From Human Cognition to Neural Activations: Probing the Computational Primitives of Spatial Reasoning in LLMs

Anonymous ACL submission

Abstract

As spatial intelligence becomes an increasingly important capability for foundation models, it remains unclear whether large language models' (LLMs) performance on spatial reasoning benchmarks reflects structured internal spatial representations or reliance on linguistic heuristics. We address this question from a mechanistic perspective by examining how spatial information is internally represented and used. Drawing on computational theories of human spatial cognition, we decompose spatial reasoning into three primitives, relational composition, representational transformation, and stateful spatial updating, and design controlled task families for each. We evaluate multilingual LLMs in English, Chinese, and Arabic under single pass inference, and analyze internal representations using linear probing, sparse autoencoder based feature analysis, and causal interventions. We find that task relevant spatial information is encoded in intermediate layers and can causally influence behavior, but these representations are transient, fragmented across task families, and weakly integrated into final predictions. Cross linguistic analysis further reveals mechanistic degeneracy, where similar behavioral performance arises from distinct internal pathways. Overall, our results suggest that current LLMs exhibit limited and context dependent spatial representations rather than robust, general purpose spatial reasoning, highlighting the need for mechanistic evaluation beyond benchmark accuracy. [1]

1 Introduction

Large language models (LLMs) and vision–language models (VLMs) have achieved rapid progress in reasoning, planning, and interactive decision making, and are increasingly deployed in settings that require spatial understanding, such as instruction following, navigation, embodied interaction, and robotics(Wang et al., 2024a; Brohan et al., 2023; Driess et al., 2023; Guo et al., 2024). At the same time, the study of world models has emphasized the importance of structured internal representations for prediction and control(Yi et al., 2018). These developments raise a fundamental question: do foundation models acquire genuine spatial representations, or do they succeed through surface-level linguistic regularities?(Yamada et al., 2023; Du et al., 2024; Li et al., 2024a; Du et al., 2024)

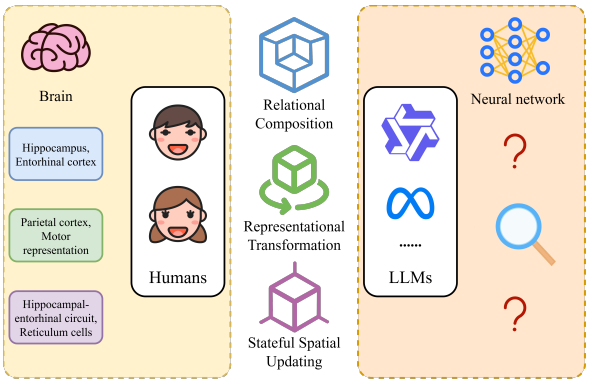

Figure 1: From human spatial cognition to spatial representations in large language models.

Recent work has begun to address this question using spatial reasoning benchmarks, reporting steady performance gains with model scale and instruction tuning(Yamada et al., 2023; Li et al., 2024a; Wei et al., 2021). However, benchmark accuracy alone provides limited insight into underlying mechanisms. Correct input–output behavior may arise from linguistic heuristics, memorization, or prompt-induced reasoning strategies, making it unclear whether models rely on structured spatial representations or shallow correlations(Holliday et al., 2014; Geirhos et al., 2020; Xie et al., 2024; Wei et al., 2022).

This stands in contrast to research on human spatial cognition, where spatial ability is characterized

[1] Our code is available at <It will be published after the paper is accepted.>

not only by behavior but by well-studied internal mechanisms, such as cognitive maps, mental rotation, and path integration. These representations explain generalization, robustness, and systematic errors. By comparison, most studies of spatial reasoning in large neural models do not directly examine internal representations, leaving the basis of their spatial behavior largely opaque(Yamada et al., 2023; Li et al., 2024a; Hewitt and Manning, 2019; Elazar et al., 2020).

The lack of mechanistic understanding has both practical and conceptual implications. As spatial reasoning becomes increasingly embedded in downstream systems, limited insight into internal representations hinders robustness assessment, interpretability, and principled comparison with biological cognition(Yamins and DiCarlo, 2016). Without representational evidence, behavioral similarities between models and humans remain speculative(Firestone, 2020).

In this work, we argue that progress on spatial reasoning in foundation models requires moving beyond benchmark-centric evaluation toward mechanistic analysis. Rather than asking only whether a model produces correct answers, we investigate whether it develops structured, compositional representations that play a functional role in behavior. Drawing inspiration from cognitive science, we decompose spatial ability into a small set of core computational primitives and design controlled task families that isolate these primitives under standard inference settings, without eliciting explicit reasoning traces.

We introduce three families of spatial tasks: relational spatial reasoning, perspective transformation, and spatial program execution. To disentangle spatial representations from linguistic form, we construct parallel tasks in English, Chinese, and Arabic. We analyze both behavior and internal representations using probing, sparse autoencoder–based feature analysis, and causal interventions. This framework enables us to localize spatial information within models, characterize representational differences across task families, and assess their dependence on language. By grounding spatial evaluation in cognitive theory and mechanistic interpretability, our study clarifies the nature and limits of spatial ability in foundation models, and provides a foundation for moving beyond benchmark accuracy.

2 Related Work

2.1 Spatial Cognition and Computational Mechanisms

Spatial cognition has long been studied in cognitive psychology and neuroscience as a core component of human intelligence. Classical psychometric frameworks distinguish abilities such as spatial perception, mental rotation, and visualization(Ekstrom et al., 1976; Shepard and Metzler, 1971), while later work emphasizes that spatial behavior arises from distinct internal representations and computational mechanisms rather than a single faculty(Eckardt, 1980; Burgess, 2008).

Research on cognitive maps shows that humans and animals construct structured, allocentric representations that support relational inference beyond immediate perception(Tolman, 1948). Studies of mental rotation and perspective taking demonstrate that these representations can undergo continuous geometric transformations(Shepard and Metzler, 1971), while navigation research highlights mechanisms for maintaining and updating spatial state over time(Etienne and Jeffery, 2004). Together, these findings motivate decomposing spatial cognition into a small set of core computational primitives. Our work adopts this computational perspective to guide task design, rather than directly replicating psychometric categories.

2.2 Spatial Reasoning in NLP and Large Language Models

Spatial reasoning in NLP has traditionally focused on interpreting spatial relations expressed in language, such as prepositions, relative positions, and instructions. With the rise of large language models (LLMs), recent work has increasingly relied on benchmarks that evaluate spatial reasoning via output accuracy(Suzgun et al., 2022).

These studies show that LLMs can achieve nontrivial performance on text-based spatial tasks, especially with chain-of-thought prompting or explicit intermediate reasoning. However, they also report limitations such as sensitivity to surface form variation, degraded performance under increased compositionality, and poor generalization(Wei et al., 2022; Li et al., 2024b). Despite these observations, most work remains behavioral, offering limited insight into whether correct outputs reflect structured spatial representations or shallow linguistic heuristics(Geirhos et al., 2020). In contrast, our work targets internal representations

rather than output accuracy alone.

2.3 Spatial Reasoning Benchmarks for Large Language Models

A growing body of work proposes benchmarks to evaluate spatial reasoning in LLMs, focusing on spatial relations, compositional descriptions, and navigation-like inference in text(Yamada et al., 2023; Rizvi et al., 2024). Synthetic and semi-natural datasets such as StepGame(Shi et al., 2022) and SpatialEval(Wang et al., 2024b) are widely used to probe multi-step spatial inference, typically using accuracy-based evaluation.

Across benchmarks, recent studies report improved performance with prompting strategies, but also reveal substantial weaknesses, including sensitivity to paraphrasing, compositional complexity, and distribution shift(Sharma, 2023). Similar patterns appear in multimodal and vision-language models, where explicit reference structures or visualization-style reasoning can boost performance but are often tightly coupled to the evaluation setup(Wu et al., 2024; Liao et al., 2024). Surveys highlight a persistent gap between behavioral success and evidence for structured, compositional spatial representations(Liu et al., 2025; Zheng et al., 2025). Overall, existing benchmarks largely emphasize output behavior, providing limited insight into underlying representations.

2.4 Probing and Mechanistic Analysis of LLMs

A parallel line of research investigates what information is encoded inside neural language models(Maennel et al., 2020; Krause and Reimann, 2024; Yamada et al., 2023; Nanda and Bloom, 2022; Bloom et al., 2024). Linear probing tests whether variables can be decoded from hidden states, while more recent work analyzes representation geometry or identifies interpretable features using sparse autoencoders(Hewitt and Manning, 2019; Hewitt and Liang, 2019; Gupta et al., 2023; Raghu et al., 2017; Yan et al., 2024). These methods also enable causal interventions that modify internal representations and measure behavioral effects(Meng et al., 2022).

While mechanistic analysis has been applied to many linguistic and semantic phenomena, its application to spatial reasoning remains limited(Olsson et al., 2022; Elhage et al., 2022). Our work builds on these interpretability tools to analyze spatial representations in LLMs, aiming to move beyond correlational evidence toward functional and causal understanding.

In summary, prior work on spatial reasoning in LLMs has largely focused on benchmark-based behavioral evaluation, while mechanistic interpretability studies have rarely targeted spatial cognition. Our work bridges these lines by grounding task design in computational theories of spatial cognition and applying mechanistic analysis to probe whether LLMs implement core spatial primitives.

3 Task Taxonomy: A Computational Decomposition of Spatial Ability

To distinguish genuine spatial computation from surface-level linguistic pattern matching in large language models, we introduce a task taxonomy grounded in a computational decomposition of spatial ability. Rather than organizing tasks by domain or surface form, we classify them by the minimal computational primitives required for correct inference, drawing on established findings in human spatial cognition.

We identify three irreducible components of spatial computation: (1) relational composition, (2) representational transformation, and (3) stateful spatial updating. Each component defines one task family, forming a compact and mechanistically accessible test suite for probing internal spatial representations beyond linguistic heuristics. Figure 2 summarizes the three families with representative examples.

3.1 Design Principles

All task families follow four shared principles. **Abstraction:** tasks use abstract entities and relations, minimizing reliance on world knowledge. **Compositionality:** solving each task requires integrating multiple spatial constraints or operations. **Parameterizability:** difficulty is systematically controlled via factors such as entity count, steps, or dimensionality. **Mechanistic Accessibility:** intermediate spatial variables are explicitly defined, enabling probing and causal intervention.

All tasks are evaluated under standard single-pass inference, without eliciting explicit reasoning traces, and are generated using controlled rule-based procedures to ensure computational equivalence across tasks and languages (Appendix A.1).

3.2 Relational Spatial Reasoning

This task family targets relational composition, requiring models to construct a globally consistent

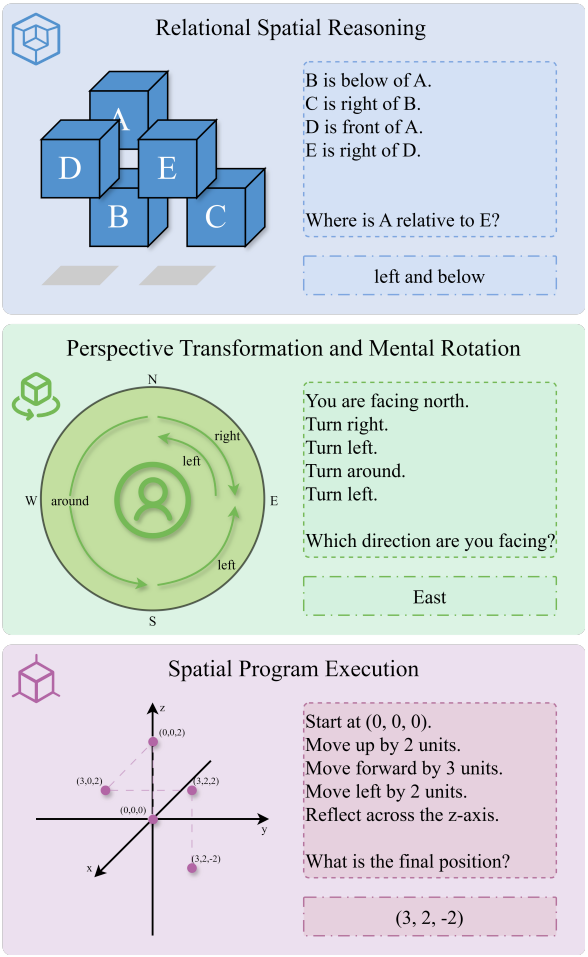

Figure 2: Illustration of the three spatial task families proposed in this work, with examples shown in English for clarity.

spatial structure from multiple pairwise relations (e.g., *A left of B*, *B above C*) and infer relations between indirectly connected entities. No metric information or procedural sequence is provided; the challenge lies entirely in relational structure building. We include both 2D and 3D variants, probing whether models form structured spatial representations rather than relying on memorized linguistic templates.

3.3 Perspective Transformation and Mental Rotation

This family isolates representational transformation. Given an initial spatial configuration or reference frame, models must apply one or more global transformations—such as rotations, reflections, or viewpoint changes—and report the resulting orientation or relation. Unlike relational reasoning, structure is assumed and must be transformed while preserving internal consistency. Both self-centered and multi-agent perspective-taking variants are included, directly probing equivariance under geometric operations.

3.4 Spatial Program Execution

The third family targets stateful spatial updating. Each instance specifies an initial position and a sequence of movement or transformation commands; the model must compute the final position after executing all steps. Performance depends on maintaining and updating a latent spatial state over time, making errors cumulative and diagnostic of state-tracking failures. This family abstracts the computational core of navigation and path integration and is especially amenable to causal analysis via intervention on intermediate states.

3.5 Alignment with Human Spatial Abilities

Although computationally motivated, the taxonomy aligns naturally with distinctions in human spatial cognition: relational reasoning with cognitive maps and spatial visualization, perspective transformation with mental rotation, and spatial program execution with dynamic updating in navigation. This alignment serves as an interpretive reference rather than a claim of equivalence.

3.6 Cross-Linguistic Task Construction

To disentangle spatial computation from language-specific cues, all task families are independently constructed in English, Chinese, and Arabic. Rather than direct translation, we ensure computational equivalence while allowing natural variation in surface realization, treating language as a controlled variable for analyzing whether spatial representations are shared or language-dependent.

4 Methodology and Experiments

4.1 Experimental Setup

Models. We evaluate spatial reasoning in two multilingual LLM families, Qwen2.5-7B(Yang et al., 2024) and Llama-3-8(Dubey et al., 2024), including both base and instruction-tuned variants. All experiments use publicly available checkpoints without task-specific fine-tuning, ensuring that observed spatial behaviors reflect general representations rather than adaptation.

Inference Protocol. All tasks are evaluated using standard single-pass inference without chain-of-thought prompting or explicit reasoning traces. Models select answers in a multiple-choice format

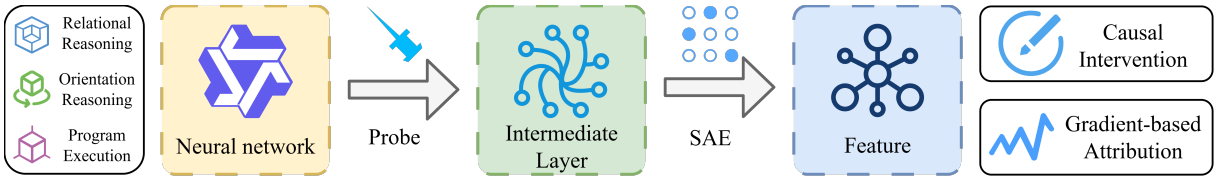

Figure 3: Overview of the proposed framework. We probe intermediate representations of a neural network, extract interpretable features using a sparse autoencoder, and analyze their roles via gradient-based attribution and causal interventions.

Task Family	Train	Test	Difficulty Range
Relational Reasoning	2 000	200	3–10
Orientation Reasoning	2 000	200	2–10
Program Execution	2 000	200	2–10
Total (per language)	**6 000**	**600**	–
Total (3 languages)	**18 000**	**1 800**	–

Table 1: Dataset statistics across task families. Each task family is constructed in English, Chinese, and Arabic with computational equivalence.

by emitting a single token. This design isolates implicit spatial representations encoded in activations, rather than reasoning strategies expressed in generated text.

Data and Evaluation. For each task family and language, we construct 2,000 training instances and 200 held-out test instances, uniformly distributed across difficulty levels. Test data are never used during probing or SAE training. Unless stated otherwise, results are reported on in-distribution test sets.

Metrics. We report behavioral accuracy and analyze performance as a function of task complexity. For representational analysis, we use linear probe R^2 scores, regression error (MAE/RMSE for continuous variables), and layer-wise trends to characterize where spatial information emerges.

Multilingual Design. All tasks are independently constructed in English, Chinese, and Arabic. Prompts are not translated; instead, computational equivalence is preserved while allowing natural linguistic variation. Models are evaluated across all languages without language-specific tuning.

Table 1 summarizes dataset statistics.

Performance does not degrade monotonically with task complexity. In relational reasoning and perspective transformation, accuracy exhibits non-monotonic fluctuations across step lengths, suggesting regime shifts in internal processing rather than simple capacity limits. In contrast, spatial program execution shows more stable degradation, indicating that coordinate-based updating is comparatively more learnable from language statistics.

Cross-linguistic evaluation reveals moderate language dependence. English and Chinese show comparable performance overall, while Arabic lags most noticeably on relational reasoning tasks. Notably, spatial program execution exhibits the smallest cross-linguistic gap, suggesting greater language invariance for coordinate-based representations. As shown in Figure 3

4.2 Probing Spatial Representations

We train linear probes to decode task-relevant spatial variables from model hidden states, including relative position vectors (Task 1), orientation vectors (Task 2), and absolute spatial coordinates (Task 3). Probes are trained on the training split and evaluated on held-out test instances.

We train linear probes to decode task-relevant spatial variables from hidden states, including relative position vectors (Task 1), orientation vectors (Task 2), and absolute coordinates (Task 3). Probes are trained on training instances and evaluated on held-out test data. Across models and tasks, spatial information consistently peaks in intermediate layers and declines sharply toward the final layers (Figure 4). For Qwen2.5-7B-Instruct, relational reasoning reaches a maximum R^2 of 0.37 at layer 19, while spatial program execution peaks at $R^2 \approx 0.25$ around layer 16. Orientation variables are weakly decodable throughout ($R^2 < 0.15$). As shown in Table 3.

This inverted-U pattern indicates that spatial representations are constructed during intermediate processing but are not preserved into the final layers responsible for token prediction. Instruction-tuned models consistently show stronger spatial representations than base models, while cross-linguistic comparisons reveal similar layer-wise emergence patterns with varying representational strength.

Model	Lang	Relational	Orientation	Program Execution
Qwen2.5-7B-Base	EN	50.0 (100/200)	23.0 (46/200)	55.5 (111/200)
	ZH	39.5 (79/200)	23.5 (47/200)	58.0 (116/200)
	AR	39.5 (79/200)	24.5 (49/200)	47.0 (94/200)
Qwen2.5-7B-Instruct	EN	49.0 (98/200)	28.0 (56/200)	62.0 (124.0/200)
	ZH	47.5 (95/200)	28.0 (56/200)	52.5 (105/200)
	AR	31.0 (62/200)	26.0 (52/200)	57.5 (115/200)
Llama3-8B-Instruct	EN	4.0 (8/200)	2.0 (4/200)	2.0 (4/200)
	ZH	0.0 (0/200)	2.5 (5/200)	0.0 (0/200)
	AR	3.5 (7/200)	3.0 (6/200)	0.5 (1/200)

Table 2: Accuracy across task families and languages (%). Values are reported as accuracy with raw counts in parentheses. For Program Execution, accuracy may exceed 100% due to partial-credit scoring. Program Execution shows the strongest and most consistent performance across languages, while Orientation Reasoning performs near chance level. Arabic exhibits lower performance on Relational Reasoning, possibly reflecting language-specific encoding challenges.

Model	Lang	Relational Reasoning	Orientation Reasoning	Program Execution
Qwen2.5-7B-Base	EN	R^2=.382 / L=25 / MAE=.55 / RMSE=.70	R^2=–.007 / L=0	R^2=.394 / L=26 / MAE=1.83 / RMSE=2.79
	ZH	R^2=.254 / L=22 / MAE=.59 / RMSE=.77	R^2=–.003 / L=0	R^2=.347 / L=24 / MAE=1.93 / RMSE=2.90
	AR	R^2=.272 / L=24 / MAE=.59 / RMSE=.76	R^2=–.011 / L=0	R^2=.347 / L=24 / MAE=1.93 / RMSE=2.90
Qwen2.5-7B-Instruct	EN	R^2=.366 / L=19 / MAE=.55 / RMSE=.71	R^2=–.008 / L=9	R^2=.402 / L=20 / MAE=1.86 / RMSE=2.80
	ZH	R^2=.243 / L=19 / MAE=.60 / RMSE=.78	R^2=–.006 / L=0	R^2=.457 / L=20 / MAE=1.65 / RMSE=2.67
	AR	R^2=.264 / L=20 / MAE=.59 / RMSE=.76	R^2=–.005 / L=0	R^2=.418 / L=20 / MAE=1.82 / RMSE=2.74
Llama3-8B-Instruct	EN	R^2=.305 / L=31 / MAE=.57 / RMSE=.75	R^2=–.001 / L=17	R^2=.168 / L=31 / MAE=2.14 / RMSE=3.27
	ZH	R^2=.207 / L=31 / MAE=.60 / RMSE=.80	R^2=–.003 / L=0	R^2=.156 / L=31 / MAE=2.19 / RMSE=3.27
	AR	R^2=.172 / L=31 / MAE=.63 / RMSE=.82	R^2=–.002 / L=4	R^2=.156 / L=31 / MAE=2.19 / RMSE=3.27

Table 3: Best-layer probing results by model, language, and task family. We report coefficient of determination (R^2), best-performing layer (L), mean absolute error (MAE), and root mean squared error (RMSE).

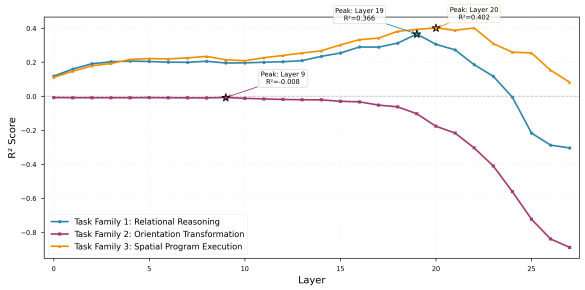

Figure 4: Layer-wise R² scores for spatial variable prediction across three task families (Qwen2.5-7B-Instruct, English). All tasks show mid-layer peaks followed by sharp declines in final layers. Task Family 1 and 3 demonstrate strong representational clarity (R² up to 0.37 and 0.40 respectively), while Task Family 2 shows minimal spatial encoding.

4.3 Sparse Autoencoder Analysis

To identify interpretable spatial features, we train sparse autoencoders (SAEs) on layers with peak probe performance for each task family. SAEs reveal a small subset of spatially selective features, typically accounting for 3–5% of all discovered features. As shown in Figure 5.

Across tasks, we observe clear axis- or direction-selective features, including units specialized for cardinal orientations (Task 2), coordinate ranges (Task 3), and relational axes (Task 1). Gradient-based attribution shows that features with stronger spatial selectivity contribute disproportionately to spatial predictions.

Feature overlap across task families is limited (12–18%), suggesting that different forms of spatial computation rely on largely distinct feature sets. Cross-linguistic analysis reveals partial feature sharing based on activation frequency, but substantially lower overlap when features are ranked by causal attribution, indicating language-specific mechanistic pathways supporting functionally similar behavior.

4.4 Causal Interventions

We perform activation patching and SAE feature ablation to assess whether spatial representations

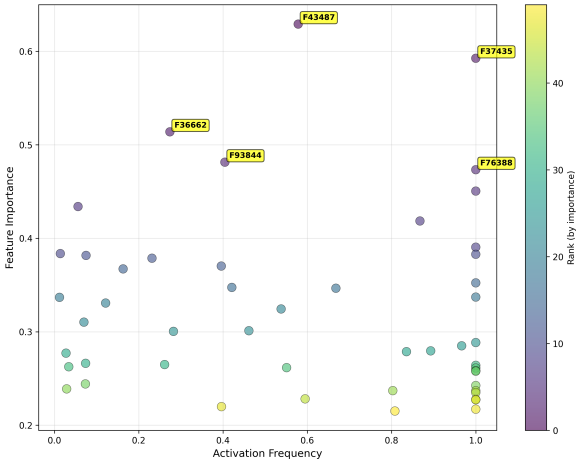

Figure 5: Feature importance versus activation frequency for Task Family 2 (Qwen2.5-7B-Istruct, Chinese). Important features are sparse and not aligned with activation frequency, indicating dissociation between usage and causal contribution.

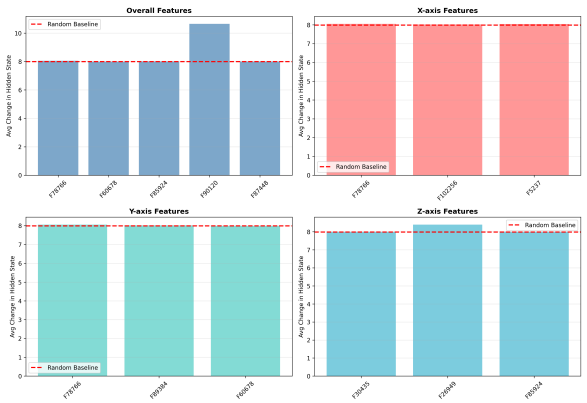

Figure 6: Intervention effects of spatial features across dimensions (Qwen2.5-7B-Instruct, Task Family 3, English). Bars show the average hidden-state change caused by intervening on spatially selective features, with the dashed line indicating a random-feature baseline.

causally influence model behavior.

In spatial program execution, patching intermediate-layer activations with counterfactual spatial states systematically shifts model outputs toward counterfactual trajectories, with strongest effects in layers 14–18. Interventions at final layers show minimal impact, despite being closest to output generation. As shown in Figure 6.

Complementary SAE feature ablation confirms functional specificity. Removing top spatial features reduces accuracy by 29% in orientation tasks and 14.5% in spatial program execution, while ablating non-spatial control features has negligible effect. These results establish that identified spatial representations are not merely decodable, but causally contribute to behavior.

5 Analyses and Discussion

5.1 What Kind of Spatial Representations Do LLMs Develop?

A central question of this work is whether large language models develop structured spatial representations analogous to cognitive maps, or whether spatial behavior arises from shallow linguistic heuristics. Our results suggest an intermediate regime: LLMs do encode spatial information, but these representations are conditional, fragmented, and weakly integrated into decision-making.

Linear probing reveals that task-relevant spatial variables are decodable from intermediate layers, with peak R^2 values up to 0.37–0.40 for relational reasoning and spatial program execution. These representations exhibit geometric coherence, supporting multi-axis integration rather than isolated symbolic associations. Crucially, causal interventions confirm functional relevance: activation patching at mid-layers reliably shifts model outputs toward counterfactual spatial trajectories.

At the same time, spatial representations are fragile and transient. Probe performance drops sharply in final layers, indicating that spatial structure constructed during intermediate processing is not preserved through output generation. Moreover, spatial features are largely task-specific, with limited overlap across relational, orientation, and coordinate-based reasoning. This fragmentation contrasts with biological cognitive maps, which support multiple spatial computations through a shared representational substrate.

Taken together, these findings indicate that LLMs exhibit *conditional and partial spatial encoding*: spatial structure is constructed when task format and linguistic cues make it accessible, but is neither persistent nor unified across tasks.

5.2 Failure Modes Across Spatial Tasks

Distinct failure patterns across task families further illuminate the limits of LLM spatial reasoning.

Relational reasoning: compositional collapse. In relational tasks, accuracy degrades non-monotonically with chain length, with sharp drops at intermediate complexity. Errors frequently preserve one spatial axis while losing another, suggesting partial breakdowns in multi-constraint composition. Performance recovery at higher complexity

likely reflects a shift toward coarse heuristics rather than robust relational inference.

Orientation transformation: representational absence. Orientation tracking performs near chance and exhibits near-zero probe R^2 across all layers. Error patterns reveal strong distributional bias toward frequent directions (e.g., north, east), indicating reliance on linguistic priors rather than geometric state tracking. Unlike position or coordinates, heading direction does not appear to be represented as a manipulable internal variable.

Spatial programs: arithmetic dominance. Spatial program execution shows the strongest behavioral performance, but error analysis reveals that a substantial fraction of failures are purely arithmetic. Once arithmetic errors are factored out, spatial performance aligns more closely with relational reasoning. This suggests that success in coordinate-based tasks partly reflects numerical competence rather than dedicated spatial computation.

5.3 Language Dependence of Spatial Representations

Our multilingual evaluation reveals a layered pattern of language dependence. Behaviorally, performance varies by up to 14 points across languages, most prominently in relational reasoning. Coordinate-based tasks show the smallest cross-linguistic gap, consistent with numerical formats being more language-invariant.

At the representational level, spatial information emerges at similar layers across languages, indicating a shared computational architecture. However, SAE analysis shows limited overlap in causally important features across languages. While different languages achieve comparable probe performance, they do so using distinct internal features, reflecting *mechanistic degeneracy*: similar outputs supported by different internal pathways.

These findings challenge the assumption that spatial reasoning in LLMs is inherently language-agnostic. Linguistic encoding substantially shapes both the accessibility and implementation of spatial computation.

5.4 Implications for Spatial Reasoning in LLMs

Spatial representations are necessary but insufficient. The gap between representational strength and behavioral accuracy indicates that encoding spatial variables alone does not guarantee reliable reasoning. Bottlenecks arise in propagating and integrating spatial information into final decision layers.

Implicit vs. explicit reasoning. Evaluating models without chain-of-thought reveals latent spatial structure, but this structure is fragile. We hypothesize that explicit reasoning stabilizes spatial representations by externalizing intermediate states, acting as a compensatory scaffold rather than merely revealing hidden competence.

Task format matters. Performance differences across task families highlight the role of format. Coordinate-based tasks benefit from arithmetic pathways, while relational and orientation tasks demand abstract composition that LLMs struggle to sustain. High performance on one format does not imply general spatial understanding.

6 Conclusion

Large language models achieve strong performance on spatial benchmarks, yet it remains unclear whether this reflects genuine spatial reasoning or reliance on linguistic heuristics. Motivated by cognitive neuroscience, we adopt a mechanistic perspective and introduce a compact task taxonomy that isolates three core spatial primitives to analyze multilingual models across languages. We find that task relevant spatial information is encoded in intermediate layers and can causally influence behavior, but these representations are fragile, transient, and highly task specific. Spatial structure typically emerges in mid layers but is weakly preserved in final decisions, and different spatial tasks rely on largely disjoint internal features, providing limited evidence for unified spatial representations. Cross linguistic analysis reveals mechanistic degeneracy, where the emergence of spatial information is largely language invariant but the internal pathways supporting spatial computation vary across languages, with greater invariance observed in coordinate based tasks. Overall, these findings place current LLMs between shallow pattern matching and robust spatial cognition as characterized in neuroscience, suggesting that more human like spatial reasoning will require approaches that better preserve and integrate spatial representations across processing stages and motivating mechanistic evaluation beyond benchmark accuracy.

Limitations

This study analyzes spatial representations in two multilingual language model families at the 7–8B scale, and findings may differ for larger models or alternative architectures. The proposed tasks isolate three spatial primitives and do not cover more complex spatial abilities such as planning or embodied interaction. Experiments are limited to English, Chinese, and Arabic, and spatial representations may interact differently with other languages. Causal conclusions rely on activation patching and feature ablation, which provide partial causal characterization, and representational analyses focus on linearly decodable information, potentially missing nonlinear or distributed structure. Finally, tasks are abstract and text-only, limiting conclusions about real-world or multimodal spatial reasoning.

References

Joseph Bloom, Curt Tigges, Anthony Duong, and David Chanin. 2024. Saelens. https://github.com/decoderesearch/SAELens.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Krzysztof Choromanski, Tianli Ding, Danny Driess, Kumar Avinava Dubey, Chelsea Finn, Peter R. Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, and 27 others. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. *ArXiv*, abs/2307.15818.

Neil Burgess. 2008. Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, 1124(1):77–97.

Danny Driess, F. Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Ho Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, and 3 others. 2023. Palm-e: An embodied multimodal language model. In *International Conference on Machine Learning*.

Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. 2024. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Annual Meeting of the Association for Computational Linguistics*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony S. Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, and 510 others. 2024. The llama 3 herd of models.

Michael J. Eckardt. 1980. The hippocampus as a cognitive map. *Journal of Nervous and Mental Disease*, 168:191–192.

Ruth B. Ekstrom, John W. French, and Harry H. Harman. 1976. Manual for kit of factor-referenced cognitive tests.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2020. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, and 1 others. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

A. S. Etienne and Kate J. Jeffery. 2004. Path integration in mammals. *Hippocampus*, 14.

Chaz Firestone. 2020. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117:26562 – 26571.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665 – 673.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, N. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. In *International Joint Conference on Artificial Intelligence*.

Sharut Gupta, Joshua Robinson, Derek Lim, Soledad Villar, and Stefanie Jegelka. 2023. Structuring representation geometry with rotationally equivariant contrastive learning. *ArXiv*, abs/2306.13924.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. *ArXiv*, abs/1909.03368.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *North American Chapter of the Association for Computational Linguistics*.

Trenton W Holliday, Joanna R. Gautney, and Lukas Friedl. 2014. Right for the wrong reasons. *Current Anthropology*, 55:696 – 724.

Renate Krause and Stefan Reimann. 2024. Items or relations - what do artificial neural networks learn? *ArXiv*, abs/2404.12401.

Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024a. Reframing spatial reasoning evaluation in language models: A real-world simulation benchmark for qualitative reasoning. *ArXiv*, abs/2405.15064.

Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. 2024b. Llms for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 119–126.

Yuan-Hong Liao, Rafid Mahmood, Sanja Fidler, and David Acuna. 2024. Reasoning paths with reference objects elicit quantitative spatial reasoning in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*.

Weichen Liu, Qiyao Xue, Haoming Wang, Xiangyu Yin, Boyuan Yang, and Wei Gao. 2025. Spatial reasoning in multimodal large language models: A survey of tasks, benchmarks and methods.

Hartmut Maennel, Ibrahim M. Alabdulmohsin, Ilya O. Tolstikhin, Robert J. N. Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. 2020. What do neural networks learn when trained with random labels? *ArXiv*, abs/2006.10455.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Neural Information Processing Systems*.

Neel Nanda and Joseph Bloom. 2022. Transformerlens. `https://github.com/TransformerLensOrg/TransformerLens`.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova Dassarma, T. J. Henighan, Benjamin Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, John Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *ArXiv*, abs/2209.11895.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Narain Sohl-Dickstein. 2017. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Neural Information Processing Systems*.

Md Imbesat Hassan Rizvi, Xiaodan Zhu, and Iryna Gurevych. 2024. Sparc and sparp: Spatial reasoning characterization and path generation for understanding spatial reasoning capability of large language models. In *Annual Meeting of the Association for Computational Linguistics*.

Manasi Sharma. 2023. Exploring and improving the spatial reasoning abilities of large language models. *ArXiv*, abs/2312.01054.

Roger N. Shepard and Jacqueline Metzler. 1971. Mental rotation of three-dimensional objects. *Science*, 171:701 – 703.

Zhengxiang Shi, Qiang Zhang, and Aldo Lipani. 2022. Stepgame: A new benchmark for robust multi-hop spatial reasoning in texts. In *AAAI Conference on Artificial Intelligence*.

Mirac Suzgun, Nathan Scales, Nathanael Scharli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Annual Meeting of the Association for Computational Linguistics*.

Edward Chace Tolman. 1948. Cognitive maps in rats and men. *Psychological review*, 55 4:189–208.

Jiaqi Wang, Zihao Wu, Yiwei Li, Hanqi Jiang, Peng Shu, Enze Shi, Huawen Hu, Chong-Yi Ma, Yi-Hsueh Liu, Xuhui Wang, Yincheng Yao, Xuan Liu, Huaqin Zhao, Zheng Liu, Haixing Dai, Lin Zhao, Bao Ge, Xiang Li, Tianming Liu, and Shu Zhang. 2024a. Large language models for robotics: Opportunities, challenges, and perspectives. *ArXiv*, abs/2401.04334.

Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, and Neel Joshi. 2024b. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *ArXiv*, abs/2406.14852.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *ArXiv*, abs/2109.01652.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Wenshan Wu, Shaoguang Mao, Yadong Zhang, Yan Xia, Li Dong, Lei Cui, and Furu Wei. 2024. Mind's eye of llms: Visualization-of-thought elicits spatial reasoning in large language models. *Advances in Neural Information Processing Systems 37*.

Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. 2024. On memorization of large language models in logical reasoning. *ArXiv*, abs/2410.23123.

Yutaro Yamada, Yihan Bao, Andrew Kyle Lampinen, Jungo Kasai, and Ilker Yildirim. 2023. Evaluating spatial understanding of large language models. *ArXiv*, abs/2310.14540.

Daniel Yamins and James J. DiCarlo. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19:356–365.

Hanqi Yan, Yanzheng Xiang, Guangyi Chen, Yifei Wang, Lin Gui, and Yulan He. 2024. Encourage or inhibit monosemanticity? revisit monosemanticity from a feature decorrelation perspective. In *Conference on Empirical Methods in Natural Language Processing*.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Fengji Yi, Wenlong Fu, and Huan Liang. 2018. Model-based reinforcement learning: A survey.

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

A Supplementary Materials

This appendix provides additional details on task generation, experimental configurations, and extended results that complement the main text.

A.1 Task Generation Details

We provide detailed algorithms for generating each task family. All algorithms use controlled randomization with fixed seeds to ensure reproducibility and computational equivalence across languages.

A.1.1 Algorithm 1: Relational Spatial Reasoning

Algorithm 1 generates multi-hop spatial reasoning instances. The key design principle is to ensure that answering the query question requires composing multiple pairwise relations, with no direct relation between the source and target entities.

Spatial Encoding: We use a 3D coordinate system where each atomic relation corresponds to a unit vector: left/right $= (\pm 1, 0, 0)$, behind/front $= (0, \pm 1, 0)$, below/above $= (0, 0, \pm 1)$. The vector representation enables systematic computation of transitive relations.

Reasoning Requirement: To guarantee multi-hop reasoning, the target entity (last in sequence) cannot directly reference the source entity (first in sequence), ensuring at least two inference steps.

A.1.2 Algorithm 2: Orientation Reasoning

Algorithm 2 generates orientation reasoning instances that require tracking heading direction through a sequence of turn actions.

Direction Encoding: Cardinal directions are encoded as angles in the standard mathematical convention: east = 0°, north = 90°, west = 180°, south = 270°. For probing analysis, directions are further encoded as unit vectors $(\cos\theta, \sin\theta)$ to enable continuous regression.

Turn Actions: We define three turn actions with deterministic effects: "Turn right" (clockwise 90°), "Turn left" (counterclockwise 90°), and "Turn around" (180°). These correspond to rotation matrices applied to the current heading vector.

Stateful Reasoning: Unlike Task Family 1, which involves relational composition over a static configuration, Task Family 2 requires maintaining and updating a latent state (current orientation) across sequential operations.

Algorithm 1 Generation of Multi-hop Spatial Reasoning Instances

Require: Number of samples N, hop range $[S_{\min}, S_{\max}]$
Ensure: Dataset $\mathcal{D}$

Define atomic spatial relations $\mathcal{R}$ and their vector encodings
Initialize empty dataset $\mathcal{D}$
for $i = 1$ **to** N **do**
  Sample hop count $S \sim \textsc{Uniform}(S_{\min}, S_{\max})$
  Create entity sequence $(e_0, e_1, \ldots, e_S)$
  Set position $\mathbf{p}(e_0) \leftarrow \mathbf{0}$
  Initialize fact set $\mathcal{K} \leftarrow \emptyset$
  for $j = 1$ **to** S **do**
    Sample reference entity $r_j \in \{e_0, \ldots, e_{j-1}\}$
    Sample relation $r \in \mathcal{R}$
    $\mathbf{p}(e_j) \leftarrow \mathbf{p}(r_j) + \mathbf{v}(r)$
    Add fact (e_j is r of r_j) to $\mathcal{K}$
  end for
  Let source $s \leftarrow e_0$, target $t \leftarrow e_S$
  Compute relative vector $\Delta \leftarrow \mathbf{p}(t) - \mathbf{p}(s)$
  Derive gold relation $a \leftarrow \textsc{Rel}(\Delta)$
  Construct question q from $\mathcal{K}$ and (s, t)
  Sample distractor options D from $\mathcal{R} \setminus \{a\}$
  Shuffle options $O \leftarrow \{a\} \cup D$
  Add instance (q, a, O, Δ, S) to $\mathcal{D}$
end for
return $\mathcal{D}$

Algorithm 2 Generation of Orientation Reasoning Instances

Require: Number of samples N, step range $[S_{\min}, S_{\max}]$
Ensure: Dataset $\mathcal{D}$

Define cardinal directions $\mathcal{D} = \{\texttt{north}, \texttt{east}, \texttt{south}, \texttt{west}\}$
Define turn actions $\mathcal{A}$ with rotation offsets
Initialize empty dataset $\mathcal{D}$
for $i = 1$ **to** N **do**
  Sample number of steps $S \sim \textsc{Uniform}(S_{\min}, S_{\max})$
  Sample initial orientation $o_0 \sim \mathcal{D}$
  Set current orientation $o \leftarrow o_0$
  Initialize action list $\mathcal{K} \leftarrow \emptyset$
  for $j = 1$ **to** S **do**
    Sample turn action $a_j \sim \mathcal{A}$
    Update orientation $o \leftarrow \textsc{Rotate}(o, a_j)$
    Append a_j to $\mathcal{K}$
  end for
  Let final orientation $o_S \leftarrow o$
  Construct question q from $(o_0, \mathcal{K})$
  Set answer $a \leftarrow o_S$
  Encode target vector $\mathbf{t} \leftarrow (\cos\theta, \sin\theta)$ for o_S
  Sample distractor options O from $\mathcal{D} \setminus \{a\}$
  Shuffle answer and distractors into multiple-choice options
  Add instance $(q, a, O, \mathbf{t}, S)$ to dataset
end for
return $\mathcal{D}$

A.1.3 Algorithm 3: Spatial Program Execution

Algorithm 3 generates spatial program execution instances that require computing the final position after applying a sequence of geometric transformations.

Action Space: We include five types of operations: (1) *Move* – translate along cardinal directions; (2) *Reflect* – mirror across coordinate axes; (3) *Rotate* – rotate around axes by 90°/180°/270°; (4) *Scale* – multiply all coordinates by a factor; (5) *Translate* – add offset vector.

Coordinate System: All operations are defined in a 3D Cartesian coordinate system starting at origin $(0, 0, 0)$. The coordinate system uses right-handed convention with x (left/right), y (backward/forward), z (down/up).

Cumulative State: Each operation modifies the current position, and the final answer depends on the cumulative effect of all operations. Errors in intermediate steps propagate to the final result, making this task diagnostic of stateful computation and error accumulation.

Quality Control: To prevent extreme coordinate values that could arise from repeated scaling, we implement rejection sampling and constrain maximum coordinate magnitudes to ± 50 units.

A.2 Full Probing Results

A.2.1 Layer-wise Probing Patterns

Across all experiments, we observe consistent patterns in how spatial information is distributed across model layers:

- **Inverted-U profile:** Spatial information peaks in intermediate layers and declines sharply in final layers, indicating that spatial representations are constructed during intermediate processing but not preserved through

Algorithm 3 Generation of Spatial Procedure Execution Instances

Require: Number of samples N, step range $[S_{\min}, S_{\max}]$
Ensure: Dataset $\mathcal{D}$

Define action space $\mathcal{A}$ (move, reflect, rotate, scale, translate)
Initialize empty dataset $\mathcal{D}$
for $i = 1$ **to** N **do**
  Sample number of steps $S \sim \textsc{Uniform}(S_{\min}, S_{\max})$
  Initialize position $\mathbf{p}_0 \leftarrow (0, 0, 0)$
  Set current position $\mathbf{p} \leftarrow \mathbf{p}_0$
  Initialize action sequence $\mathcal{K} \leftarrow \emptyset$
  for $j = 1$ **to** S **do**
    Sample spatial action $a_j \sim \mathcal{A}$
    Update position $\mathbf{p} \leftarrow \textsc{Apply}(\mathbf{p}, a_j)$
    Append a_j to $\mathcal{K}$
  end for
  Let final position $\mathbf{p}_S \leftarrow \mathbf{p}$
  Construct question q from $(\mathbf{p}_0, \mathcal{K})$
  Set answer $a \leftarrow \mathbf{p}_S$
  Sample distractor positions O by perturbing $\mathbf{p}_S$
  Shuffle correct answer and distractors into multiple-choice options
  Add instance $(q, a, O, \mathbf{p}_S, S)$ to dataset
end for
return $\mathcal{D}$

output generation.

- **Task-specific variations:** Different task families show peak spatial encoding at different layer positions, suggesting distinct computational stages for different types of spatial reasoning.
- **Cross-linguistic consistency:** The layer-wise emergence pattern is similar across languages, though the magnitude of decodable information varies.

Detailed layer-by-layer probing results are presented in the main paper (Table 3).

A.3 SAE Hyperparameters and Visualizations

A.3.1 SAE Training Configuration

All sparse autoencoders were trained using the SAE-Lens library with the following configuration:

- **Architecture:** Standard one-layer SAE with ReLU activation
- **Expansion factor:** 32× (e.g., 3584 → 114,688 features for Qwen2.5-7B)
- **Training samples:** 2,000 task instances per language
- **Training tokens:** Approximately 300,000–420,000 tokens depending on language
- **Batch size:** 4,096 tokens
- **Learning rate:** 3×10^{-4} with linear warmup (1,000 steps) and cosine decay
- **L1 coefficient:** $\lambda = 0.001$ (selected via validation sweep over $\{0.0001, 0.0005, 0.001, 0.005\}$)
- **Optimizer:** AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- **Training steps:** 300–400 batches until convergence
- **Target layer:** Peak probe layer for each task family

A.3.2 SAE Quality Metrics

Task	Lang	MSE	R^2	Sparsity	L0
Task 1	EN	**0.136**	**0.9996**	0.0025	288.3
	ZH	0.798	0.9965	0.0023	268.2
	AR	0.252	0.9989	**0.0019**	**219.5**
Task 2	EN	0.147	0.9994	0.0026	301.4
	ZH	0.823	0.9962	0.0024	276.8
	AR	0.268	0.9987	**0.0020**	**227.3**
Task 3	EN	**0.142**	**0.9995**	0.0027	294.7
	ZH	0.765	0.9968	0.0025	271.5
	AR	0.241	0.9990	**0.0021**	**223.8**

Table 4: SAE reconstruction quality and sparsity metrics for Qwen2.5-7B-Instruct. MSE: mean squared error; R^2: explained variance; Sparsity: fraction of active features; L0: average number of active features per sample. Best values per task family in bold.

Reconstruction MSE ranges from 0.136 (EN, Task 1) to 0.798 (CN, Task 1), with explained variance consistently above 0.996 except for Chinese Task 1 (0.9965). Average L0 (number of active features) ranges from 219 (AR, Task 1) to 301 (EN, Task 2), indicating successful sparsity enforcement.

A.3.3 Feature Activation Distributions

Most features activate rarely, with a heavy-tailed distribution where a small proportion of features account for the majority of activation mass. This

confirms that SAEs successfully learn sparse representations.

A.3.4 Feature Selectivity Analysis

We compute feature selectivity using gradient-based attribution (gradient × activation). Analysis of SAE features reveals that spatially selective features are not necessarily the most frequently activated, indicating dissociation between usage frequency and causal importance for spatial reasoning.

A.4 Prompt Templates (English)

All tasks follow a multiple-choice format with four options. Below are representative prompt templates for each task family in English.

A.4.1 Task Family 1: Relational Spatial Reasoning

[English]
`System:` You are a helpful assistant.

`User:` Given the following spatial facts, answer the question by selecting one option.

D is behind C.
B is left of A.
C is behind B.
E is right of D.

Where is E relative to A?

A. right and behind
B. left and front
C. behind
D. left and behind

`Assistant:` The answer is

A.4.2 Task Family 2: Orientation Reasoning

[English]
`System:` You are a helpful assistant.

`User:` Follow the instructions step by step and answer the question by selecting one option.

You are facing north.
Turn right.
Turn left.
Turn around.
Turn right.

Which direction are you facing now?

A. north
B. east
C. south
D. west

`Assistant:` The answer is

A.4.3 Task Family 3: Spatial Program Execution

[English]
`System:` You are a helpful assistant.

`User:` Execute the spatial operations step by step and answer the question by selecting one option.

Start at (0, 0, 0).
Move forward by 3 units.
Move right by 2 units.
Reflect the position across the z-axis.
Move up by 1 unit.

What is the final position?

A. (2, 3, -1)
B. (2, 3, 1)
C. (-2, 3, 1)
D. (2, -3, 1)

`Assistant:` The answer is

A.4.4 Inference Protocol

For all experiments:

- Models generate a single token continuation after the prompt
- The most likely single-character token (A/B/C/D) is selected as the answer
- No chain-of-thought or explicit reasoning is elicited

- Temperature is set to 0 (greedy decoding)
- Maximum generation length is 1 token

This protocol ensures that evaluation targets implicit spatial representations encoded in activations rather than verbalized reasoning strategies.

A.5 Dataset Statistics and Quality Control

A.5.1 Data Generation Process

All datasets were generated using rule-based procedures with controlled randomization. For each task family:

1. **Entity/Action Sampling:** Entities (Task 1), directions (Task 2), and operations (Task 3) are sampled uniformly at random.
2. **Constraint Verification:** Generated instances are verified to ensure they require multi-step reasoning and do not allow shortcut solutions.
3. **Answer Distribution:** Correct answers are balanced across option positions (A/B/C/D) with uniform distribution (25% each).
4. **Difficulty Stratification:** Instances are stratified by complexity level (number of steps) with uniform distribution across difficulty range.

A.5.2 Cross-Linguistic Consistency

To ensure computational equivalence across languages, we verify:

- **Structural isomorphism:** Same reasoning steps and answer patterns
- **Token count variance:** Chinese prompts are typically 15–20% shorter due to character-level encoding; Arabic prompts are 10–15% longer
- **Template diversity:** All languages use the same number of prompt templates
- **Random seed alignment:** Parallel instances across languages use the same random seed, ensuring matched difficulty levels

A.6 Additional Experimental Details

A.6.1 Computational Resources

All experiments were conducted on NVIDIA A100 GPUs (40GB). Typical resource requirements:

- Model inference (2,000 samples): 10–15 minutes
- Linear probe training (all layers): 30–45 minutes
- SAE training (one layer, one task): 2–3 hours
- Intervention experiments: 1–2 hours per configuration

Total compute: approximately 500 GPU-hours across all models, languages, and tasks.

A.6.2 Reproducibility

We provide:

- Complete data generation code with fixed random seeds
- Exact model checkpoints and inference configurations
- Probe training scripts with hyperparameter specifications
- SAE training configurations and learned feature dictionaries
- Intervention protocols and analysis notebooks

All code will be released upon publication to ensure full reproducibility.

A.7 Intervention Experiment Details

A.7.1 Activation Patching Protocol

For each intervention experiment:

1. **Baseline forward pass:** Run model on original input, collect activations at target layer
2. **Counterfactual forward pass:** Run model on counterfactual input (e.g., different spatial trajectory), collect activations at same layer
3. **Patching:** Replace activations at target layer with counterfactual activations
4. **Measurement:** Continue forward pass and measure output shift toward counterfactual prediction

We report:

- **Shift magnitude:** KL divergence between original and patched output distributions
- **Prediction change:** Whether patching changes the top-1 prediction
- **Layer specificity:** Effect size as a function of intervention layer

A.7.2 SAE Feature Ablation Protocol

For each feature ablation experiment:

1. **Feature identification:** Rank SAE features by gradient-based attribution
2. **Selective ablation:** Zero out top-k features for varying k values
3. **Reconstruction:** Decode modified SAE activations back to hidden states
4. **Evaluation:** Measure accuracy drop on test set

Feature ablation experiments confirm that identified spatial features causally contribute to model behavior, with accuracy systematically decreasing as more top-ranked features are removed.

A.7.3 Counterfactual Construction

For spatial program execution (Task 3), we construct counterfactuals by:

- Flipping a single operation in the sequence (e.g., "move left" $\rightarrow$ "move right")
- Computing the resulting position divergence $\|\mathbf{p}_{\text{actual}} - \mathbf{p}_{\text{counterfactual}}\|$
- Selecting pairs with divergence in range [3, 10] units (neither too similar nor too different)

For relational reasoning (Task 1), counterfactuals are constructed by:

- Flipping one spatial relation (e.g., "A left of B" $\rightarrow$ "A right of B")
- Verifying that the modified configuration is logically consistent
- Computing answer change (same vs. different final relation)