EUCLIFOLD: PROBING 3D EUCLIDEAN PRIOR IN VLMS VIA COGNITIVELY-STRATIFIED FOLDING TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans leverage robust 3D spatial priors to align perception with the physical world, enabling flexible and intelligent behavior. While Vision-Language Models (VLMs) exhibit impressive zero-shot performance, it remains unclear whether they possess genuine spatial reasoning capabilities, as standard evaluations are confounded by dataset bias and spurious correlations. To address this, we introduce **EucliFold**, a synthetic visual question-answering benchmark focused on cube net folding in Euclidean space—a domain that enables precise analysis while requiring genuine spatial understanding. We propose a **cognitively-stratified evaluation framework** that decomposes spatial reasoning into three hierarchical levels: **Perception** (grounding sensory input to spatial representations), **Operation** (manipulating representations according to instructions), and **Imagination** (autonomous spatial problem-solving under geometric constraints). This decomposition isolates genuine spatial reasoning from superficial pattern matching. To mitigate evaluation biases, we employ **Winograd-style accuracy** using minimal-pair contrastive samples. Our evaluation reveals that state-of-the-art VLMs demonstrate reasonable perceptual capabilities but fail significantly at operational and imagination-level spatial reasoning, suggesting reliance on statistical patterns rather than genuine geometric understanding. Ablation studies confirm the effectiveness of our cognitively-stratified decomposition and bias-resistant evaluation methodology. EucliFold provides a rigorous testbed for probing emergent spatial priors in future models and demonstrates how systematic cognitive decomposition can reveal nuanced capability gaps in VLMs.

1 INTRODUCTION

Vision-Language Models (VLMs) have demonstrated remarkable zero-shot generalization capabilities across diverse tasks (Liu et al., 2023b; Chen et al., 2024), suggesting the emergence of sophisticated internal representations and reasoning mechanisms. Recent research has provided compelling evidence that transformer-based models (Vaswani et al., 2017) can develop internal representations that align with real-world structure (Gurnee & Tegmark, 2023) and human perception (Abdou et al., 2021; Huh et al., 2024), learn generalizable solutions (Zhong et al., 2023; Huang et al., 2024), and exhibit emergent behaviors (Brown et al., 2020; Wei et al., 2022). However, the spatial reasoning capabilities of VLMs remain relatively weak, compared with human (Ma et al., 2024; Liu et al., 2023a; Tang et al., 2025; Tong et al., 2025), and current spatial ability evaluation datasets struggle to quantitatively assess whether VLMs possess generalizable spatial priors. This study explores the quantitative measurement of such priors.

Spatial ability evaluation datasets can be broadly categorized into two types: curated datasets and synthetic datasets. The former (Yu et al., 2023; Liu et al., 2024; Ma et al., 2024; Tang et al., 2025) provide valuable insights into real-world applicability but suffer from naturalistic confounds and spurious correlations that obscure the sources of model failures. Conversely, synthetically generated datasets enable controlled evaluation and have revealed important insights—such as VLMs' significantly weaker three-dimensional spatial reasoning compared to two-dimensional reasoning (Mayer et al., 2025; Zhang et al., 2025)—yet existing benchmarks lack a principled decomposition of spatial reasoning into distinct cognitive levels. This limitation conflates low-level perceptual abilities

with high-level cognitive reasoning, making it difficult to assess whether VLMs possess genuine, generalizable spatial priors or merely exploit superficial statistical regularities.

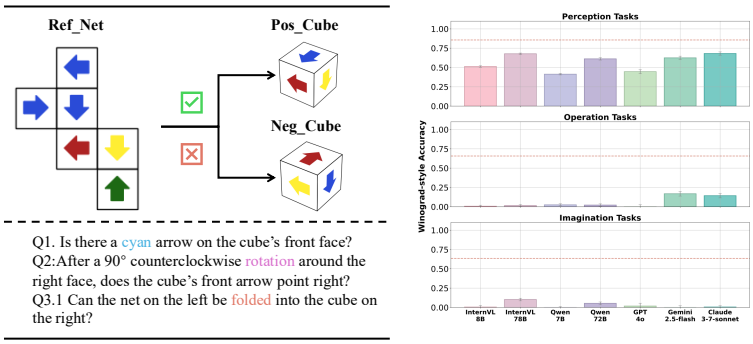

Figure 1: Euclifold Tasks and Performance

To address these challenges, we propose a evaluation framework grounded in cognitive science principles. We introduce a theoretically motivated decomposition of spatial prior into three qualitatively distinct levels: **Perception** (grounding sensory input to spatial representations), **Operation** (manipulating spatial representations according to external instructions), and **Imagination** (autonomous spatial problem-solving under geometric constraints). This hierarchical framework enables precise characterization of VLMs' spatial capabilities and identifies the specific cognitive levels at which generalization succeeds or fails.

For our empirical investigation, as shown in Figure 1, we focus on cube net folding tasks within Euclidean space and developed a synthetic dataset **EucliFold**. This choice is motivated by several key advantages: Euclidean representations provide mathematical precision for rigorous analysis; cube folding scales from basic perception to complex imagination-level reasoning; and systematic task variation enables controlled evaluation while maintaining sufficient complexity to reveal meaningful capability differences.

To address VLMs' intrinsic biases—including perception bias (Wang et al., 2024), pre-training bias (Lin et al., 2024), and response bias (Zheng et al., 2024)—we develop a bias-resistant evaluation methodology. We employ "Winograd-Style Accuracy" that compares performance on minimal-pair samples differing only in critical spatial content. This approach, inspired by the Winograd Schema Challenge (Levesque et al., 2012; Thrush et al., 2022), isolates genuine spatial reasoning from statistical artifacts by measuring the difference between true-belief and false-belief response patterns.

Our evaluation reveals that while current VLMs achieve reasonable performance on perceptual tasks, they struggle significantly with operational spatial reasoning and fail almost entirely at imagination-level tasks requiring autonomous spatial problem-solving. These findings suggest that current VLMs lack robust internal spatial representations and rely heavily on superficial pattern matching rather than genuine geometric understanding.

We make four contributions: (1) a systematic decomposition of spatial reasoning into distinct cognitive levels; (2) a controlled synthetic dataset EucliFold eliminating confounds while maintaining complexity; (3) bias-resistant evaluation distinguishing high-level Euclidean prior from low-level pattern matching; (4) comprehensive analysis revealing systematic VLM spatial reasoning gaps.

2 Spatial Priors of Three Cognitive Levels

Understanding spatial reasoning in artificial systems requires decomposing the underlying cognitive processes. Drawing from cognitive science research on spatial cognition and mental imagery (Shepard & Metzler, 1971), we propose that robust 3D spatial reasoning emerges from three hierarchical capabilities: **Perception**, **Operation**, and **Imagination**. Each level exhibits qualitatively distinct

characteristics and builds upon the previous one to enable increasingly sophisticated spatial reasoning.

Perception: Grounding Spatial Concepts. At the foundational level, Perception encompasses grounding multimodal inputs into coherent internal spatial representations. The robustness of these internal spatial representations defines the perception-level spatial priors of humans or artificial models. Tasks such as spatial relation extraction (Liu et al., 2023a) evaluate the quality of perception-level spatial priors. *Generalizability* at this level stems from cross-modal consistency and distribution-invariant spatial relations that depend on relative geometric relationships rather than specific sensory modalities.

Operation: Manipulating Spatial Representations. Building upon perceptual grounding, Operation involves systematic manipulation of spatial representations according to external instructions or rules. This requires knowledge of spatial transformation dynamics and geometric mappings. Tasks such as dynamic prediction (Yi et al., 2020) evaluate the quality of operation-level spatial priors. *Generalizability* emerges from understanding predictable spatial dynamics that transfer across novel contexts. Crucially, operational competence provides the foundation for self-prediction when applying transformations—a prerequisite for imagination-level reasoning.

Imagination: Autonomous Spatial Problem-Solving. At the highest level, Imagination represents autonomous generation and evaluation of spatial operations under spatial constraints. *Generalizability* stems from internalizing fundamental geometric principles, enabling flexible reasoning over open-ended operation sets and novel problem configurations.

This hierarchical framework has critical implications for evaluation: systems may exhibit superficially impressive higher-level performance while relying on brittle, correlation-based strategies at lower levels. Such systems fail when encountering distribution shifts or requiring genuine geometric reasoning rather than pattern matching.

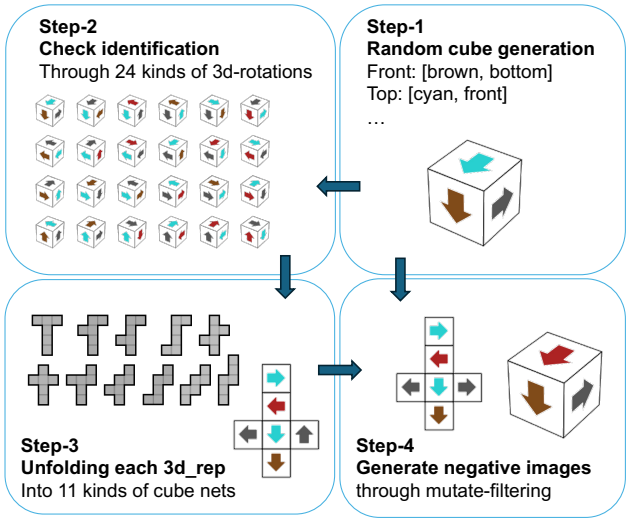

Figure 2: EucliFold Data Generation Process

3 EUCLIFOLD DATASET

3.1 SPATIAL REPRESENTATION DESIGN AND IMAGE GENERATION

The tasks in EucliFold center on cube net folding, chosen for three key properties: (1) it requires object manipulation and matching within three-dimensional Euclidean space, (2) it admits multiple valid spatial operation sequences without a fixed methodology, yet each solution pathway involves non-trivial reasoning, and (3) it maintains simplicity at both perceptual and cognitive levels without introducing excessive extraneous complexity.

Building upon the cube folding framework, we designed cube surface patterns that maintain diversity while avoiding excessive cognitive burden. Each cube face displays an arrow oriented parallel to the cube's edges (yielding four possible orientations), with each arrow rendered in one of eight possible colors (consistent with the established CLEVR benchmark Johnson et al. (2017)). Under this design, there exist at least $\frac{(8\times4)^6}{24} > 44,739,242$ distinct cubes and $(8 \times 4)^3 = 32,768$ different three-dimensional cube views.

To minimize spurious correlations at lower cognitive levels, EucliFold employs a systematic generation pipeline (Figure 2) that produces the required 3D cube views and 2D cube net images through uniform sampling based on symmetry principles at each stage. More details are in A

3.2 Contrastive Question Pairs

We employ fixed templates (see Appendix C) to test VLMs, varying only key vocabulary or image content. By minimizing differences between positive and negative samples, we control for VLMs' inherent biases, particularly response bias Zheng et al. (2024).

Perception Tasks All images in EucliFold are generated using the Python package *matplotlib*, which significantly differs from the image input distribution that most VLMs encounter during training. To ensure narrative precision, we use fixed text templates to describe problems, which also differs from everyday conversational language. This distributional shift poses challenges to VLMs' perceptual generalization capabilities.

Perception-level tasks examine whether VLMs can achieve robust spatial concept understanding under EucliFold's language-image distribution. We generate images $I(\text{3d\_rep})$, text $T(\text{3d\_rep})$, and incorrect text $T_{\text{neg}}(\text{3d\_rep})$ based on three-dimensional spatial representations, asking models to judge image-text matching.

Ideally, we test whether models can map both images and text to correct internal spatial concepts. Limited by interaction modalities, we can only estimate spatial grounding through image-text matching: $\textbf{Score}[I(\textbf{3d\_rep}), T(\textbf{3d\_rep})] \sim \textbf{Score}[I(\textbf{3d\_rep}), \textbf{3d\_rep}] \times \textbf{Score}[T(\textbf{3d\_rep}), \textbf{3d\_rep}]$

Through uniform sampling of *3d_rep*, we attempt to offset spurious matches based on irrelevant visual content and restore matches based on spatial concepts.

We design two perception tasks, *color recognition* and *orientation recognition*, to test whether models understand cube orientations in three-dimensional space and whether they understand the three-dimensional orientations corresponding to two-dimensional patterns on cube surfaces. The *color recognition* and *orientation task* uses the following statement template:

> *There is a* **{*color/wrong color*}** *arrow on the* {*visible face*} *face of the cube.*
> *The arrow on the* {*visible face*} *face of the cube is pointing towards the* {*orientation*}.

Since VLMs achieve extremely high accuracy in color recognition, we use color as an indicator to test models' perception and understanding of different cube surfaces. Building on cube surface perception, we can contrastively analyse the *orientation tasks* to judge whether models can correctly perceive the arrow directions on cube surfaces.

Operation Tasks Operation tasks examine whether VLMs can understand three-dimensional spatial rotations of cubes. While other spatial operations could be examined, we choose three-dimensional rotation to maintain consistency with the text-image distribution of perception problems.

We generate images $I(\text{3d\_rep})$ based on three-dimensional spatial representations, operation text $T_{\text{op}}(\text{3d\_op})$ based on three-dimensional rotation operations, and result text $T(\text{3d\_op}(\text{3d\_rep})) = T(\text{final\_3d\_rep})$ based on final representations. Negative samples are $T(\text{wrong\_final\_3d\_rep})$. We expect models to complete text-image matching in stages:

$$\begin{aligned}&\text{Score}[I(\text{3d\_rep}), T_{\text{op}}(\text{3d\_op}) + T(\text{3d\_op}(\text{3d\_rep}))] \\ &\sim \text{Score}[I(\text{3d\_rep}), \text{3d\_rep}] \times \text{Score}[T(\text{3d\_op}), \text{3d\_op}] \\ &\quad \times \text{Score}[\text{3d\_op}(\text{3d\_rep}), \text{final\_3d\_rep}] \times \text{Score}[T(\text{final\_3d\_rep}), \text{final\_3d\_rep}]\end{aligned} \tag{1}$$

The key operation is Score[3d_op(3d_rep), final_3d_rep], measuring models' ability to perform operations based on spatial instructions. Through uniform sampling of *3d_op* and *3d_rep*, we aim to offset spurious correlations at the perceptual level Score[I(3d_rep), 3d_rep] and Score[T(final_3d_rep), final_3d_rep], as well as spurious correlations based on specific linguistic configurations (e.g., based solely on T_{op}(3d_op)). This allows us to examine models' understanding of the **spatial operation concepts** Score[T(3d_op), 3d_op] and the dynamic functions of them Score[3d_op(3d_rep), final_3d_rep]. the statement of the **operation task** is as following:

> *If the cube rotates {90/270} degrees counterclockwise around its {visible face} face, the arrow on the initial {visible face} face of the cube will point to the* ***{orientation/wrong orientation}***.

Imagination Tasks We design two imagination tasks: *folding* and *matching* to examine whether models can understand cubes and their nets as constant objects in three-dimensional space that maintain consistency after arbitrary reasonable transformations.

The folding task examines the correspondence between cubes and their nets. Under EucliFold's three-dimensional cube view settings, each net has 24 possible final folding configurations, and the folding actions to achieve each configuration are arbitrary. Our text provides no feasible folding action sets or traversal strategies.

The matching task examines whether two nets can be folded into identical cubes. This task also has many feasible spatial operation schemes, such as folding both nets separately and then performing rotational matching, or traversing all local adjacency relationships. The statements for folding and matching tasks, respectively:

> *The cube in Image-2 can be formed by folding the net shown in Image-1.*
> *The cube net in Image-1 and the cube net in Image-2 can be folded into identical cubes.*

We control image content to offset shallow perceptual correlations. For folding tasks, we generate reference net images I_{net}(2d_rep(cube)) and correct folded images I_{cube}(3d_rep(cube)). For negative samples, randomly different cubes have excessive differences that allow models to easily exclude negative samples. Therefore, we employ a mutate-filtering approach to select perceptually similar images that do not belong to the same cube: $I_{\text{cube\_neg}}(\text{3d\_rep}) = I_{\text{cube}}(\text{3d\_rep\_similar})$ where 3d_rep_similar $\notin$ 3d_rep_set(cube).

Through uniform sampling of cube, *3d_rep*, and *2d_rep*, we aim to minimize the impact of spurious correlations at the perceptual level and examine models' ability for autonomous three-dimensional spatial matching.

$$\begin{aligned} &\text{Score}[I_{\text{cube}}(\text{3d\_rep(cube)}), I_{\text{net}}(\text{2d\_rep(cube)})] \\ &\sim \text{Score}[I_{\text{cube}}, \text{cube}] \times \text{Score}[I_{\text{net}}, \text{cube}] \\ &\quad \times \textbf{Score[cube\_spatial\_transformation]} \end{aligned} \tag{2}$$

The matching task does not inherently require more qualitative abilities than the folding task but poses greater pressure on working memory. The folding task only requires traversal matching of three cube faces, while matching requires traversal matching of six faces.

3.3 Winograd-Style Accuracy

For VLM evaluation, besides designing task distributions to eliminate superficial spurious correlations, we must also eliminate response bias Zheng et al. (2024). Specifically, VLM answer tokens True/False are influenced by the joint attention of all preceding tokens, many of which are unnecessary. These specific irrelevant contexts create strong tendencies toward True or False tokens.

To minimize the impact of such tendencies on measurement results, we adopt and adjust the metric from Winoground Thrush et al. (2022). We measure models' relative beliefs under specific contexts through answer differences between paired positive and negative samples with minimal necessary differences.

Specifically, for each pair of questions (pos, neg), we count two types of answer combinations:

- **True-belief:** (correct, correct), or equivalently, (True positive, True negative)
- **False-belief:** (incorrect, incorrect), or equivalently, (False negative, False positive)

The remaining two answer combinations cannot reflect the differences between paired questions, indicating that VLMs cannot effectively distinguish between them. We use the difference between true-belief and false-belief to measure model capability, termed:

$$\text{Winograd-Style Accuracy} = \mathbb{P}[\text{true\_belief}] - \mathbb{P}[\text{false\_belief}] \tag{3}$$

For confidence interval calculation, we assume $\mathbb{P}[\text{true\_belief}]$ and $\mathbb{P}[\text{false\_belief}]$ are independent. We calculate $\text{VAR} = \hat{p} \times (1 - \hat{p})/n$ separately, thus $\text{VAR}(\text{Winograd-Style Accuracy}) = \text{VAR}(\text{true\_belief}) + \text{VAR}(\text{false\_belief})$. Finally, we use normal distribution approximation Z-scores to obtain confidence intervals.

4 Experiments and Analysis

Model Selection. We choose GPT-4o (Hurst et al., 2024), Gemini-2.5-flash (Team et al., 2023), and Claude-3.5-sonnet (Anthropic, 2024) as representative high-performance closed-source VLMs. We select the Qwen-VL-2.5 series (Bai et al., 2025) as representative general-purpose VLMs and the InternVL-2.5 series (Chen et al., 2024) as representative post-trained visual reasoners.

Experimental Settings. For each model, we employ system prompts to control chain-of-thought reasoning (Kojima et al., 2022), followed by true-false answers. More specific parameter configurations are detailed in Appendix D. Each datapoint in this section represent around 1,200 samples, for detail, see Appendix B.

4.1 General Performance and the Effect of Scaling

As shown in Figure 1, overall, most of the open-source and closed-source models demonstrate relatively high score at the perception level, though gaps remain compared to human performance. At the operation level, only Gemini and Claude show performance significantly above chance level, yet still far weaker than humans. At the imagination level, only InternVL-78B and Qwen-72B perform significantly above chance level, still substantially below human performance.

Perception Tasks. The color recognition task at the perception level primarily judges whether models can distinguish different faces of cubes. We ask models about the color of specific cube faces (e.g., top face arrow color) and set negative colors from the other two visible faces as distractors. As shown in Figure 3, except for extremely small models, most achieve near-perfect performance.

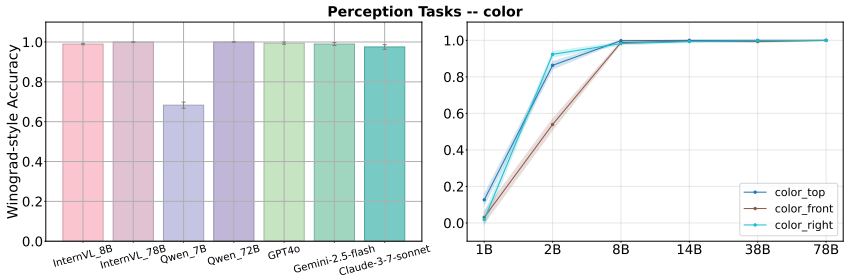

Figure 3: Performance scaling on color recognition tasks across different model sizes.

The arrow orientation recognition task primarily judges whether models can understand the orientation of two-dimensional objects in three-dimensional space. As shown in Figure 4, accuracy shows an increasing trend with model scale but does not reach perfect levels. This may be due to VLMs' insufficient accuracy in perceiving fine-grained content in images (Fu et al., 2024), or lack of internal three-dimensional spatial concepts, preventing proper grounding of image content to sufficiently discriminative three-dimensional spatial representations.

On the other hand, models show significant differences in perception accuracy across different cube faces. Since color task accuracy saturates at 8B parameters, indicating that models can perfectly distinguish the three faces, the orientation perception accuracy differences may be due to varying distortions of the arrows, resulting in different perception difficulties.

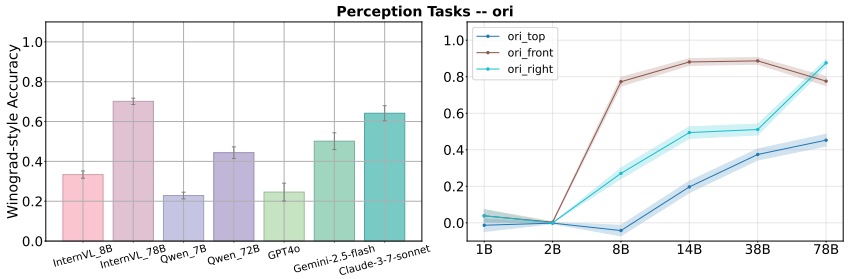

Figure 4: Performance scaling on orientation recognition tasks across different model sizes and cube faces.

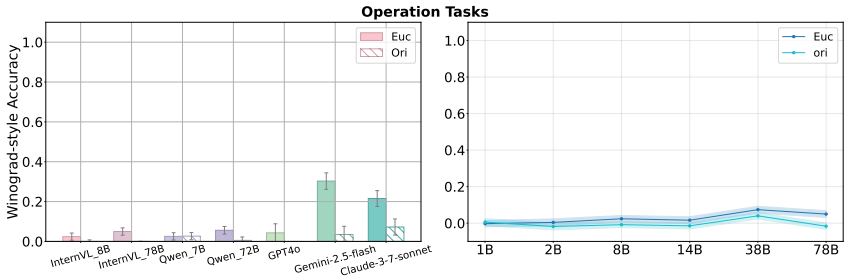

Figure 5: Comparison of direct answer vs. chain-of-thought prompting on folding tasks.

4.1.1 Operation Tasks: Prompting

As illustrated in Figure 5, only two close-source models (Gemini-2.5-flash and Claude-3.7-sonnet) perform above chance-level. The performance gap between perception and operation tasks indicates that while models can recognize static spatial configurations to some extent, they struggle with dynamic spatial transformations.

Ablation Study on Prompting. Parallel to orientation tasks, we change the description style to Euclidean terms, forming another set of questions (Euc in short). The results in Figure 5 indicate a significant performance gain when shifting from natural-style language (ori) to mathematical style (Euc). Although we implement multiple strategies to avoid various biases in VLM evaluation, we cannot fully control prompt-induced biases in VLMs' internal chain-of-thought preferences.

4.1.2 Imagination Tasks: Failure in Complex Spatial Reasoning

As shown in Figure 1, folding tasks show that only the InternVL series can exceed chance level, possibly benefiting from similar tasks in InternVL's post-training process that enable solving partial problems.

As mentioned earlier, InternVL's folding ability may stem from learning a specific non-generalizable strategy rather than possessing genuine generalizable spatial imagination capabilities. We verify this hypothesis through two ablation studies.

Ablation Study on Chain-of-Thought Prompting We test InternVL's accuracy on folding tasks using direct prompting. Since generalizable spatial folding strategies must involve multi-step traversal and enumeration, direct-answer approaches cannot encode such variable-length strategies (Merrill et al., 2022). Therefore, as shown in Figure 6, when direct-answer format approaches chain-of-thought accuracy, it must be utilizing a fixed strategy to complete partial spatial matching.

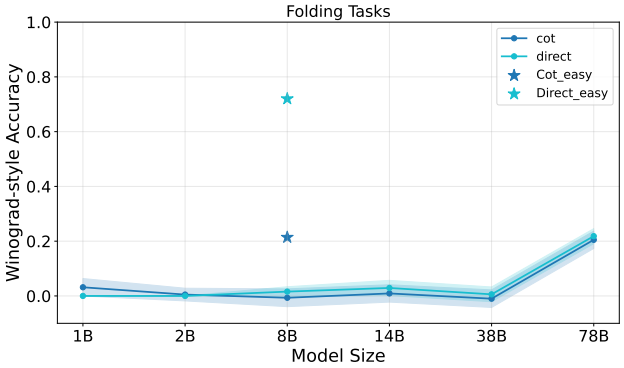

Figure 6: Ablation study on folding tasks: direct vs. chain-of-thought prompting and leaked vs. proper negative samples.

Ablation Study on easy negative samples We design leaked negative samples for folding tasks, where negative cube samples contain colors not present in the net, allowing models to use simple strategies to judge negative samples (comparing overall color sets between cube and net). Results show that even 8B models achieve accuracy significantly above chance level on leaked negative sample settings.

These two ablation experiments demonstrate that InternVL lacks imagination-level Euclidean priors and emphasize the necessity of precisely controlling negative samples.

5 Related Work

5.1 Large Language Model-based Visual Language Models

With the rapid development of large language models (LLMs) and the prohibitive cost of training large models from scratch, researchers increasingly build Visual Language Models (VLMs) by integrating visual encoders with pre-trained LLMs. This approach inherits world knowledge and reasoning capabilities from the underlying language model. Pioneer studies such as BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023b), and MiniGPT-4 (Zhu et al., 2023) demonstrate significantly more robust instruction-following and broader zero-shot capabilities compared to previous train-from-scratch VLMs like CoCa (Yu et al., 2022). These advances have substantially expanded the scope of zero-shot visual question answering, catalyzing the development of comprehensive benchmarks such as MME (Zhang et al., 2021), MMMU (Yue et al., 2024), and MMBench (Liu et al., 2024) to evaluate VLM capabilities across diverse domains. Evaluations reveal that while current VLMs excel at OCR and visual grounding, they struggle with mathematical reasoning and real-world understanding (Chen et al., 2024; Bai et al., 2025). Spatial reasoning tasks particularly expose these limitations, requiring both geometric understanding and logical inference—two areas where current VLMs show systematic weaknesses.

5.2 Spatial Reasoning Benchmarks for VLMs

While general-purpose VLM benchmarks like MMBench (Liu et al., 2024) contain spatial understanding tasks, several benchmarks specifically target VLM spatial capabilities. 3DSRBench (Ma et al., 2024) collects 2,772 human-annotated questions about 3D spatial reasoning, revealing that leading VLMs achieve only 50% accuracy compared to 90% human performance. LEGO-Puzzles (Tang et al., 2025) curates 1,100 questions from generated LEGO images, showing GPT-4o achieves 60% accuracy while humans reach 93.6%. VSI-Bench (Yang et al., 2025) evaluates VLM cognitive mapping with video inputs. Beyond evaluation, some benchmarks aim to improve spatial abilities through fine-tuning. Spatial Aptitude Training (SAT) (Ray et al., 2024) and Sparkle (Tang et al., 2024) demonstrate non-trivial performance gains but remain far from human-level performance. These findings suggest systematic limitations in VLM spatial reasoning. Our work probes

a fundamental source of this deficiency: the lack of robust 3D Euclidean priors that enable flexible spatial understanding.

6 Conclusion and Discussion

Our evaluation reveals fundamental limitations in current VLMs' spatial reasoning capabilities. While state-of-the-art models demonstrate reasonable performance at the Perception level, they fail dramatically at Operation and Imagination levels, achieving only X% and Y% accuracy respectively compared to near-perfect human performance. This stark capability gap suggests that VLMs rely primarily on statistical pattern matching rather than developing genuine geometric understanding of 3D transformations. Our cognitively-stratified framework effectively isolates these different levels of capability, revealing that spatial reasoning deficits are not uniform but concentrated in higher-order operations requiring mental manipulation of spatial representations. These findings align with cognitive neuroscience research that spatial pirors are mainly on distinct neural circuits (O'Keefe & Dostrovsky, 1971; Moser et al., 2008), while challenging the assumption that scaling data and parameters alone will bridge the human-AI gap in spatial reasoning.

This work contributes EucliFold, a cognitively-inspired benchmark for evaluating 3D spatial reasoning in VLMs, along with a bias-resistant evaluation methodology that minimizes confounding factors. Our three-level decomposition framework offers a principled approach to capability assessment that could be adapted to other cognitive domains. The systematic nature of current failures across different model architectures suggests that achieving human-level spatial intelligence may require architectural innovations or training paradigms that explicitly incorporate geometric inductive biases rather than incremental improvements to existing approaches. While our cube net domain provides rigorous controlled evaluation, future work should investigate generalization to other spatial reasoning tasks and explore whether explicit geometric training can address the fundamental limitations we identify. Our methodology demonstrates how insights from cognitive science can inform AI evaluation, potentially leading to more robust benchmarks for assessing genuine reasoning capabilities beyond pattern matching.

Author Contributions

If you'd like to, you may include a section for author contributions as is done in many journals. This is optional and at the discretion of the authors.

Acknowledgments

Use unnumbered third level headings for the acknowledgments. All acknowledgments, including those to funding agencies, go at the end of the paper.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. Can language models encode perceptual structure without grounding? a case study in color. *arXiv preprint arXiv:2109.06129*, 2021.

AI Anthropic. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source

multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2024.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

Yufei Huang, Shengding Hu, Xu Han, Zhiyuan Liu, and Maosong Sun. Unified view of grokking, double descent and emergent abilities: A perspective from circuits competition. *arXiv preprint arXiv:2402.15175*, 2024.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. *KR*, 2012(13th):3, 2012.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. Revisiting the role of language priors in vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 29914–29934, 2024.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023a.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

Julius Mayer, Mohamad Ballout, Serwan Jassim, Farbod Nosrat Nezami, and Elia Bruni. ivispar–an interactive visual-spatial reasoning benchmark for vlms. *arXiv preprint arXiv:2502.03214*, 2025.

William Merrill, Ashish Sabharwal, and Noah A Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.

Edvard I Moser, Emilio Kropff, and May-Britt Moser. Place cells, grid cells, and the brain's spatial representation system. *Annu. Rev. Neurosci.*, 31(1):69–89, 2008.

John O'Keefe and Jonathan Dostrovsky. The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*, 1971.

Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Roger N Shepard and Jacqueline Metzler. Mental rotation of three-dimensional objects. *Science*, 171(3972):701–703, 1971.

Kexian Tang, Junyao Gao, Yanhong Zeng, Haodong Duan, Yanan Sun, Zhening Xing, Wenran Liu, Kaifeng Lyu, and Kai Chen. Lego-puzzles: How good are mllms at multi-step spatial reasoning? *arXiv preprint arXiv:2503.19990*, 2025.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning. *arXiv preprint arXiv:2410.16162*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, et al. Code2logic: Game-code-driven data synthesis for enhancing vlms general reasoning. *arXiv preprint arXiv:2505.13886*, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

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

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*, 2020.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions*, 2022, 2022.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Jiahuan Zhang, Shunwen Bai, Tianheng Wang, Kaiwen Guo, Kai Han, Guozheng Rao, and Kaicheng Yu. Ascending the infinite ladder: Benchmarking spatial deformation reasoning in vision-language models. *arXiv preprint arXiv:2507.02978*, 2025.

Yunhang Shen Yulei Qin Mengdan Zhang, Xu Lin Jinrui Yang Xiawu Zheng, Ke Li Xing Sun Yunsheng Wu, Rongrong Ji Chaoyou Fu, and Peixian Chen. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2021.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *ICLR*, 2024.

Ziqian Zhong, Ziming Liu, Max Tegmark, and Jacob Andreas. The clock and the pizza: Two stories in mechanistic explanation of neural networks. *Advances in neural information processing systems*, 36:27223–27250, 2023.

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

A Appendix: Data Generation Pipeline Details

There are 5 steps to generate all configurations and images. **(0) Spatial representation design.** We establish a three-dimensional representation (*3d_rep*) and two-dimensional representation (*2d_rep*) to encode cube spatial states and net configurations, respectively.

(1) Distinct cube generation. We randomly generate arrow patterns on cube surfaces, then apply the 24 rotational group transformations of *3d_rep* to produce 24 equivalent representations. After comparing against existing cube representations and filtering duplicates, we obtain a collection of distinct cubes.

(2) Positive 3D view generation. Each cube corresponds to 24 possible *3d_rep* configurations. Since only three faces are visible in any view, we filter cubes that would produce duplicate 3D view images.

(3) Positive 2D net generation. Each 3D-rotational variant of a cube corresponds to 11 distinct two-dimensional nets (excluding 2D-rotational and 2D-mirror symmetries), yielding $24 \times 11 = 264$ possible net configurations per cube.

(4) Negative sample generation. To produce sufficiently challenging negative samples, we employ a *mutation-validation* approach rather than random generation, ensuring that negative images maintain plausible appearance similar to positive images while violating geometric constraints. The detailed methodology is presented in the following section.

B Appendix: Data Generation Pipeline Details

For data generation, we first create 50 distinct cubes, then generate three-dimensional cube views for each of the 24 three-dimensional representations per cube, yielding 1,200 cube images in total. Upon inspection, the image duplication rate is below 5%. For each image, we generate corresponding positive and negative samples for perception and operation questions, resulting in 1,200 pairs per question type.

For folding questions, we randomly rotate each cube image, then randomly select a method to unfold it, obtaining a net image as reference. We then randomly mutate the color or shape of one face to generate a negative cube image, thus obtaining paired samples. For matching questions, we randomly select 24 nets for each cube, then perform mutations, also yielding 1,200 question pairs.

C Appendix: Question Templates

Template for one-image tasks:

Image: {image of a 3D cube view}
Question: *Based on the image and the description of the image, is the following statement True or False?*
Description: *The image shows a cube with three visible faces (top, front, right), each face has an arrow on it.*
Statement: {question-specific statement}
Answer: {possible chain-of-thought} **{True or False}**

Template for two-images tasks:

Image-1: {image-1}
Image-2: {image-2}
Question: *Based on Image-1 and Image-2, is the following statement True or False?*
Statement:{question-specific statement}
Answer: {possible chain-of-thought} **{True or False}**

D Appendix: VLMs Evaluation Configurations

System Prompt Settings. As for the main experiment, we use the same chain-of-though style System Prompt for all models. The *perception* and *operation* tasks contains one image, *imagination* tasks contains two images.

System Prompt for *chain-of-though* setting

The following is a True/False question based on **{an image/two images}**. Analyze the image and the question carefully, then determine if the statement is True or False. Provide your reasoning step by step.

System Prompt for *direct-answer* setting

The following is a True/False question based on **{an image/two images}**. Directly output only 'True' or 'False' as your answer. Do not provide any reasoning, explanation, or additional text.

System Prompt for *short-style* setting

The following is a True/False question based on **{an image/two images}.**

System Prompt for *base-style* setting

You are a helpful assistant.

Chain-of-Thought Settings. We use temperature=0.0 for chain-of-thought text generation and temperature=1.0 for direct text generation.

E Appendix: Human Performance

We sample 60 examples each from perception, operation, folding, and matching tasks, totaling 240 samples. We constructed a visualization webpage for testing, selecting university student volunteers as subjects. Each subject randomly answered 16 questions (4 from each category), and only complete responses were considered valid. We received 19 valid response batches, totaling 304 questions. The final distribution of answered questions was 54 from perception, 50 from operation, 54 from folding, and 50 from matching.

For human subjects, seeing paired questions simultaneously provides additional useful information, so we do not use paired questions to test humans, assuming humans have no response bias. We estimate corresponding Winograd-Style Accuracy through human positive accuracy and negative accuracy:

$$\begin{aligned} \mathbb{P}[\text{true\_belief}] &\sim \mathbb{P}[\text{pos\_acc}] \times \mathbb{P}[\text{neg\_acc}] \\ \mathbb{P}[\text{false\_belief}] &\sim (1 - \mathbb{P}[\text{pos\_acc}]) \times (1 - \mathbb{P}[\text{neg\_acc}]) \\ \text{Human Winograd-Style Accuracy} &= \mathbb{P}[\text{true\_belief}] - \mathbb{P}[\text{false\_belief}] \end{aligned} \tag{4}$$

F APPENDIX: LLM USAGE

Large Language Models (LLMs) were used as auxiliary tools for code refinement and text polishing. All LLM-generated content, including code and written text, was rigorously reviewed and validated by at least one author.