6: A qualitative example of turn-based maze reasoning. The model first needs to decompose the path into cardinal directions, counts direction changes, classifies the type of turn (right vs. left), and finally determines the total number of right turns along the path.

A Concrete Case in CogVSR Dataset



Main Question:
The figure represents a 5×5 grid (5 rows and 5 columns) containing various animals, where each 1×1 square is a block and each block contains an animal from {`cat`, `dog`, `elephant`, `giraffe`, `rabbit`}. Based on the grid, how many blocks contain a rabbit?

Available options: A. 6 B. 5 C. 4 D. 0.

Ground Truth Answer: A. 6

Generated Reasoning Steps:

1. **Subquestion:** How many blocks in row 1 contain a rabbit?
   Answer: 1
   [SKILL: HIGH-LEVEL VISUAL PERCEPTION]
2. **Subquestion:** How many blocks in row 2 contain a rabbit?
   Answer: 3
   [SKILL: HIGH-LEVEL VISUAL PERCEPTION]
3. **Subquestion:** How many blocks in row 3 contain a rabbit?
   Answer: 1
   [SKILL: HIGH-LEVEL VISUAL PERCEPTION]
4. **Subquestion:** How many blocks in row 4 contain a rabbit?
   Answer: 0
   [SKILL: HIGH-LEVEL VISUAL PERCEPTION]
5. **Subquestion:** How many blocks in row 5 contain a rabbit?
   Answer: 1
   [SKILL: HIGH-LEVEL VISUAL PERCEPTION]
6. **Subquestion:** What is the total number of blocks in the entire grid that contain a rabbit?
   Answer: 6
   [SKILL: MATH REASONING]
7. **Subquestion:** What is the final answer?
   Answer: A
   [SKILL: DECISION]

Figure 7: A qualitative example of grid-based counting. The model need to first counts rabbits row by row and then aggregates these subcounts to obtain the total number of rabbit-containing blocks in the grid.

B.1 CASE STUDY IN COGVSR

We present three examples from the test set of the CogVSR dataset in Figure 5, Figure 6 and Figure 7.

C PROMPT DETAILS

Prompts for Chain-of-Thought Generation. Below is the specific prompt used to generate the Chain-of-Thought reasoning for the visual spatial reasoning tasks.

Generating CoT Prompt Template

Role Definition:
You are an expert visual spatial reasoning assistant. Given the following question and the correct answer, provide a detailed step-by-step chain of thought reasoning that leads to the correct answer.

Input Context:
Question: {question}
Correct Answer: {correct answer}

Task Instructions:
Please provide a detailed step-by-step analysis of how to solve this problem. Your response should:

1. Analyze what you see in the image.
2. Break down the problem into logical steps.
3. Conclude with the correct answer.

Output Constraints:
Format your response as a clear paragraph, each step is one sentence.

Subquestion Generation Prompt. To enable more granular control over the reasoning process, we employ a specific prompt designed to decompose complex queries into manageable sub-steps. The prompt explicitly instructs the model to map each step to a specific cognitive skill while adhering to a strict JSON output format. The full prompt content is provided below.

Generating Subquestion Prompt

Input Data:
Here is the question:
<question>
{question}
<question>
Here is the chain-of-thought:
<chain-of-thought>
{cot}
<chain-of-thought>

Note:

- Your task is to break the question down into detailed subquestions, ensuring each subquestion can be answered using only one specific cognitive skill.
- You need to create a structured and explicit reasoning process that simulates critical thinking while maintaining clarity and precision.
- The subquestion needs to be easy to answer and the answer needs to be concise.
- The information of chain-of-thought cannot be used directly if it doesn't exist in main query.
- Each subquestion should be derived solely from the main query and the preceding subquestion.
- You CAN NOT retrieval information from chain-of-thought, but you can retrieval from question.
- Each subquestion should be designed to map to exactly one coginitive skill.
- Your output should be formatted as a list of JSON objects, where each object represents a subquestion, its answer, and the required cognitive skill.

- You should use the most efficient logic to analyze the problem and minimize the number of subquestions.

Output format:

```
[
  {
    "subquestion": "<Subquestion text>",
    "answer": "<Concise answer>",
    "cognitive_skill": "<Assigned cognitive skill>"
  },
  {
    "subquestion": "<Subquestion text>",
    "answer": "<Concise answer>",
    "cognitive_skill": "<Assigned cognitive skill>"
  }
]
```

Your answer:

Final Answer Generation Prompt. Once the reasoning steps are complete, we utilize a strict prompt to consolidate the information and produce the final output. This prompt is designed to enforce a specific JSON format and prevent the model from adding extraneous commentary.

Generating Answer Prompt

Input Data:
Here is the question:
<question>
{question}
<question>
Here is the prior knowledge in chain-of-thought (CoT) format:
<prior_knowledge>
{cot}
<prior_knowledge>

Instructions:

- Answer the question carefully.
- You can use the information in prior_knowledge to help you answer the question.
- Your response should be clear and concise.
- Do not include any explanation, commentary, or code.
- Do not output anything after the closing square bracket.

Output Format:
Only output your final answer using this format:

```
[
  {{"answer": "<Your answer here>"}}
]
```

Your answer:

Top-k Token Selection Prompt

Role: You are a token selection assistant. Your task is to identify the tokens in a model-generated answer that carry the most task-relevant semantic information for answering the given question.
Input:
Question:

`{question}`

Generated Answer:

`{answer}`

Indexed Answer Tokens:

`{indexed_tokens}`

Number of tokens to select:

`{3}`

Instructions:

- Select exactly 3 tokens whose semantic content is most directly relevant to answering the question.
- Prioritize answer-bearing information such as entities, spatial relations, directions, attributes, quantities, comparisons, and decision terms.
- Prefer tokens that are necessary for understanding the core meaning of the answer.
- Avoid articles, punctuation, auxiliary verbs, and other function words unless they are semantically necessary.
- Do not use external knowledge or infer reasoning that is not expressed in the question or answer.
- Do not use or infer any cognitive-function or skill label when selecting tokens.
- Select tokens only from the provided indexed token list. Do not modify, merge, or rewrite tokens.
- Return exactly 3 token indices in their original order.

Output Format:

```
{
  "selected_indices": [<index_1>, ..., <index_k>]
}
```

Return only the JSON object without additional explanation.

D ANNOTATIONS

To ensure the quality and reliability of the decomposed subQAF triplets in the CogVision dataset, we design a rigorous multi-stage annotation pipeline, combining expert review and model-based verification. The goal is to verify the logical validity of subquestions, the correctness of their associated cognitive function labels, and the accuracy of the answers.

Stage 1: Validating Subquestion Decomposition In the first stage, we evaluate whether the generated subquestions are logically sound and align with natural human reasoning. For each QA pair, three expert annotators (with backgrounds in linguistics or cognitive science) independently assess the validity of each subquestion. A subquestion is marked `true` if it meaningfully contributes to answering the main question and follows a logical reasoning trajectory. Otherwise, it is marked `false`.

We apply the following filtering criteria:

- **AI-Human Agreement**: If any annotator considers fewer than 60% of the subquestions valid, the entire QA decomposition is discarded.
- **Inter-Annotator Agreement**: A subquestion is deemed invalid if at least two annotators mark it as `false`. If over 40% of the subquestions in a QA pair are invalid under this rule, the whole QA pair is removed.

This filtering ensures that the retained QA decompositions follow coherent, cognitively plausible reasoning chains.

Stage 2: Verifying Cognitive Function Labels In the second stage, annotators evaluate the correctness of the function label f_i assigned to each subQAF triplet (q_i, a_i, f_i). Three annotators independently mark each label as `true` or `false`. When discrepancies occur, annotators collaboratively reassign the correct cognitive label to ensure alignment with the underlying mental operation.

This step ensures that the categorization of subquestions accurately reflects established distinctions between information retrieval, semantic understanding, logical reasoning, and other cognitive processes.

Stage 3: Answer Verification via Model and Human Review In the final stage, we verify the correctness of each answer a_i using manual procedure. A human annotator checks and corrects the answer if necessary. Given the generally objective nature of answers, only one annotator is required for this task.

E COGVSR STATISTICS

The statistics for the CogVSR dataset is shown in Table 6.

Table 6: Distribution of cognitive skills across 3,759 subquestions.

Cognitive Skill	**Count**	**Percentage**
Spatial Perception	1,026	27.29%
High-level Visual Perception	941	25.03%
Relational Reasoning	576	15.32%
Decision Making	498	13.25%
Knowledge Recall	272	7.24%
Information Extraction and Understanding	198	5.27%
Math Reasoning	138	3.67%
Low-level Visual Perception	106	2.82%
Total	3,759	100.00%

F ATTENTION HEAT MAP FOR OTHER MODELS

We present the heatmaps 8 9 10 11 12 for the remaining five models in this subsection. The results reveal a notable universality in the sparsity patterns of attention heads across different architectures.

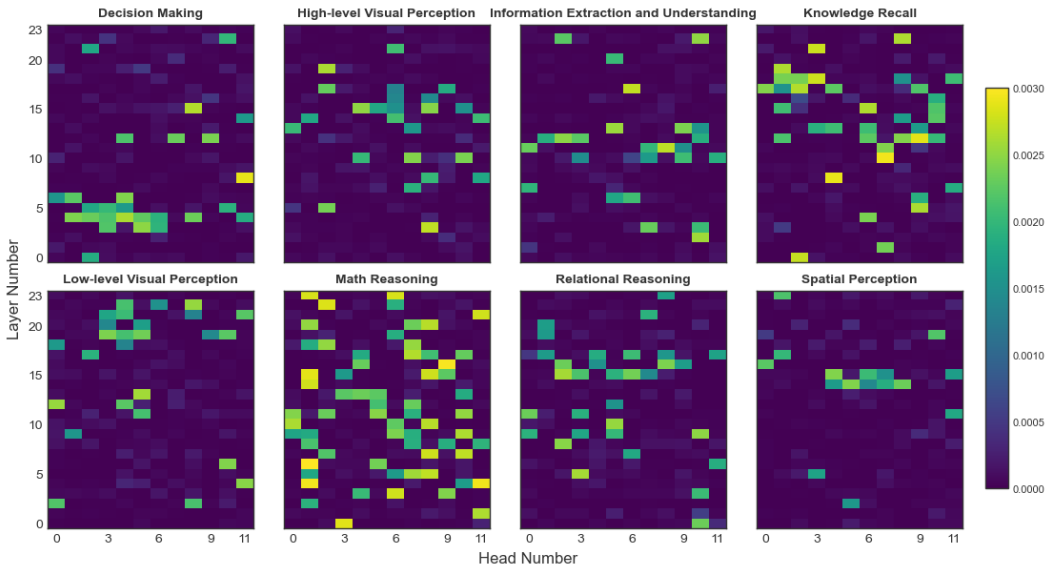


Figure 8: The existence of cognitive heads in Intern3VL-2B. The x-axis represents the head index, while the y-axis indicates the layer index.

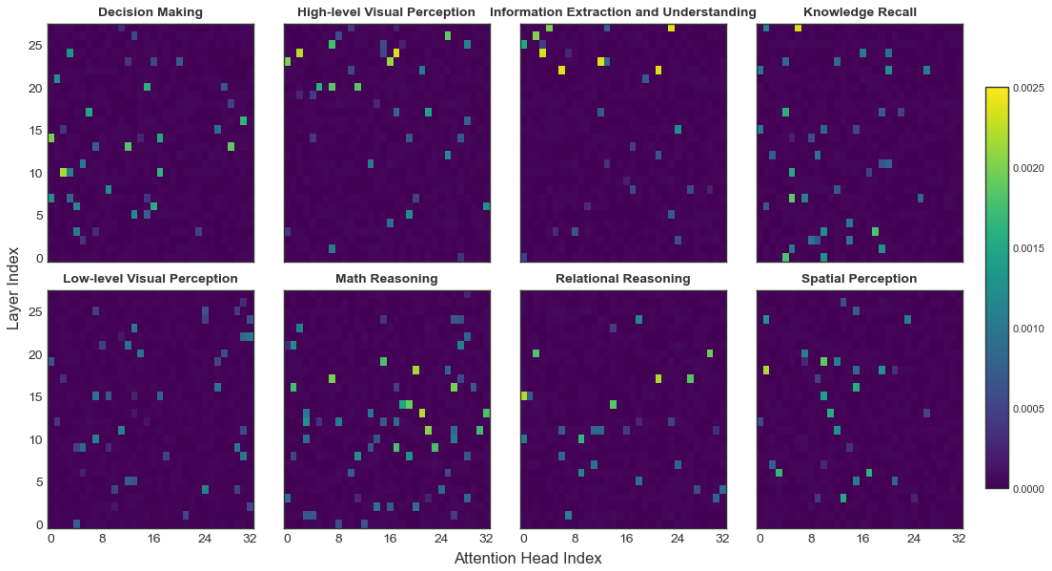


Figure 9: The existence of cognitive heads in LLaMA3.2-11B. The x-axis represents the head index, while the y-axis indicates the layer index.

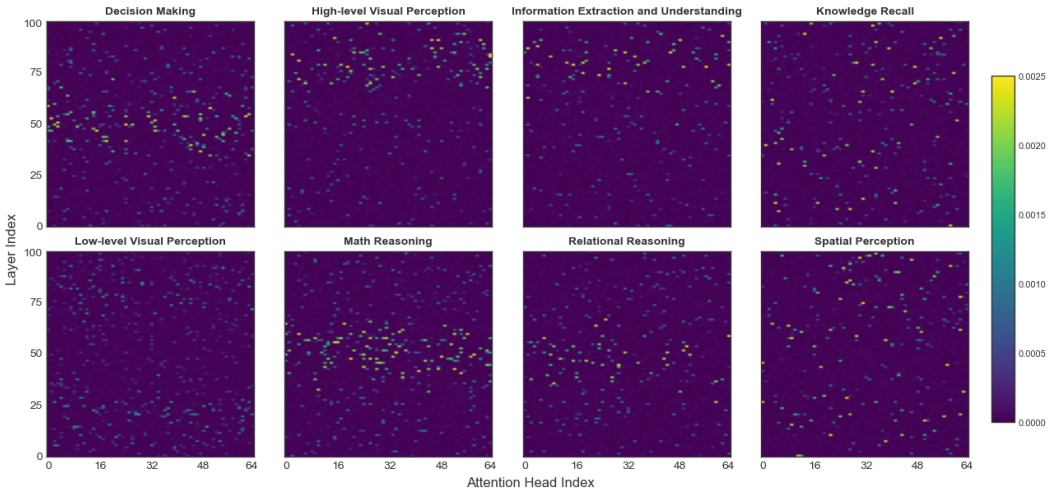


Figure 10: The existence of cognitive heads in LLaMA3.2-90B. The x-axis represents the head index, while the y-axis indicates the layer index.

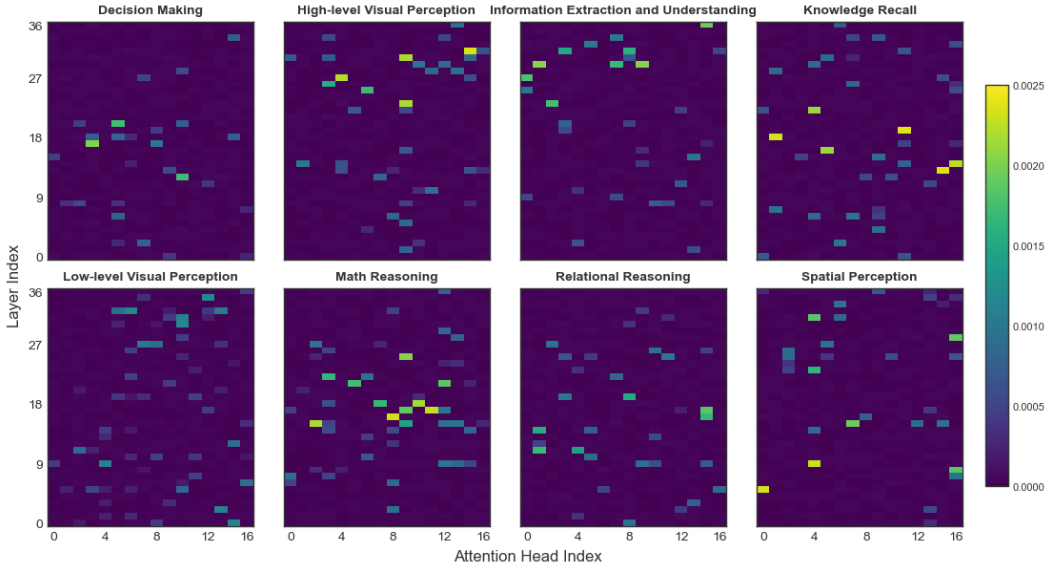


Figure 11: The existence of cognitive heads in Qwen2.5VL-3B. The x-axis represents the head index, while the y-axis indicates the layer index.

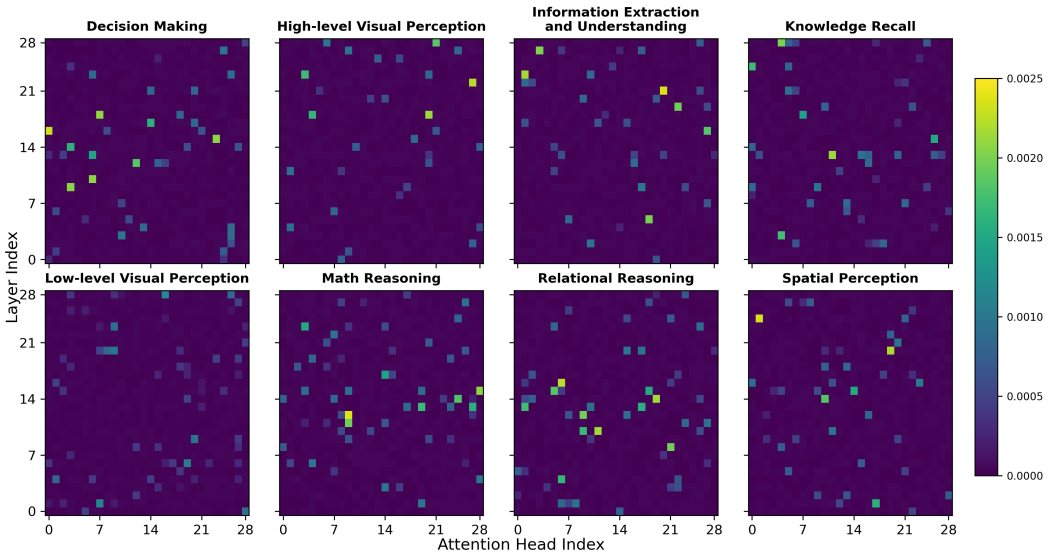


Figure 12: The existence of cognitive heads in Intern3VL-8B. The x-axis represents the head index, while the y-axis indicates the layer index.

G Ablation Study

In this section, we provide a raw output sample demonstrating the token-level analysis. Specifically, we extract the top-3 probability tokens during the reasoning phase to monitor the model's decision boundaries.

Topk Tokens Selection Example



Context:
Consider the real-world 3D location of the objects. Which object is further away from the camera? (A. grey motorcycle, B. green fruits)

Model Output Data:


```
[
  {
    "subquestion": "Using perspective rules, which object is farther
        ?",
    "answer": "The green fruits are farther away.",
```

```
    "cognitive_skill": "Relational Reasoning",
    "Top_3_Tokens": [
       "green",
       "fruit",
       "farther",
    ]
  }
]
```


We also explore alternative strategies for extracting representations, including using the first token, last token, full tokens. The corresponding results shown in Table G show that top-k token is the best.

It has been suggested that cognitive functions may vary by layer depth (Ma et al., 2025c; Zheng et al.), meaning that even if two attention heads across different layers exhibit similar activation patterns aligned with a cognitive function, they might still play distinct roles due to their positional context in the model. Thus, layer information is crucial for interpreting functional specialization. We include the layer vector as a coarse signal to differentiate these roles. We compare using with LayerBias versus without LayerBias. Masking heads with LayerBias causes a larger accuracy drop, showing the value of including layer information.

Table 7: Effect of token selection and layer-bias augmentation on cognitive-head intervention results for Qwen2.5-VL-7B. We compare different token representations with and without the layer-bias vector. For each setting, we report performance after masking either randomly selected heads or identified cognitive heads. Lower values indicate stronger intervention effects.

Token	LayerBias	Inter_Head	Spatial-related Cognitive Functions				Other Cognitive Functions											
			Spatial		Relational		Low-Level		High-Level		Info		Recall		Math		Decision	
			llm	acc	llm	acc	llm	acc	llm	acc	llm	acc	llm	acc	llm	acc	llm	acc
First	without	random	80.72	76.18	82.35	78.01	84.10	80.06	87.95	83.74	89.80	85.44	88.02	83.70	82.49	78.20	86.20	81.96
		cognitive	74.86	75.31	78.64	70.92	62.31	58.47	77.12	72.40	81.36	76.15	79.88	74.96	72.45	67.31	71.24	66.50
	with	random	80.15	76.02	81.93	77.70	84.25	80.18	88.12	83.95	90.10	85.69	88.17	83.92	82.60	78.30	86.35	82.05
		cognitive	70.42	72.54	75.28	66.46	58.04	53.86	74.35	69.22	78.22	72.83	76.40	70.55	68.90	63.87	67.45	62.30
Last	without	random	80.38	76.05	82.04	77.62	84.02	79.91	88.05	83.80	89.65	85.20	88.12	83.64	82.31	78.05	86.12	81.80
		cognitive	86.25	85.12	81.46	73.08	70.10	65.42	82.56	77.14	84.40	79.25	82.35	77.08	75.18	70.44	75.92	70.50
	with	random	80.61	76.30	82.22	77.91	84.40	80.10	88.30	84.02	90.00	85.55	88.31	83.86	82.70	78.25	86.40	82.00
		cognitive	84.33	83.26	78.54	70.18	66.20	61.85	79.42	74.28	81.15	75.90	78.66	73.50	72.00	67.20	72.35	67.10
Full	without	random	80.20	75.88	82.00	77.55	84.18	80.02	88.14	83.92	89.92	85.50	88.05	83.76	82.58	78.18	86.32	81.97
		cognitive	72.38	64.75	74.22	67.41	59.86	54.82	75.60	70.20	80.44	75.05	78.30	72.66	70.82	65.70	68.95	63.74
	with	random	80.34	76.10	82.08	77.72	84.35	80.15	88.20	84.05	90.05	85.60	88.24	83.85	82.66	78.36	86.46	82.10
		cognitive	67.46	59.28	70.28	63.33	55.10	50.34	72.08	66.55	77.25	71.64	74.90	69.30	67.05	62.10	65.30	60.15
Top-k	without	random	79.85	75.70	81.92	77.60	84.10	79.95	88.05	83.92	89.75	85.40	88.10	83.82	82.50	78.20	86.20	81.92
		cognitive	63.92	60.38	68.77	64.42	52.38	48.15	72.55	67.80	77.96	72.84	75.54	70.35	67.41	62.30	64.72	59.86
	with	random	80.04	75.92	82.11	77.86	84.39	80.21	88.27	84.10	90.02	85.73	88.39	84.03	82.76	78.44	86.51	82.19
		cognitive	**57.69**	**53.42**	**63.85**	**59.57**	**47.92**	**43.18**	**68.31**	**63.04**	**73.77**	**68.55**	**71.23**	**65.98**	**62.84**	**57.63**	**59.41**	**54.22**

H MLP

We train a two-layer multi-class MLP for cognitive function classification. The first layer applies a shared linear projection to each multi-head representation vector, reducing each to a 64-dimensional embedding. These embeddings are then flattened and concatenated into a single vector of size $64 \times numberofheads$. This vector is fed into a hidden layer with 512 units, followed by a ReLU activation and a dropout with a rate of 0.3. The final output layer maps the 512-dimensional hidden representation to the set of cognitive function labels.

The model is trained using the Adam optimizer with a learning rate of 10^{-4} and a cross-entropy loss. Training proceeds for 100 epochs. The test accuracy of our classification method across all VLM models is summarized in the Table 8.

Table 8: The test accuracy (%) of probing method on different LLMs.

Dataset	InternVL3-2B	InternVL3-8B	Qwen2.5VL-7B	Qwen2.5VL-3B	LLaMA3.2-11B	LLaMA3.2-90B
CogVSR	78.74	75.88	82.15	84.22	89.64	83.44

More Results of MLP Probing VS Logistic Regression. We conduct comparative experiments using two classification methods—MLP and logistic regression—on Qwen2.5-VL-7B. We evaluate their effectiveness from two perspectives: (1) classification performance and (2) negative intervention experiments, where we mask the top-30 cognitive heads identified by each method. As shown in Table 9, MLP achieves higher probing performance. For intervention results, lower post-masking scores indicate stronger functional relevance of the identified heads.

Table 9: Comparison between MLP probing and logistic-regression probing.

Probe	Macro-F1 ↑	Micro-F1 ↑	Post-Mask Acc ↓	Random Acc	Δ↑
Logistic Regression	72.5	76.8	65.7	81.0	15.3
MLP	**80.4**	**84.1**	**58.2**	81.1	**22.9**

I Spatial Head Augmentation (SHA) Prompt

To enhance visual grounding, we propose the Spatial Head Augmentation (SHA) method. We utilize *Gemini 2.5-Flash* to generate precise 2D bounding boxes and object segmentation masks. This process ensures that the model focuses on the specific entities referenced in the textual context.

SHA Generation Prompt

Input Context:
Given this image with the following context:
Question: {question}
Answer: {answer}

Task Instructions:
Detect and provide segmentation masks for all prominent objects explicitly mentioned in the question and answer.

Output Format:
Output a JSON list of segmentation masks where each entry contains:
- `"box_2d"`: the 2D bounding box as [ymin, xmin, ymax, xmax] normalized to 0-1000
- `"mask"`: the segmentation mask as a base64-encoded PNG
- `"label"`: descriptive text label for the object

Constraints:
Focus on the objects mentioned in the question and answer. Only label the explicitly mentioned object like car, bird, person, beach etc., not ”the thing”, ’it’, ’the opposite one’.

J Masked Ratio vs Performance

We gradually mask out the number of cognitive heads and see how the model's behavior change. The results can be found in the Figure 13.

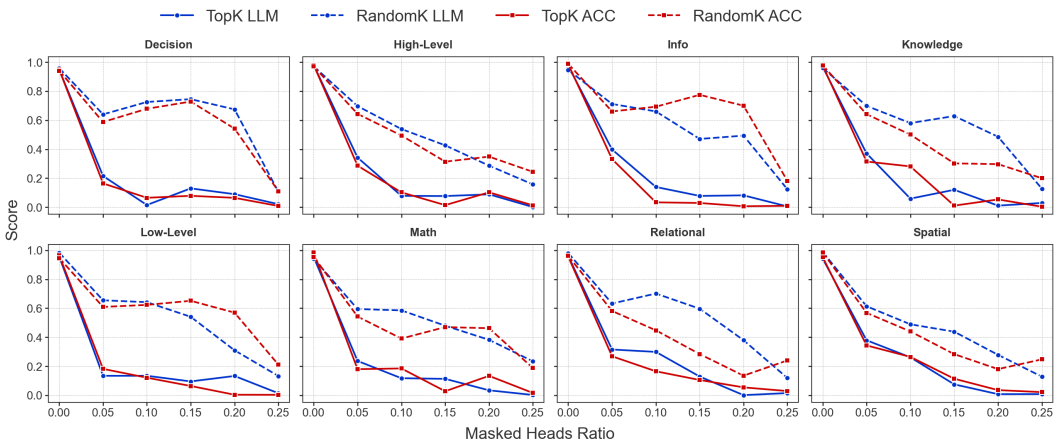


Figure 13: The performance of Qwen2.5-VL-7B after masking out top K cognitive heads vs K random heads on inference, high-level visual reception, low-level visual reception, and decion-masking.

K IMPORTANCE CURVE

We ranked the importance scores and identified the elbow point, as illustrated in Figure 14.

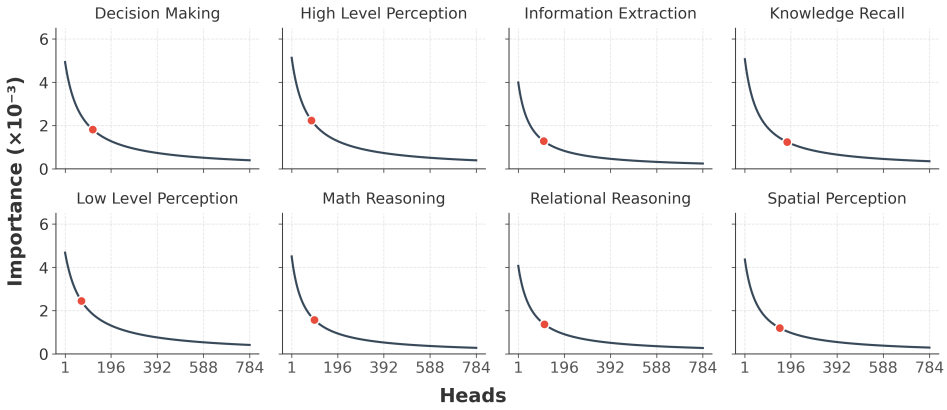


Figure 14: Importance curve for eight functions of Qwen2.5-VL-7B.

L MORE ANALYSIS OF SHA

We further analyze the robustness of activated spatial and relational heads under different input augmentations by measuring the similarity of head-importance maps using Pearson correlation. As shown in Table 10, adding bounding box information yields high consistency with the original setting (spatial: 0.80, relational: 0.78), which further increases when combined with masking (spatial: 0.84, relational: 0.81). This suggests that these input perturbations act as stimuli that activate spatially related heads with higher importance scores, while the underlying functional heads remain intrinsic to the model. Additionally, the relatively high Pearson correlation coefficients across different augmentations (e.g., Shape and Color) indicate that the activation patterns of functional heads remain stable under spatially relevant input perturbations.

Table 10: Similarity of activated spatial and relational heads under input augmentations. We report Pearson correlation over head-importance maps.

Comparison	Spatial r	Relational r
Original vs. +BBox	0.80	0.78
Original vs. +BBox+Mask	0.84	0.81
Shape vs. Color	0.74	0.71
Shape vs. +BBox+Mask	0.69	0. 74

M HYPERPARAMETER ANALYSIS

We vary the intervention strength α and observe that positive interventions ($\alpha > 0$) yield generally stable and consistent performance improvements. In contrast, negative interventions ($\alpha < 0$) lead to performance degradation, indicating that suppressing activation along the identified directions adversely affects spatial reasoning. These results further support the functional relevance of the identified attention heads. Notably, we achieve the strongest gains at $\alpha = 0.1$.

Table 11: Effect of intervention strength α on CogVSR spatial functions. $\alpha = 0$ denotes no intervention.

α	Spatial Perception	Relational Reasoning
-0.2	78.62	80.44
-0.1	79.31	81.26
0.0	80.04	82.11
0.05	81.12	83.02
0.1	**81.89**	**83.76**
0.2	81.54	83.41
0.3	80.87	82.68