

How Do LLMs and VLMs Understand Viewpoint Rotation Without Vision? An Interpretability Study

Zhen Yang¹, Ping Jian^{*1}, Zhongbin Guo¹, Zuming Zhang¹,
Chengzhi Li¹, Yonghong Deng¹, Xinyue Zhang¹, Wenpeng Lu²

¹School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

²Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China
{bityangzhen, pjian}@bit.edu.cn

Abstract

Over the past year, spatial intelligence has drawn increasing attention. Many prior works study it from the perspective of *visual*-spatial intelligence, where models have access to visuo-spatial information from visual inputs. However, in the absence of visual information, whether linguistic intelligence alone is sufficient to endow models with spatial intelligence, and how models perform relevant tasks with *text*-only inputs still remain unexplored. Therefore, in this paper, we focus on a fundamental and critical capability in spatial intelligence from a linguistic perspective: viewpoint rotation understanding (VRU). Specifically, LLMs and VLMs are asked to infer their final viewpoint and predict the corresponding observation in an environment given textual description of viewpoint rotation and observation over multiple steps. We find that both LLMs and VLMs perform poorly on our proposed dataset while human can easily achieve 100% accuracy, indicating a substantial gap between current model capabilities and the requirements of spatial intelligence. To uncover the underlying mechanisms, we conduct a layer-wise probing analysis and head-wise causal intervention. Our findings reveal that although models encode viewpoint information in the hidden states, they appear to struggle to bind the viewpoint position with corresponding observation, resulting in a hallucination in final layers. Finally, we selectively fine-tune the key attention heads identified by causal intervention to improve VRU performance. Experimental results demonstrate that such selective fine-tuning achieves improved VRU performance while avoiding catastrophic forgetting of generic abilities¹.

^{*}Corresponding author.

¹Our dataset and code will be released at https://github.com/Young-Zhen/VRU_Interpret

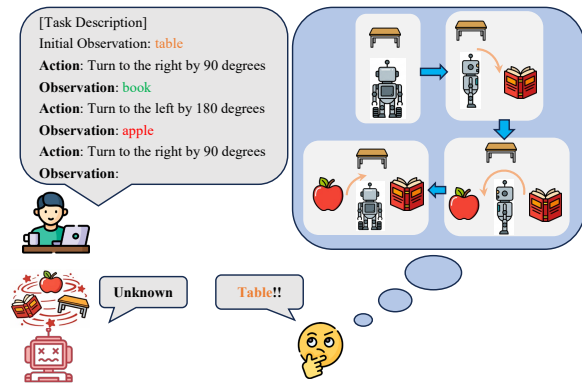


Figure 1: For textual viewpoint rotation understanding, human can easily imagine the spatial scenario and achieve 100% accuracy, while LLMs and VLMs struggle to output the correct answer.

1 Introduction

With the rapid development of large language models (LLMs) and vision-language models (VLMs), spatial intelligence has attracted increasing attention (Feng et al., 2025). Spatial intelligence generally involves perceiving and mentally manipulating spatial relationships (Yang et al., 2025b), and requires myriad capabilities such as relational reasoning (Wang et al., 2024), distance estimation (Zhang et al., 2025b), viewpoint transformation between egocentric and allocentric perspectives (Li et al., 2025b; Xue et al., 2025), etc. In this work, we focus on a fundamental and critical capability in spatial intelligence: **viewpoint rotation understanding** (VRU) (Zhang et al., 2025a). Specifically, we investigate whether models can accurately infer the final position and the corresponding observation in an environment, after undergoing a sequence of viewpoint rotations and receiving observations.

Many prior studies (Yang et al., 2025b; Li et al., 2025b; Zhang et al., 2025c; Ma et al., 2024) have taken efforts on benchmarking and improving the spatial intelligence using visual data. However, it is noteworthy that spatial intelligence exists irre-

spective of sensory modality, even a blind person can perceive space through other senses (Gardner, 1983). Regrettably, whether VLMs and LLMs can comprehend viewpoint rotation (i.e., VRU) without visual information still remains underexplored. To address this gap, we first propose a simple textual viewpoint rotation understanding dataset **VRUBench**, in which models are provided only with textual descriptions of multi-step viewpoint rotations and their associated observations, and are required to predict the observation after the final rotation (as illustrated in Figure 1.). Results show that human can easily achieve 100% accuracy on VRUBench while the majority of current LLMs and VLMs perform poorly. Even state-of-the-art (SOTA) models such as Qwen3-VL (Bai et al., 2025a), which demonstrates remarkable performance across numerous challenging tasks, achieves only $\approx 60\%$ accuracy, highlighting a substantial gap between current model capability and the requirements of spatial intelligence.

Despite the poor performance of both LLMs and VLMs, we still obtain the following valuable findings: i) **Firstly, VLMs consistently outperform LLMs** (e.g., Qwen3-VL-8B vs Qwen3-8B), indicating that training with visual data enhances models’ spatial perception ability, even when no visual input is available at inference time. ii) **Secondly, for both VLMs and LLMs, enabling models to “think-then-answer” achieves better performance than the no-reasoning setting.** This observation complements the conclusion of prior work (Yang et al., 2025b), which reports that eliciting model’s reasoning capability through chain-of-thought (CoT) (Wei et al., 2022) does not improve the performance on their *visual*-spatial dataset VSI-Bench. In contrast, our results indicate that under a *text*-only setting, reasoning-based methods can somewhat enhance model performance, suggesting an inherent difference between text-only and vision-based spatial perception.

To uncover the intrinsic mechanisms underlying the poor VRU performance, which could provide insights for building advanced models with stronger spatial intelligence, we conduct comprehensive interpretability studies including layer-wise and head-wise analysis. Concretely, to evaluate the capabilities of LLMs and VLMs for encoding the information of viewpoint rotation, we conduct a layer-wise probing analysis on VRUBench. We find that models have a strong capability to encode the direction and angle of the viewpoint ro-

tation at each step, while the capability of encoding absolute orientation gradually emerges in early-to-middle layers and diminishes in later layers.

To figure out what happens in the later layers, a causal intervention method, known as *path patching* (Wang et al., 2023) is utilized to conduct head-wise interpretation. Through path patching, a small fraction of attention heads located in middle-to-late layers are identified as responsible for VRU performance (i.e., key heads), where the attention patterns are also interpreted in a human-understandable manner. Our findings reveal that after the model acquires the viewpoint information through early-to-middle layers, the **proposal head** extracts all possible candidate answers then the final answer is selected by **answer decision head**. We also find a special head, **unknown head**, which reflects a cautious behavior in answer prediction and may be induced by alignment training (Bai et al., 2025b; Zhu et al., 2025). In other words, in the later layers, the model gradually transitions from orientation perception to the decision-making stage, which explains aforementioned decline in orientation-probing accuracy. However, the poor performance in VRU demonstrates that these heads fail to effectively bind the perceived viewpoint orientation from the preceding layers with the corresponding observation at that orientation, resulting in hallucinated answer decisions.

As aforementioned, since the key heads do not function effectively to choose correct answers, it motivates us to selectively fine-tune these heads to improve the VRU performance. Experimental results demonstrates that selective fine-tuning can achieve enhanced VRU performance with only 50% of the GPU hours required for full fine-tuning, while preserving the general capabilities.

2 Related Work

Spatial Intelligence Spatial intelligence exists irrespective of sensory modality, even a blind person can perceive space through other senses (Gardner, 1983). Regrettably, researches on spatial intelligence primarily focus on *visual*-spatial intelligence (Yang et al., 2025b; Ramakrishnan et al., 2025; Guo et al., 2025; Tang et al., 2024) while *textual*-spatial intelligence still lacks a sufficient study. The most relevant existing studies have largely concentrated on benchmarking and enhancing LLM performance in understanding the *spatial relations* within *static, single-viewpoint* scenarios (Shi et al.,

2022; Li et al., 2024; Yamada et al., 2024), leaving dynamic and viewpoint-shifting scenarios underexplored. This motivates our study of the proposed VRUBench, where spatial understanding must incorporate changes over time and viewpoint. Importantly, so far, no systematic research has analyzed the mechanisms underlying spatial intelligence.

Mechanistic Interpretability Interpreting the inner mechanisms of LLMs and VLMs has attracted increasing attention in recent years (Madsen et al., 2023). Despite treating models as black box, these works tend to uncover the intrinsic cause of models behaviors, such as hallucination (Gao et al., 2025a; Wang et al., 2025), mathematical ability (Zhang et al., 2024b; Yu and Ananiadou, 2024; Li et al., 2025a), safety (Huang et al., 2025; Qi et al., 2025; Zhou et al., 2025) etc. However, the underlying mechanisms about how models achieve *textual-spatial intelligence* in dynamic, viewpoint-shifting scenarios still remain mysterious, motivating the interpretability analysis in this work. Furthermore, a deeper understanding of model mechanisms has also shed light on model improvement, including better architectural design (Fu et al., 2023), token/weight pruning (Jiang et al., 2025; Gao et al., 2025b) and post hoc enhancement (Li et al., 2023; Chuang et al., 2024; Ju et al., 2024a). Benefiting from advances in interpretability, some studies have attempted to perform targeted interventions on the behavior of specific attention heads based on interpretability results. For example, Inference-Time Intervention (ITI) (Li et al., 2023) proposes that adding steering vectors to certain attention heads can improve the faithfulness of model outputs, while Yin et al. (2024) proposes to train the steering vectors on downstream task and add them to the hidden representations at selected heads. Our work generally follows the “interpret-then-improve” paradigm (Zhang et al., 2024b), where models are not only being interpreted but also improved based on the interpretability findings.

3 Task Definition and Dataset

3.1 Task Definition

For text-only VRU, the prompt P for LLMs and VLMs can be formalized as:

$$P = I \oplus O_0 \oplus \overbrace{A_1 \oplus O_1 \oplus A_2 \oplus \dots \oplus A_n}^{n \text{ steps}}, \quad (1)$$

where \oplus denotes the string concatenation operation, I is the task instruction, O_0 , A_i and O_i describe

the initial observation, direction and angle of i -th viewpoint rotation, and the observation after i -th rotation in the environment, respectively. Examples are provided in Figure 1 and Appendix A. After the final rotation A_n , models \mathcal{M} are tasked to predict the corresponding observation:

$$\hat{O}_n = \mathcal{M}(\cdot | P). \quad (2)$$

In Sec.4, we test 2 evaluation settings: **directly outputting the answer** and **think-then-answer**. For the former, we extract the first token generated by models as \hat{O}_n . For the latter, models are asked to output the answers between `<ans></ans>` tags after chain-of-thought (CoT), and the \hat{O}_n is extracted from the tags, detailed in Appendix B.2.

We adopt the accuracy of observation prediction as the evaluation metric:

$$Acc = \frac{\sum_{i \in \mathcal{D}} \delta(\hat{O}_n^i, O_n^i)}{|\mathcal{D}|}, \quad (3)$$

where $\delta(x, y)$ is an indicator function that equals 1 for $x = y$ and 0 otherwise, $|\mathcal{D}|$ is the dataset size.

3.2 Dataset Details

For dataset synthesis, we construct a simulated environment containing 100 objects that may appear in real indoor rooms. The viewpoint rotation angle θ here is restricted to $\theta \in \{0^\circ, 90^\circ, 180^\circ, 270^\circ, 360^\circ\}$, which ensures that, after rotating to the nearest adjacent viewpoint, objects observed in the previous view do not simultaneously appear in the rotated view, thereby avoiding ambiguity in observation prediction. At every step, model randomly rotates its viewpoint by θ degrees to the left or right. If the model has never visited the resulting viewpoint before, we randomly sample one object from the 100 candidates and present it as the observation. Otherwise, we replay the previously observed object. After the final action, the ground-truth of observation is set as the object corresponding to the final viewpoint orientation if the viewpoint is visited before otherwise “unknown”. Finally, the constructed dataset **VRUBench** consists of 19,591 instances, including 4,614 samples with 2-step rotations, 4,977 with 3-step rotations, 5,000 each with 4-step and 5-step.

4 Evaluation

4.1 Experimental Settings

Since research on specialized spatial intelligence models is still in its early stages, this paper pri-

	Models	2-step (%)	3-step (%)	4-step (%)	5-step (%)	Avg. (%)
LLMs	L2-7B-chat	5.44	17.22	26.24	25.64	18.90
	L3.1-8b-Instruct	35.39	37.71	38.66	34.74	36.65
	Q2.5-7B	58.00	41.33	36.18	33.28	41.89
	Q2.5-14B	61.77	60.64	59.42	58.00	59.92
	Q2.5-32B	88.56	74.20	67.54	62.28	72.84
	Q3-4B (-thinking)	35.37 (43.84)	31.75 (45.69)	33.24 (45.66)	34.40 (43.84)	33.66 (44.77)
	Q3-8B (-thinking)	41.66 (72.02)	38.22 (52.82)	40.66 (44.64)	43.56 (40.42)	41.02 (52.09)
	Q3-32B (-thinking)	50.04 (81.32)	51.94 (63.69)	53.92 (49.44)	56.18 (39.32)	53.08 (57.99)
VLMs	Q2.5-VL-3B	52.73	40.00	33.42	25.50	37.62
	Q2.5-VL-7B	53.32	49.11	48.38	44.22	48.67
	Q2.5-VL-32B	68.14	61.94	55.62	54.30	59.84
	Q3-VL-4B (-thinking)	47.83 (50.65)	48.60 (45.11)	49.06 (40.30)	48.20 (36.98)	48.44 (43.11)
	Q3-VL-8B (-thinking)	64.33 (85.57)	61.28 (62.38)	57.04 (55.48)	54.58 (50.94)	59.21 (63.16)
	Q3-VL-32B (-thinking)	77.50 (97.90)	71.03 (96.44)	67.54 (96.16)	64.42 (95.82)	69.98 (96.55)
	G3-Flash (-thinking)	84.15 (93.15)	77.22 (90.32)	73.17 (85.71)	69.73 (76.65)	75.91 (86.32)
	Human Performance	100	100	100	100	100

Table 1: The accuracy of various LLMs and VLMs in i -step ($i = 2, 3, 4, 5$) viewpoint rotation tasks. We also compute the average accuracy (**Avg.**) to evaluate the overall performance. Due to table length constraints, we use L to denote **LLaMA**, Q to denote **Qwen**, and G to denote **Gemini**.

mainly focuses on generic LLMs and VLMs. For LLMs, we experiment with LLaMA2-7B-chat (Touvron et al., 2023), LLaMA3.1-8B-Instruct (Team, 2024), Qwen2.5 series (Yang et al., 2024), Qwen3 series (Yang et al., 2025a). For VLMs, we test Qwen2.5-VL series (Bai et al., 2025b), Qwen3-VL series (Bai et al., 2025a) and the latest frontier model Gemini3-Flash (Google, 2025). Besides, for Qwen3 and Qwen3-VL, which support reasoning mode, the results with the *thinking* mode enabled are reported as “-thinking”, which allows us to assess the impact of explicit reasoning on VRUBench. Further details are provided in Appendix B. To ensure fairness and consistency, we strictly maintained uniform hyperparameter settings across all evaluated models. Specifically, we adopted a greedy decoding strategy for text generation, setting the temperature to 0.0 and do_sample to False. Furthermore, the generation length and thinking budget constraints are set as: when thinking was enabled, the thinking_budget was set to 1024 (see Appendix B.2), and the max_new_tokens was capped at 2048. In contrast, when thinking was disabled, the thinking_budget was set to 0, and max_new_tokens was strictly limited to 5. All other hyperparameters not explicitly mentioned were kept at their default configurations.

4.2 Results and Discussion

LLMs/VLMs vs. Human As shown in Table 1, human can easily achieve 100% accuracy in VRU while both LLMs and VLMs perform poorly. Considering that the VRU task represents only one of the capabilities required for spatial intelligence, the overall unsatisfactory performance in VRU indicates that current models are still far from possessing genuine spatial intelligence.

LLMs vs. VLMs Apart from the human evaluation results, we also observe systematic differences across model types, that is, VLMs generally outperform LLMs. This advantage is observed not only across different model families (e.g., Qwen2.5-VL-7B vs. LLaMA2-7B), but also consistently between VLMs and their corresponding LLM backbones sharing the same architecture (e.g., Qwen2.5-VL-7B vs. Qwen2.5-7B). It demonstrates that although no visual information is available at inference time, VLMs trained with visual data still exhibit stronger VRU ability than LLMs that have never been exposed to visual data.

Takeaway I: Training with visual data can benefit the performance in textual spatial task.

Notably, a very recent study (Guo et al., 2026) has compared text-only vs. visual training on more complex visual datasets. They converted 17 complex subtasks within visual-spatial intelligence, e.g., navigation, embodied perception, into text descriptions and compared the performance of LLMs and VLMs. The final results align perfectly with our observations, that is, VLMs consistently outperform LLMs on these *complex* text-only spatial tasks, which further demonstrates Takeaway I.

Reasoning vs. Non-reasoning Despite the marginal improvement in some cases (e.g., Qwen3-VL-8B-thinking outperforms Qwen3-VL-8B by only 3.95), enabling the models to reason before outputting answers generally leads to better performance compared to directly outputting answers. Appendix B.3 presents the model outputs when enabling the reasoning mode. It can be observed that, in this mode, models do try to reason based on the viewpoint rotation description rather than just follow templates in the data, which confirms the genuine boost to VRU task performance provided by the reasoning mechanism. Combining with the findings in Yang et al. (2025b), which shows that reasoning methods such as CoT are not effective to improve the performance on *visual*-spatial intelligence, it further demonstrates the inherent difference between text-only and visual-based spatial intelligence.

Effect of Scaling Law It can be observed that the overall performance improves as the model size increases, demonstrating that the scaling law (Kaplan et al., 2020; Hoffmann et al., 2022) still holds in the VRU task.

5 How Do VLMs and LLMs Simulate Viewpoint Rotation?

The generally poor performance of both VLMs and LLMs on the VRU task motivates us to investigate how they internally simulate viewpoint rotation, with the goal of identifying the underlying causes of their limitations. Here, we primarily focus on the mechanisms of intrinsic, implicit reasoning (i.e., under the “directly outputting answers” setting), which has yielded a variety of valuable findings for understanding model behaviors (Hanna et al., 2023; Zhang et al., 2024b; Yu and Ananiadou, 2025). Interpretability studies of the explicit reasoning process such as CoT follow another line of research, therefore fall outside the scope of this work. We

leave it as a future work.

5.1 Layer-wise Probing Analysis

5.1.1 Probing Method

We first adopt the probing method to investigate whether models encode **i) the direction and angle at every rotation step**, and **ii) the absolute orientation after each rotation**. Specifically, we feed each sample in the proposed dataset into LLMs and VLMs, and extract the hidden representation associated with the last token at each action A_i (the token is “\n”) from all layers, as illustrated in Figure 2(a). Subsequently, for the probing target i) and ii), we annotate each representation with different label respectively. For instance, when probing for i), the label is the direction and angle described in A_i . As for ii), the label is the absolute orientation after A_i . Data-label pairs for different probing targets can be found in Appendix C.1.

Finally, we obtain the representations, denoted as R_l where $l \in \{1, 2, \dots, L\}$ indexes the layer, and their corresponding direction/angle/orientation labels Y . Then a linear ² probing classifier \mathcal{F}_l is trained to map each R_l to labels Y . By measuring the test set performance of \mathcal{F}_l , we can infer the extent to which the hidden layer encodes the direction/angle/orientation (Ju et al., 2024b, 2025).

5.1.2 Do Models Encode Direction & Angle?

As shown in Figure 2(b) and (c), the probing accuracy for both direction and angle exceeds 99% in most layers. It is not surprising since the information of rotation direction and angle is explicitly provided in A_i , the models could aggregate the information by self-attention mechanism (Vaswani et al., 2017) even when the texts about direction (left/right) or angle (0/90/180/270/360) are tokenized into separate tokens (Yu and Ananiadou, 2024). In summary, **LLMs and VLMs do encode the rotation direction and angle at each step**.

5.1.3 Do Models Encode Orientation?

Regrettably, when probing the absolute orientation after the viewpoint rotation, as shown in Figure 2(b) and (c), both LLMs and VLMs struggle to achieve high accuracy across all layers. However, VLMs still exhibit overall better performance compared to LLMs, which is consistent with the better

²We also explored non-linear probes such as MLPs in preliminary experiments, but observed that the conclusions drawn from probing were consistent with linear ones. Therefore, we ultimately choose to use the linear probes. Check Appendix C.2 for further clarification.

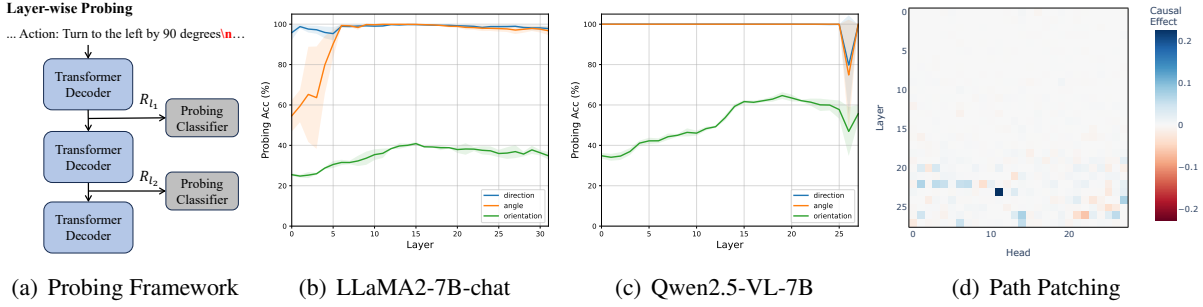


Figure 2: (a)-(c): Illustration of layer-wise probing and the probing results (direction/angle/orientation) on LLaMA2-7B-chat and Qwen2.5-VL-7B. (d): Path patching results of Qwen2.5-VL-7B on proposed VRUBench. For each head, a darker color indicates a larger causal effect, which also reflects the importance of the head in VRU.

performance of VLMs in VRU. It indicates that the capability of encoding orientation is positively correlated with the model’s final performance on the VRU task. In brief, **VLMs exhibit a stronger ability to encode the information of absolute viewpoint orientation than LLMs.**

Another interesting observation is that the capability of encoding orientation within VLMs gradually emerges in early-to-middle layers (1-20) but diminishes in the later layers (21-28). Note that orientation perception is only an intermediate step toward predicting the final observation, i.e., there remains a gap between orientation and the final answer. The decrease in orientation probing accuracy therefore suggests that a pattern shift may occur in the middle-to-late layers. To verify this, we subsequently conduct a head-wise causal intervention to uncover what happened in the later layers.

5.2 Head-wise Causal Intervention

5.2.1 Path Patching

Path patching is an interpretability technique that tests the causal role of specific internal computation paths in a model by selectively replacing activations along those paths and observing the effect on the output (Goldowsky-Dill et al., 2023; Zhang and Nanda, 2024; Tigges et al., 2024). It typically involves the following steps: 1) running model on clean data D_{cl} and corrupted data D_{cor} , 2) intervening targeted head’s activation with corrupted one while freezing others, 3) measuring the causal effect after patching under specific metrics. Here, we construct the corrupted data as flipping the rotation direction at A_n , i.e., the last rotation step. And the metric associated with causal effect is defined as the logit difference between the answers of clean

data and corrupted data:

$$\text{logit}_* = \mathcal{M}(t_{cl} | \cdot) - \mathcal{M}(t_{cor} | \cdot), \quad (4)$$

where $*$ denotes the clean/corrupted/patched run, t_{cl} and t_{cor} are the first token of the answer to clean and corrupted data, respectively. Note that when the answer does not change after flipping the rotation direction, logit_* will always be 0. To avoid this, all the clean-corrupted data pairs that share the same answers are filtered out. Finally, the causal effect is defined as:

$$\phi_i = \frac{\text{logit}_{pt} - \text{logit}_{cl}}{\text{logit}_{cor} - \text{logit}_{cl}} \quad (5)$$

$$\Phi = \frac{\sum_{i=1}^{|\Omega|} \phi_i}{|\Omega|}, \quad (6)$$

where ϕ_i is the causal effect for i -th clean-corrupted data pair, $|\Omega|$ denotes the size of such data pair after filtering, and Φ represents the averaged causal effect. More details about path patching are provided in Appendix D. For simplicity, we primarily report the results of Qwen2.5-VL-7B in following discussion, while the results of other models can be found in Appendix E.

5.2.2 Key Heads Identification

Figure 2(d) depicts the results of path patching (i.e., Φ), where the intensity of the color reflects the strength of causal effect. We can find that only a small fraction of heads have a relatively significant effect on the answer generation. In other words, the key attention heads in VRU is particularly sparse, which is consistent with previous findings (Zhang et al., 2024b; Yang et al., 2025c). More importantly, **these key heads are mainly located in the middle-to-upper layers (21-28) while the heads in early layers almost exhibit zero causal effect.**

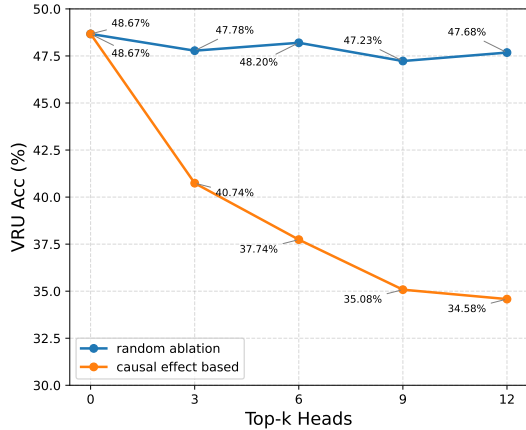


Figure 3: The VRU performance after knocking out randomly-selected K heads (“random ablation”), and ablating the top- K heads sorted by their causal effects.

Recalling that the information encoding viewpoint orientation also declines after layer 20 (Section 5.1.3), it suggests that these key heads must play a pivotal role in this process.

5.2.3 Validation

To further validate the faithfulness of the identified key heads, we conduct additional validation experiments by measuring the performance degradation when knocking out these heads, and comparing it with randomly knocking out. Settings of the knocking out experiments are detailed in Appendix F. Figure 3 illustrates the results after ablating the top- k heads sorted by causal effect. It shows that, as the key heads are gradually knocked out, the performance drops significantly, while remaining stable in random ablation. It demonstrates that the identified heads do play an important role in VRU task, and have strong causal significance for VRU performance. Given that the importance and faithfulness of these heads have been validated, to uncover the function of the identified heads during VRU, we subsequently conduct an attention pattern analysis to check what content the key heads attended to.

5.2.4 Attention Pattern Analysis

The sparsity of the key heads enables us to analyze their attention patterns manually. Figure 4 visualizes the attention patterns with specific question of the key heads identified by path patching. More results about attention pattern are supplemented in Appendix G. Regarding the function of these key heads, it can be summarized as follows:

Proposal Head We find that the head 22.1 consistently focuses on all the candidate answers in prompts, including the token “unknown” when the final answer cannot be determined based on the observation histories. In other words, such head acts as a proposer that extracts all the possible answers and delivers the results to subsequent components.

Answer Decision Head After receiving the candidate answers from 22.1, the answer decision head 26.14 and 23.11 (Appendix G) precisely increases its attention to the chosen output (i.e., “plant”) while simultaneously reducing attention to other candidates (“unknown” and “trash_can”), resulting in a pronounced focus on the final answer.

Unknown Head Surprisingly, even after the decision heads have precisely focused on the final answer, the subsequent heads do not directly maintain attention on the selected answer until the final output. Instead, we observe that the head 27.14 exhibits a strong attention to the token “unknown” just before the final layers. Moreover, we find that when the model is instructed to output an alternative token (semantically unrelated) in cases where the final observation cannot be determined, it loses such special attention. But the same pattern re-emerges when it is replaced with “不知道”, the Chinese translation of “unknown” (shown in Appendix H). These observations demonstrate that the head is not simply focusing on what to respond when the answer is indeterminate. Instead, its distinctive attention to “unknown” and “不知道” reflects, to some extent, an inherent decision-making bias of the head. That is, a preference for cautious responses that acknowledge uncertainty rather than providing potentially incorrect definitive answers. An evidence is that after ablating the head 27.14, the overall proportion of outputting “unknown” significantly drops from **65.78%** to **40.73%**. The reason could be that after undergoing alignment training, the demand of safeguard drives the model to develop specialized attention heads that explicitly handle uncertainty by generating “unknown” (Ding et al., 2025; Zhang et al., 2024a).

5.2.5 Answer Decision Mechanisms in VRU

In summary, the key heads in the later layers primarily function as determining the final output, indicating that the model transitions from viewpoint orientation perception to answer selection in the later layers. This finding supports the hypothesis in Section 5.2.2 that a pattern shift does occur in

Models	Train Speed	Tuned Params.	Spatial Intelligence				Generic Ability			
			VRUBench		SpinBench*		MMLU*		BBH*	
			Acc.	Δ	Acc.	Δ	Acc.	Δ	Acc.	Δ
Qwen2.5-VL-3B	-	-	37.6	-	46.7	-	62.5	-	25.5	-
+ Full SFT	10sam./sec.	3.0B	88.5	+50.9	46.2	-0.5	61.5	-1.0	19.7	-5.8
+ Selective SFT	18sam./sec.	0.03B	80.1	+42.5	47.0	+0.3	62.9	+0.4	26.2	+0.7
Qwen2.5-VL-7B	-	-	48.7	-	44.8	-	60.3	-	49.2	-
+ Full SFT	5sam./sec.	7.0B	96.3	+47.6	47.3	+2.5	55.6	-4.7	35.8	-13.4
+ Selective SFT	11sam./sec.	0.06B	78.7	+30.0	48.4	+3.6	60.3	+0.0	48.4	-0.8

Table 2: Overall performance of selective and full fine-tuning. * represents OOD dataset.

7 Conclusion

In this paper, we focus on a fundamental and critical ability in spatial intelligence: viewpoint rotation understanding (VRU). Firstly, we propose a synthesized dataset to benchmark the VRU capability of both VLMs and LLMs, which yields limited performance. Through layer-wise and head-wise interpretability analysis, we not only uncover the internal mechanisms about how models performs VRU without vision, but also identify a small subset of attention heads that play a key role in this process. Beyond interpretability, the enhanced performance by selectively fine-tuning the key heads further demonstrates the potential of our findings in model improvement. Our work serve as an initial mechanistic interpretability of spatial intelligence, laying a solid foundation for future studies on more complex tasks in spatial intelligence.

Limitations

This work conducts comprehensive interpretability studies and provides valuable insights for the spatial intelligence capability of current LLMs and VLMs. Despite these, there are still some limitations of our work, which could be summarized into the following two aspects. Firstly, LLMs and VLMs are demonstrated to be sensitive to the prompt phrasing, exploring the sensitivity in VRU task may yield additional findings. However, this falls beyond the primary scope of this paper, thus we leave it for future works. Secondly, due to computational resource constraints, the selective and full fine-tuning were conducted on models no larger than 7B. We hope that future research could extend it to larger-scale models.

Ethics Statement

This paper explores the mechanisms underlying a textual-spatial intelligence ability and proposes to improves it with selective fine-tuning. The 100 objects used in constructing the proposed dataset, VRUBench, are manually reviewed to ensure that they do not contain any personally identifying information or malicious content. Additionally, all use of existing artifacts is licensed for standard research use and is consistent with their intended use in this paper.

The inner mechanisms uncovered in this paper provide insight for understanding model behaviors in spatial intelligence. Nonetheless, the findings in this paper may be misused to maliciously manipulate model behavior at inference time, thus we emphasize the importance of monitoring LLMs’ behavior during inference. Furthermore, we do not see any other potential risks.

Acknowledgements

This work is supported by the grants from the National Natural Science Foundation of China (No. 62172044 and No. 62376130). The authors would like to thank the organizers of ACL 2026 and the reviewers for their helpful suggestions.

References

- Guillaume Alain and Yoshua Bengio. 2018. [Understanding intermediate layers using linear classifier probes](#). *Preprint*, arXiv:1610.01644.
- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang,

- Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibao Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. 2025a. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan 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. 2025b. [Qwen2.5-vl technical report](#). *CoRR*, abs/2502.13923.
- Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2024. [Dola: Decoding by contrasting layers improves factuality in large language models](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- James M. Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational Psychology Review*, 3(3):149–210.
- Zhikai Ding, Shiyu Ni, and Keping Bi. 2025. [Do LVLMs know what they know? a systematic study of knowledge boundary perception in LVLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 19832–19848, Suzhou, China. Association for Computational Linguistics.
- Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, Songwei Li, Yunke Zhang, Yuming Lin, Tong Li, Jingtao Ding, Chen Gao, Fengli Xu, and Yong Li. 2025. [A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science](#). *CoRR*, abs/2504.09848.
- Daniel Y. Fu, Tri Dao, Khaled Kamal Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. 2023. [Hungry hungry hippos: Towards language modeling with state space models](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Cheng Gao, Huimin Chen, Chaojun Xiao, Zhiyi Chen, Zhiyuan Liu, and Maosong Sun. 2025a. [H-neurons: On the existence, impact, and origin of hallucination-associated neurons in llms](#). *Preprint*, arXiv:2512.01797.
- Leo Gao, Achyuta Rajaram, Jacob Coxon, Soham V. Govande, Bowen Baker, and Dan Mossing. 2025b. [Weight-sparse transformers have interpretable circuits](#). *Preprint*, arXiv:2511.13653.
- Howard Gardner. 1983. *Frames of Mind: The Theory of Multiple Intelligences*, tenth anniversary, second paperback edition. Basic Books.
- Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. [Localizing model behavior with path patching](#). *CoRR*, abs/2304.05969.
- Google. 2025. Gemini 3 flash: frontier intelligence built for speed.
- Zhongbin Guo, Jiahe Liu, Yushan Li, Wenyu Gao, Zhen Yang, Chenzhi Li, Xinyue Zhang, and Ping Jian. 2025. [Beyond flatlands: Unlocking spatial intelligence by decoupling 3d reasoning from numerical regression](#). *Preprint*, arXiv:2511.11239.
- Zhongbin Guo, Zhen Yang, Yushan Li, Xinyue Zhang, Wenyu Gao, Jiacheng Wang, Chengzhi Li, Xiangrui Liu, and Ping Jian. 2026. [Can llms see without pixels? benchmarking spatial intelligence from textual descriptions](#). *Preprint*, arXiv:2601.03590.
- Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. [How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. [Training compute-optimal large language models](#). *CoRR*, abs/2203.15556.
- Chao Huang, Zefeng Zhang, Juewei Yue, Quangang Li, Chuang Zhang, and Tingwen Liu. 2025. [Safety alignment should be made more than just A few attention heads](#). *CoRR*, abs/2508.19697.
- Lei Jiang, Zixun Zhang, Yuting Zeng, Chunzhao Xie, Tongxuan Liu, Zhen Li, Lechao Cheng, and Xiaohua Xu. 2025. [DCP: Dual-cue pruning for efficient large vision-language models](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21202–21215, Suzhou, China. Association for Computational Linguistics.

- Tianjie Ju, Yijin Chen, Xinwei Yuan, Zhuosheng Zhang, Wei Du, Yubin Zheng, and Gongshen Liu. 2024a. [Investigating multi-hop factual shortcuts in knowledge editing of large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 8987–9001. Association for Computational Linguistics.
- Tianjie Ju, Zhenyu Shao, Bowen Wang, Yujia Chen, Zhuosheng Zhang, Hao Fei, Mong-Li Lee, Wynne Hsu, Sufeng Duan, and Gongshen Liu. 2025. [Probing then editing response personality of large language models](#). In *Second Conference on Language Modeling*.
- Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024b. [How large language models encode context knowledge? A layer-wise probing study](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 8235–8246. ELRA and ICCL.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *CoRR*, abs/2001.08361.
- Diederik P. Kingma and Jimmy Ba. 2015. [Adam: A method for stochastic optimization](#). In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability.
- Chengzhi Li, Heyan Huang, Ping Jian, Zhen Yang, and Yaning Tian. 2025a. [Improving temporal understanding logic consistency in video-language models via attention enhancement](#). *Preprint*, arXiv:2510.08138.
- Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqiao Zhang, Yongliang Shen, Weiming Lu, and Yueting Zhuang. 2025b. [Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models](#). *CoRR*, abs/2505.21500.
- Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024. [Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the stepgame benchmark](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 18500–18507. AAAI Press.
- Kenneth Li, Oam Patel, Fernanda B. Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. [Inference-time intervention: Eliciting truthful answers from a language model](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.
- Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M. de Melo, Alan L. Yuille, and Jieneng Chen. 2024. [3dsrbench: A comprehensive 3d spatial reasoning benchmark](#). *CoRR*, abs/2412.07825.
- Andreas Madsen, Siva Reddy, and Sarath Chandar. 2023. [Post-hoc interpretability for neural NLP: A survey](#). *ACM Comput. Surv.*, 55(8):155:1–155:42.
- Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, and Peter Henderson. 2025. [Safety alignment should be made more than just a few tokens deep](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Krähenbühl, and Vladlen Koltun. 2025. [Does spatial cognition emerge in frontier models?](#) In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. [Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts](#). In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11321–11329. AAAI Press.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, and et al. 2023. [Beyond the imitation game: Quantifying and extrapolating the capabilities of language models](#). *Trans. Mach. Learn. Res.*, 2023.
- Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2024. [Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning](#). *CoRR*, abs/2410.16162.
- Llama Team. 2024. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Curt Tigges, Michael Hanna, Qinan Yu, and Stella Biderman. 2024. [LLM circuit analyses are consistent across training and scale](#). In *Advances in Neural*

- Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024.*
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. [Llama 2: Open foundation and finetuned chat models](#). *CoRR*, abs/2307.09288.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.
- Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. [MLLM can see? dynamic correction decoding for hallucination mitigation](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. 2024. [Is A picture worth A thousand words? delving into spatial reasoning for vision language models](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. [Interpretability in the wild: a circuit for indirect object identification in GPT-2 small](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Boyi Wei, Kaixuan Huang, Yangsibo Huang, Tinghao Xie, Xiangyu Qi, Mengzhou Xia, Prateek Mittal, Mengdi Wang, and Peter Henderson. 2024. [Assessing the brittleness of safety alignment via pruning and low-rank modifications](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-thought prompting elicits reasoning in large language models](#). In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.
- Qiyao Xue, Weichen Liu, Shiqi Wang, Haoming Wang, Yuyang Wu, and Wei Gao. 2025. [Reasoning path and latent state analysis for multi-view visual spatial reasoning: A cognitive science perspective](#). *Preprint*, arXiv:2512.02340.
- Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. 2024. [Evaluating spatial understanding of large language models](#). *Trans. Mach. Learn. Res.*, 2024.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025a. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. 2025b. [Thinking in space: How multimodal large language models see, remember, and recall spaces](#). In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pages 10632–10643. Computer Vision Foundation / IEEE.

- Zhen Yang, Ping Jian, and Chengzhi Li. 2025c. [Option symbol matters: Investigating and mitigating multiple-choice option symbol bias of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 1902–1917. Association for Computational Linguistics.
- Fangcong Yin, Xi Ye, and Greg Durrett. 2024. [Lofit: Localized fine-tuning on LLM representations](#). In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.
- Le Yu, Bowen Yu, Haiyang Yu, Fei Huang, and Yongbin Li. 2024. [Language models are super mario: Absorb-ing abilities from homologous models as a free lunch](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Zeping Yu and Sophia Ananiadou. 2024. [Interpreting arithmetic mechanism in large language models through comparative neuron analysis](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 3293–3306. Association for Computational Linguistics.
- Zeping Yu and Sophia Ananiadou. 2025. [Understanding and mitigating gender bias in llms via interpretable neuron editing](#). *CoRR*, abs/2501.14457.
- Fred Zhang and Neel Nanda. 2024. [Towards best practices of activation patching in language models: Metrics and methods](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. [R-tuning: Instructing large language models to say ‘i don’t know’](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 7113–7139. Association for Computational Linguistics.
- Hong Zhang, Jiayu Wan, Jing Zhang, Ding Yuan, Xu-liang Li, and Yifan Yang. 2025a. [P2ftrack: Multi-object tracking with motion prior and feature posterior](#). *ACM Trans. Multim. Comput. Commun. Appl.*, 21(1):22:1–22:22.
- Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. 2025b. [From flatland to space: Teaching vision-language models to perceive and reason in 3d](#). *CoRR*, abs/2503.22976.
- Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. 2024b. [Interpreting and improving large language models in arithmetic calculation](#). In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.
- Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Junqi Zhao, Allison Koenecke, Boyang Li, and Lu Wang. 2025c. [SPHERE: unveiling spatial blind spots in vision-language models through hierarchical evaluation](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11591–11609. Association for Computational Linguistics.
- Yuyou Zhang, Radu Corcodel, Chiori Hori, Anoop Cherian, and Ding Zhao. 2025d. [Spinbench: Perspective and rotation as a lens on spatial reasoning in vlms](#). *CoRR*, abs/2509.25390.
- Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. [On the role of attention heads in large language model safety](#). In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.
- Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and et al. 2025. [Internv13: Exploring advanced training and test-time recipes for open-source multimodal models](#). *CoRR*, abs/2504.10479.

A Samples From VRUBench

2-step rotation

content:

You are standing inside a room and can observe objects from your current viewpoint. You will take actions to rotate your viewpoint to the left or right, after which both your viewpoint and the objects in view changes accordingly. The actions taken are specified by "Action:", and the corresponding objects you see are specified by "Observation:". After taking the final action, predict the resulting observation. If the observation after the final action cannot be determined based on previous observations, output unknown.

Initial Observation: avocado
Action: Turn to the right by 270 degrees
Observation: router
Action: Turn to the left by 90 degrees
Observation:
ground-truth: unknown
Qwen2.5-VL-7B: unknown

3-step rotation

content:

< task description >
Initial Observation: receiver
Action: Turn to the right by 180 degrees
Observation: clock
Action: Turn to the left by 180 degrees
Observation: receiver
Action: Turn to the right by 0 degrees
Observation:
ground-truth: receiver
Qwen2.5-VL-7B: clock

4-step rotation

content:

< task description >
Initial Observation: plate
Action: Turn to the left by 90 degrees
Observation: faucet
Action: Turn to the right by 180 degrees
Observation: wardrobe
Action: Turn to the left by 360 degrees
Observation: wardrobe

Action: Turn to the left by 90 degrees
Observation:
ground-truth: plate
Qwen2.5-VL-7B: unknown

5-step rotation

content:

< task description >
Initial Observation: avocado
Action: Turn to the right by 270 degrees
Observation: router
Action: Turn to the left by 90 degrees
Observation: remote
Action: Turn to the right by 0 degrees
Observation: remote
Action: Turn to the right by 180 degrees
Observation: avocado
Action: Turn to the right by 270 degrees
Observation:
ground-truth: router
Qwen2.5-VL-7B: router

B Evaluation Details

B.1 System Prompt

For the prompt template, we utilize the default system prompts of LLaMA, Qwen models. For instance, the final messages for Qwen2.5-VL-7B are as follows:

```
"role": "system"  
"content":  
"You are a helpful assistant"  
"role": "user"  
content:  
< task description >  
< Actions and Observations >
```

B.2 Implementation Details of Reasoning Mode

When the reasoning mode of Qwen3 and Qwen3-VL is enabled (denoted as "-thinking" in Table 1), the hyperparameter `thinking_budget` is set to 1024, indicating that Qwen3 models generate their reasoning content of up to 1024 tokens between the `<think>` and `</think>` tokens before producing the final answer. As specified in Section 3.1, in order to accurately extract the final answer generated by models, we instruct Qwen3 and Qwen3-VL to output their answers between `<ans></ans>` tags:

Prompt for think-then-answer

```
"role": "system"
"content":
"You are a helpful assistant"
"role": "user"
content:
You are standing inside a room and
can observe objects from your current
viewpoint. You will take actions to rotate
your viewpoint to the left or right, after
which both your viewpoint and the objects
in view changes accordingly. The actions
taken are specified by "Action:", and the
corresponding objects you see are specified
by "Observation:". After taking the final
action, predict the resulting observation.
If the observation after the final action
cannot be determined based on previous
observations, output unknown. Output
your final answer wrapped in <ans> and
</ans> tags.

< Actions and Observations >
```

We evaluate most of the models in Table 1 using the official Qwen API, except Qwen3-VL-4B (-thinking), for which no official API is available. When running Qwen3-VL-4B locally, we modify the prompt as follows, using the official implementation designed to elicit the reasoning capability of Qwen3 models (Yang et al., 2025a; Bai et al., 2025a):

Prompt for Qwen3-VL-4B

```
"role": "system"
"content":
"You are a helpful assistant"
"role": "user"
content:
/think You are standing inside a room ...
Output your answer between <ans></ans>
tags.

< Actions and Observations >
```

To control the length of thinking, we reproduced the implementation related to the `thinking_budget` hyperparameter, where the model is forced to terminate the thinking phase by inserting the `</think>` token to generated content once the thought content reaches a length of 1024 tokens.

B.3 Reasoning Examples

Figure 12-13 present the results when thinking mode is enabled. It can be observed that the models extract information from input and try to solve the problems from a mathematical perspective, ultimately leading to the improved performance in Table 1.

B.4 Human Performance

The human performance in Table 1 are obtained from manual assessments conducted by twenty volunteers who are master students. We randomly sample 10,000 cases from the proposed VRUBench and invite a pool of 20 human evaluators. To ensure reliability, every single case was independently evaluated by two different annotators. That is, each annotator is responsible for 1,000 cases. We evaluated the consistency of the human responses and achieved a Krippendorff's Alpha (Krippendorff, 2011) of 1.0, which strongly demonstrates the inter-annotator reliability.

C Layer-wise Probing

C.1 Dataset for Probing

For probing *direction* (the probing label is shown after ▷):

Labels for Probing Direction

```
content:
< task description >
Initial Observation: avocado
Action: Turn to the right by 270 degrees
▷ right
Observation: router
Action: Turn to the left by 90 degrees
▷ left
Observation:
```

For probing *angle*:

Labels for Probing Angle

```
content:
< task description >
Initial Observation: avocado
Action: Turn to the right by 270 degrees
▷ 270°
Observation: router
Action: Turn to the left by 90 degrees
▷ 90°
Observation:
```

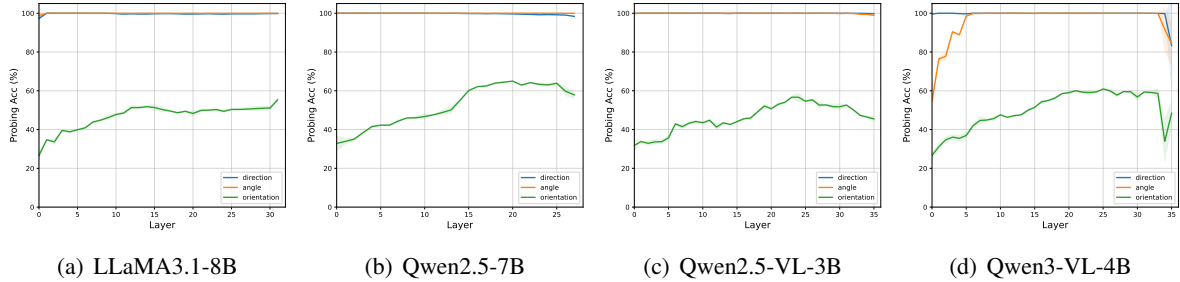


Figure 5: The probing results of other LLMs and VLMs.

For probing *orientation*:

Labels for Probing Orientation

content:

< task description >

Initial Observation: avocado

Action: Turn to the right by 270 degrees

▷ 270°

Observation: router

Action: Turn to the left by 90 degrees

▷ 180°

Observation:

When training the probes, direction prediction is formulated as a binary classification task (left/right), angle as a 5-way classification task (0° / 90° / 180° / 270° / 360°), and orientation as a 4-way classification task (the four cardinal directions, indexed with 0° / 90° / 180° / 270°). We train each probe three times and report the mean and standard deviation of the results in Figure 2 and 5.

C.2 Probing Architecture

In the main experiments of probing, we utilize linear classifiers as probes to explore whether LLMs and VLMs encode direction/angle/orientation. Here, we provide further clarification about the choice of linear probes rather than non-linear ones.

Theoretically, linear probes impose a minimal inductive bias and therefore serve as a more faithful diagnostic of what is already encoded in the representation (Alain and Bengio, 2018). In contrast, non-linear probes may introduce extra expressive capacity that allows the probe to solve the task independently, thereby confounding the interpretation of probing accuracy as evidence of representational content.

Empirically, in our preliminary experiments, we find that although non-linear probes could improve the overall probing accuracy across various LLMs

and VLMs, the conclusions drawn from the probing keep consistent with linear probes. For instance, the probing accuracy of LLaMA2-7B and Qwen2.5-VL-7B in Figure 2 would increase simultaneously when probing with MLPs. However, the overall accuracy of Qwen2.5-VL-7B still outperforms LLaMA2-7B, and increases across layers 1-20 while declining over layers 21-28, which aligns with the findings using linear probes (Section 5.1). That is, non-linear probes do not affect the conclusion of probing, which is also consistent with findings in previous study (Ju et al., 2025).

Therefore, based on aforementioned theoretical and empirical analysis, we ultimately opt to use linear probes during the main experiments.

C.3 Probing Results on Other LLMs and VLMs

The probing results of other LLMs (LLaMA3.1-8B and Qwen2.5-7B) and VLMs (Qwen2.5-VL-3B and Qwen3-VL-4B) are shown in Figure 5, which exhibit consistent observations with Figure 2.

D Details about Path Patching

Transformers (Vaswani et al., 2017) can be conceptualized such that the residual connections form the main information pathway (the residual stream, illustrated in Figure 6), while the attention modules and MLPs act as bypass pathways that add their computations to the residual stream. Path patching is a causal intervention technique designed to identify and quantify the contribution of specific internal computational pathways to a model’s final prediction (Goldowsky-Dill et al., 2023; Wang et al., 2023). The core idea is to treat internal components (e.g., attention heads, MLP blocks, or residual streams) as nodes in a computational graph and to test whether information transmitted along two nodes (Sender \rightarrow Receiver) is causally responsible for the model’s behavior.

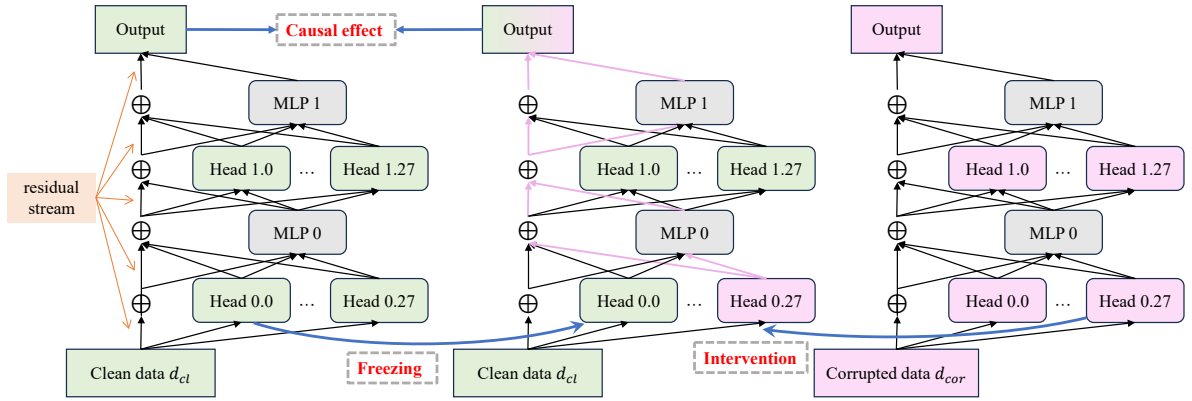


Figure 6: An illustration of the practical application of *path patching*.

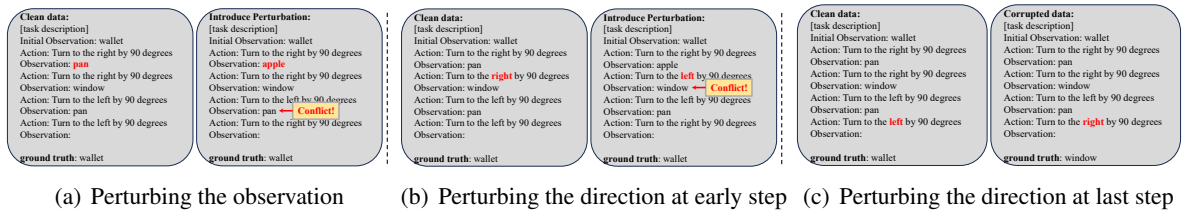


Figure 7: Perturbing the observation (a) or rotation direction at early step (b) could introduce logical inconsistencies in the textual description. Hence, we perturb the rotation direction at the last step to avoid such conflicts and construct the clean-corrupted data pairs (c).

Formally, given a clean input d_{cl} and a corrupted input d_{cor} , path patching performs controlled interventions on selected intermediate activations while holding all other components fixed, as depicted in Figure 6. Let h denote a Sender node (e.g., an attention head) and y denote a Receiver node (e.g., the output logits). The goal is to determine whether information flowing along the path $h \rightarrow y$ is necessary for producing the observed output.

The procedure consists of three main steps:

Activation collection. The model is first run on both d_{cl} and d_{cor} , and the activations of all relevant internal components are recorded. These cached activations serve as sources for later interventions.

Selective intervention. A hard intervention is applied to the sender node h by replacing its activation on d_{cl} with the activation on d_{cor} , where the effect will be further propagated to the Receiver node along with a set of computational paths. Simultaneously, all other components not on the tested path are frozen to their activations under d_{cl} . This isolates the causal effect of the sender node by ensuring that the only information differing between the two runs flows through the designated path.

Effect measurement. The modified activations are propagated forward through the network to produce new output logits. The causal importance of the path $h \rightarrow y$ is then quantified by the causal effect metrics (e.g., Eq.6). If intervening on h induces a substantial change in the output while all other components remain fixed, this provides evidence that the path $h \rightarrow y$ is causally important for the model’s computation on the task.

E More Results of Path Patching

We further report the results of path patching on LLaMA2-7B-chat, Qwen2.5-7B and Qwen2.5-VL-3B in Figure 8. It can be observed that these models with different model families and sizes exhibit similar path patching results presented in Section 5.2.2, where key heads are distributed sparsely and in the middle-to-upper layers.

For the construction of corrupted data, path patching generally aims to apply minimal perturbation to the clean input, and ideally, such perturbations should preserve the token length to ensure comparability of internal activations. In our dataset, two elements are theoretically perturbable: the rotation (angle and direction) and the object being observed. However, rotation angles such as 0, 90

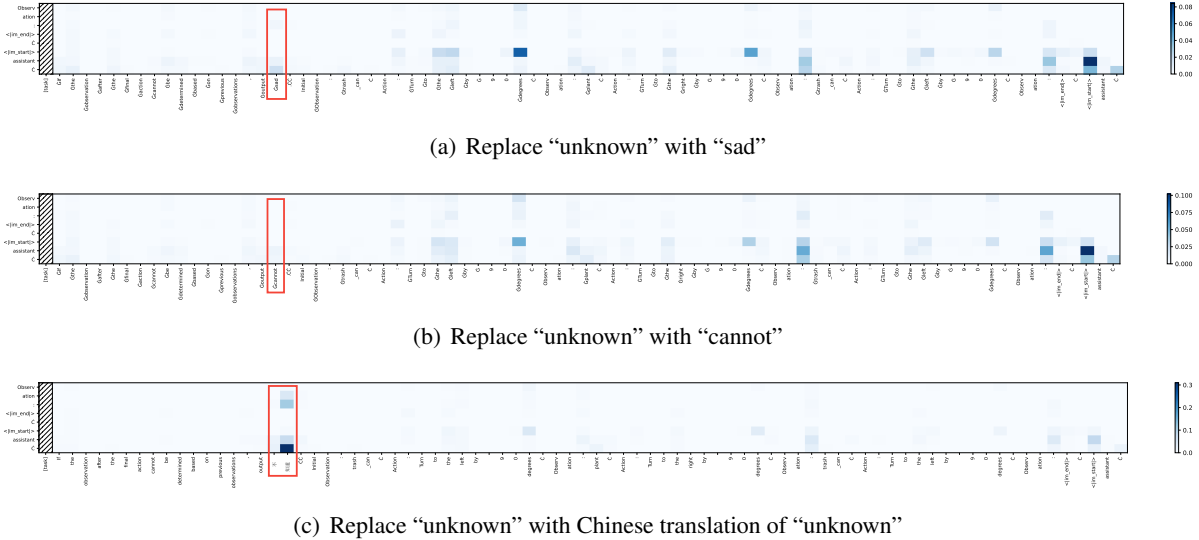


Figure 10: Attention pattern cases of unknown head when the words “unknown” is replaced with other alternative tokens. The head no longer exhibits strong attention to semantically unrelated replacements such as “sad” or “cannot.” However, when “unknown” is replaced with its Chinese translation (“不知道”), which has the same semantic, the head again shows a similarly high level of attention to “不知道”.

H Unknown Head with Alternative Token

As discussed in Section 5.2.4, we instruct the models to output other words instead of “unknown” when the final observation cannot be determined by previous observations, in order to comprehensively understand the mechanisms of the Unknown Head 27.14. As shown in Figure 10, when the word “unknown” is replaced with semantically unrelated tokens, the head 27.14 does not keep the same particular attention on the alternative tokens, but keeps a high attention to “不知道”, reasonably demonstrating the function of the unknown head as discussed in Section 5.2.4. That is, the head indeed encodes a cautious response strategy, namely, acknowledging that the answer is unknown rather than simply attending to the token when the answer cannot be determined based on observation history.

I Details of Selective Fine-tuning

I.1 Experimental Details

Algorithm Firstly, we split the parameter matrices $W_{K/Q/V/O}^i$ into H blocks, which is similar to head ablation in Appendix F. To achieve selective fine-tuning, only the parameter matrices $W_{K/Q/V/O}^{i,j}$ associated with key attention heads i,j is set to tunable, while freezing all other parameters. Following Yu et al. (2024); Zhang et al. (2024b), we rescale the gradients by a factor of $\frac{H}{h}$, where H denotes the total number of heads in each

Methods	Viewpoint	OR	FR	OI
Full SFT	47.3	51.06	31.91	86.49
Selective SFT	48.4	53.19	43.62	91.43
Δ	1.10	2.13	11.70	5.41

Table 3: The results of Qwen2.5-VL-7B on various sub-tasks of SpinBench, which exhibits consistent improvements. OR, FR, OI denote Object Rotation, Face Rotation, and Object Identity respectively.

layer and h denotes the number of heads updated in that layer.

Dataset As specified in Section 6, the construction of training set for selective and full SFT follows the pipeline in Section 3.2, resulting in a training set comprising 19,641 samples. It consists of 4,961 2-step samples, 4,954 3-step samples, 4,996 4-step samples and 5,000 5-step samples, which is distinct from the test set VRUBench. As for another spatial intelligence benchmark SpinBench, we extract the tasks related to viewpoint as the evaluation set, which matches the research scope, i.e., viewpoint rotation understanding, in this work. Additionally, Table 3 shows the results on extra sub-tasks of SpinBench. The selective SFT achieves consistent improvements not only on the viewpoint subset but also across other visual-spatial subtasks, which demonstrates the robustness of OOD gains

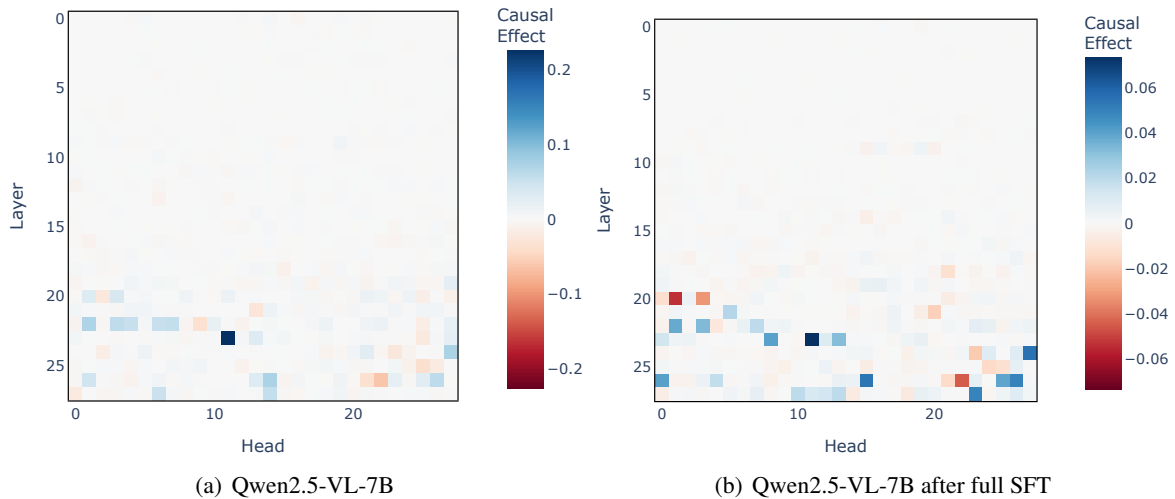


Figure 11: The results of path patching before and after full fine-tuning.

and further verifies Takeaway II.

Hyperparameters During selective and full SFT, we adopt Adam (Kingma and Ba, 2015) as the optimizer with an learning rate of 2×10^{-5} . Models are trained with a batch size of 32, warm up ratio of 0.02 and weight decay of 0.1 for one epoch.

I.2 Further Discussion about Table 2

It is noteworthy that under selective SFT, although both Qwen2.5-VL-3B and Qwen2.5-VL-7B are trained with top-32 heads, Qwen2.5-VL-3B with smaller size surprisingly outperforms the 7B model after selective SFT. The reason could be that the 7B model consists of $28 \times 28 = 784$ heads while 3B model only contains $36 \times 16 = 576$ heads, thus 32 fine-tuned attention heads account for a lower percentage of the total heads in the 7B model. Therefore, the results do not contradict the scaling law (Kaplan et al., 2020; Hoffmann et al., 2022). Another evidence is that the 7B model continues to outperform the 3B model under full fine-tuning.

Furthermore, we also take efforts in analyzing the catastrophic forgetting of generic ability after full fine-tuning in Table 2. As shown in Figure 11, we compare the causal effect (i.e., the importance in VRU) of each head before and after full fine-tuning. It can be observed that more heads emerge as key heads for VRU after full SFT, and some heads with relatively low importance also become prominent after full SFT. Prior studies have shown that a single attention head can simultaneously contribute to both general capabilities and task-specific functions (e.g., VRU) (Wei et al., 2024). Therefore, the importance changes observed in Figure 11 in-

dicates that the model transforms general-purpose heads into ones specialized for VRU, thereby resulting the extremely high performance in VRU. On the other hand, it also explains the degradation in general capabilities after full SFT.

J Computational Budget

All the experiments mentioned in this paper, including the probing analysis, path patching, selective fine-tuning, are conducted on 2 NVIDIA A800 80GB GPUs.

1st Reasoning Example

Input:

Initial Observation: window

Action: Turn to the right by 180 degrees

Observation: mirror

Action: Turn to the left by 180 degrees

Observation: window

Action: Turn to the right by 180 degrees

Observation:

(GT: mirror)

Qwen3-8B: (pred: mirror) ✓

Okay, let's try to figure out what the next observation should be...

after two turns (right and left), you're back to the original position. So if you turn right again, maybe it's the same as the first turn...

the previous time when you turned right from window, you saw mirror. So this seems like the same scenario. Therefore, the observation should be **mirror** again.

Qwen3-VL-8B: (pred: mirror) ✓

Okay, let's try to figure this out step by step. Let me read through the initial problem again.

...(restate the question)...

Let's model the direction as angles...So a right turn of 180 degrees would be rotating 180 degrees clockwise. Similarly, left is counterclockwise.

So, starting direction: 0 degrees (east).

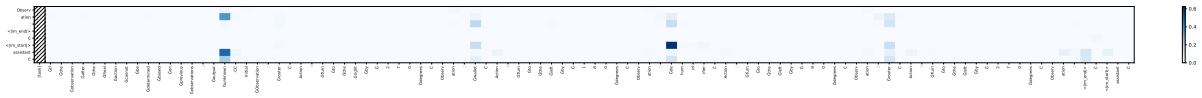
Action 1: Turn right 180 degrees $\rightarrow 0 + 180 = 180$ degrees (west). Observation: mirror.

Action 2: Turn left 180 degrees. Left turn is counterclockwise, so subtract 180. $180 - 180 = 0$ degrees (east). Observation: window.

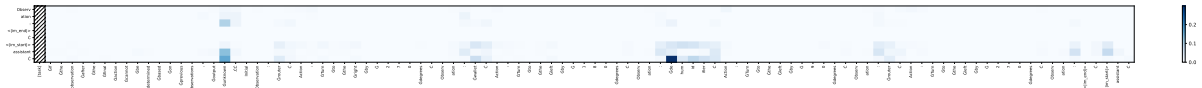
Action 3: Turn right 180 degrees. $0 + 180 = 180$ degrees (west). Observation: **mirror**.

Yes, same!

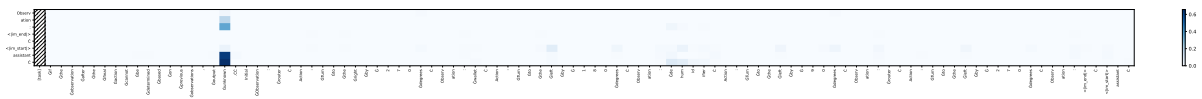
Figure 12: 1st reasoning example.



(a) Proposal Head 22.1

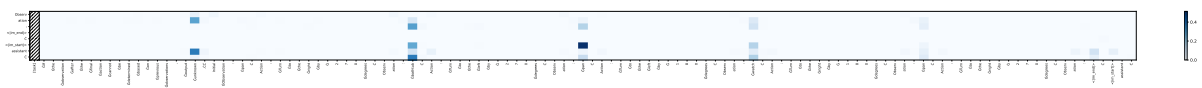


(b) Answer Decision Head 26.14

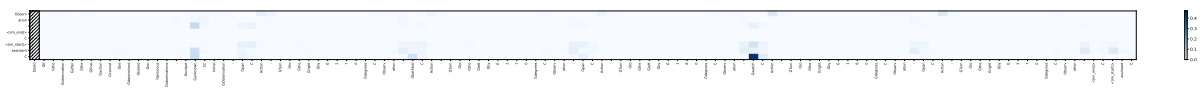


(c) Unknown Head 27.14

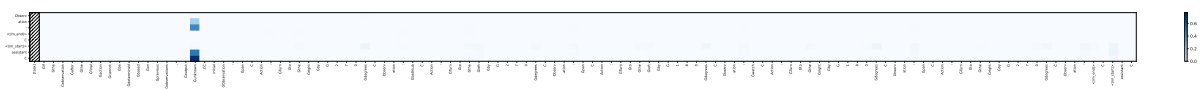
Figure 15: **Example 2** (attention pattern). Model output: *dehumidifier*.



(a) Proposal Head 22.1



(b) Answer Decision Head 26.14



(c) Unknown Head 27.14

Figure 16: **Example 3** (attention pattern). Model output: *bathtub*.