EXTENDED ABSTRACT
World2Mind: Cognition Toolkit for Allocentric Spatial Reasoning in Foundation Models

Anonymous CVPR submission

Paper ID *****

Abstract

Achieving robust spatial reasoning remains a fundamental challenge for current Multimodal Foundation Models (MFMs). Existing methods either overfit statistical shortcuts via 3D grounding data or remain confined to 2D visual perception, limiting both spatial reasoning accuracy and generalization in unseen scenarios. Inspired by the spatial cognitive mapping mechanisms of biological intelligence, we propose ***World2Mind****, a training-free spatial intelligence toolkit. At its core, World2Mind leverages 3D reconstruction and instance segmentation models to construct structured spatial cognitive maps,* ***empowering MFMs to proactively acquire targeted spatial knowledge regarding interested landmarks and routes of interest.*** *To provide robust geometric-topological priors, World2Mind synthesizes an* ***Allocentric-Spatial Tree (AST)*** *that uses elliptical parameters to model the top-down layout of landmarks accurately. To mitigate the inherent inaccuracies of 3D reconstruction, we introduce a three-stage reasoning chain comprising tool invocation assessment, modality-decoupled cue collection, and geometry-semantics interwoven reasoning. Extensive experiments demonstrate that World2Mind boosts the performance of frontier models, such as GPT-5.2, by 5%∼18%. Astonishingly, relying solely on the AST-structured text, purely text-only foundation models can perform complex 3D spatial reasoning, achieving performance approaching that of advanced multimodal models.*

1. Introduction

Although multimodal foundation models (MFMs) [1, 12, 24, 28] excel in general visual understanding and cross-modal reasoning [29], they struggle significantly in embodied AI and complex spatial reasoning tasks requiring physical interaction [15, 17, 18, 26, 34]. This deficiency stems from their over-reliance on egocentric observations and lacking the capacity to abstract global spatial topology [18, 34], trapping MFMs in an insurmountable "semantic-geometry gap" in tasks like distance estimation, viewpoint transformation, and path planning.

Current efforts to enhance MFMs' spatial reasoning primarily follow two paradigms. ***Training-based methods*** [6, 9, 20] fine-tune models on massive 3D-grounded QA pairs. However, this forces models to overfit statistical shortcuts [14, 25] rather than acquiring genuine spatial cognition, leading to poor generalization in out-of-distribution scenarios [33]. Alternatively, introducing explicit 3D modalities [7, 10, 21, 31] exacerbates inter-modal alignment challenges [35]. Meanwhile, recent ***tool-based methods*** [8, 19, 36] rely on active rendering under large-scale 3D reconstruction. These are severely bottlenecked by reconstruction quality and remain tethered to low-level visual perception, failing to abstract geometric data into structured semantics for high-level logical reasoning.

Biological Intelligence (BI) offers an ideal blueprint to break the shackles of egocentric observation. Rather than reacting passively to transient visual inputs, the biological brain intrinsically transforms egocentric views into an allocentric perspective [4]. Supported by place cells in the hippocampus [3, 11, 22] and grid cells in the entorhinal cortex [13], mammals construct a global ***cognitive map*** entirely independent of their egocentric viewpoint [23, 27]. This map forms the cornerstone for strategic mental simulation and advanced reasoning [2].

To bridge this gap between foundation models and BI in spatial representation and reasoning, we propose ***World2Mind***, a **plug-and-play spatial cognition toolkit** that equips models with human-like mental simulation capabilities. World2Mind integrates an efficient geometry-semantics alignment pipeline, leveraging pre-trained visual geometry [16, 30, 32] and instance segmentation models [5] to extract semantic voxel grids. From these, it constructs two core representations: ***1)*** a Route Cognitive Map for passability prediction, and ***2)*** a Landmark Cognitive Map for object topology. Encapsulated as an accessible toolset,

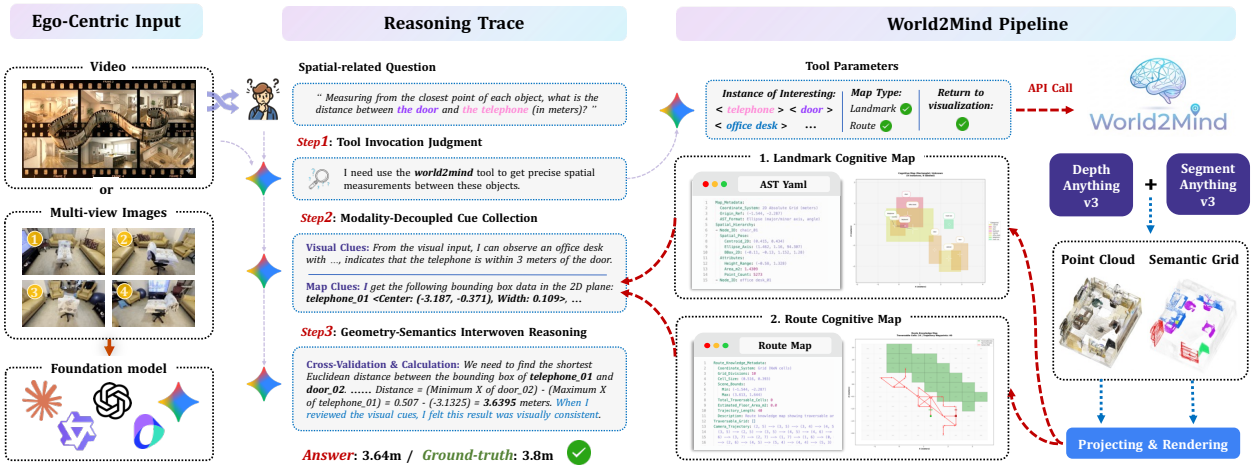

Figure 1. **Overview of foundation models performing allocentric spatial reasoning via the proposed World2Mind toolkit.** Given egocentric video or multi-view observations, the model first assesses the necessity of tool invocation and subsequently passes key parameters (*e.g.*, instances of interest) to World2Mind to drive the generation of spatial cognitive maps. World2Mind integrates an efficient pipeline for 3D reconstruction and semantic-geometry alignment, returning the required structured spatial knowledge through targeted projection and rendering mechanisms. Furthermore, the model conducts geometry-semantics interwoven reasoning based on both the raw visual observations and the geometric cues provided by World2Mind, ultimately yielding highly reliable answers.

World2Mind enables models to dynamically specify parameters (*e.g.*, instances of interest, required spatial knowledge, and map visualizations) to proactively acquire targeted allocentric spatial knowledge on demand.

To provide robust geometric-topological priors, we formally define the **Allocentric-Spatial Tree (AST)** as the core spatial representation in World2Mind. The AST is a directed acyclic graph utilizing geometrically stable landmarks (*e.g.*, beds, tables) as core nodes to hierarchically associate surrounding smaller instances. Crucially, to approximate the fuzzy nature of human cognition, the AST models spatial footprints using rectangle-elliptical parameters (bounding boxes, major/minor axes, eccentricity, and rotation angles). These designs equip models with robust, dense, and highly actionable geometric-topological priors.

However, merely offering spatial representation is insufficient to guarantee robust reasoning. In complex physical scenarios, reconstruction quality often suffers severe corruption due to occlusions or restricted viewpoints, leading to conflicts with objective geometric laws and raw visual observations. To mitigate this risk, we integrate a rigorous ***spatial reasoning chain*** into World2Mind: ***1)*** Difficulty Assessment and Tool Invocation, preventing over-computation on simple superficial queries; ***2)*** Modality-Decoupled Cue Collection, independently extracting information from egocentric vision, AST structured text, and map visualizations; and ***3)*** Geometry-Semantics Interwoven Reasoning, guiding the model to resolve cross-modal conflicts proactively and ultimately yield reliable spatial decisions.

Extensive evaluations across various spatial reasoning benchmarks demonstrate that World2Mind yields stable performance improvements of **6%–18%** for frontier models like GPT-5.2, while maintaining exceptional efficiency and reasoning interpretability. Astonishingly, leveraging the pure, high-density allocentric priors provided by the AST, text-only foundation models can execute complex 3D reasoning directly within their parameter space simply by reading the AST representation, approaching the performance of advanced multimodal models. Our findings offer a highly promising pathway to overcome the spatial cognition bottleneck in foundation models.

2. Method Overview

This section details the technical overview of the proposed world2mind, as illustrated in Fig. 1

2.1. Geometry-Semantic Alignment Pipeline

Given an egocentric video sequence or multi-view image set $\{I_t\}_{t=1}^{T}$, our primary objective is to transcend the limitations of 2D vision and construct a robust 3D semantic representation of the physical world.

➊ **Depth Estimation & Semantic Extraction.** We employ ***Depth Anything V3*** [16] for monocular depth estimation, obtaining the depth map $D_t \in \mathbb{R}^{H\times W}$ and camera pose $T_t \in SE(3)$. Concurrently, we utilize ***SAM3*** [5] to extract open-vocabulary semantic masks M_t based on a user-specified category list $\mathcal{C}$. To suppress the accumulation of long-tail errors inherent in depth estimation, we introduce a dual-level filtering mechanism based on the predicted confidence map $C_t \in [0,1]^{H\times W}$. Specifically, we formulate a

Table 1. **Main results on the VSI-Bench [34] benchmark (Tiny subset).**

Models	Avg.	Numerical Answer (%)				Multiple-Choice Answer (%)			
		Obj. Count	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Frontier models w/o. World2Mind									
GPT-5.2	46.7	52.5	34.9	67.5	50.6	42.0	40.7	34.7	51.0
Claude-4.6-Opus	38.4	46.9	18.5	62.1	26.8	40.0	47.2	34.7	30.6
Gemini-3-Pro	55.2	47.8	32.1	**71.3**	55.0	54.0	44.8	57.1	79.6
Frontier models w./ World2Mind									
GPT-5.2	54.0 (↑7.3)	47.4 (↓5.1)	33.4 (↓1.5)	63.3 (↓4.2)	52.4 (↑1.8)	**64.0** (↑22.0)	41.1 (↑0.4)	51.0 (↑16.3)	79.6 (↑28.6)
Claude-4.6-Opus	56.0 (↑17.7)	**59.0** (↑12.0)	34.3 (↑15.8)	67.3 (↑5.2)	54.8 (↑28.0)	**64.0** (↑24.0)	62.7 (↑15.6)	**65.3** (↑30.6)	40.8 (↑10.2)
Gemini-3-Pro	**61.0** (↑5.8)	51.8 (↑4.1)	**36.8** (↑4.7)	57.7 (↓13.5)	**62.6** (↑7.6)	62.0 (↑8.0)	**67.7** (↑22.9)	**65.3** (↑8.2)	**83.7** (↑4.1)

Table 2. **Results on the MindCube-Tiny [18] benchmark.**

Models	Avg.	Around	Among	Rotation
Frontier models w/o. World2Mind				
GPT-5.2	49.9	62.4	45.2	48.5
Claude-4.6-Opus	48.5	58.8	50.7	29.0
Gemini-3-Pro	75.1	77.2	68.2	93.0
Frontier models w./ World2Mind				
GPT-5.2	54.6 (↑4.7)	60.4 (↓2.0)	47.7 (↑2.5)	68.0 (↑19.5)
Claude-4.6-Opus	62.9 (↑14.4)	82.4 (↑23.6)	60.8 (↑10.1)	45.0 (↑16.0)
Gemini-3-Pro	**81.6** (↑6.5)	**86.0** (↑8.8)	**75.8** (↑7.6)	**93.5** (↑0.5)

binary validity mask $\mathcal{V}_t \in \{0,1\}^{H\times W}$ as follows:

$$\mathcal{V}_t(u,v) = \mathbb{I}\left(C_t(u,v) > \tau_{\text{pixel}}\right) \cdot \mathbb{I}\left(\mu_t > \tau_{\text{frame}}\right), \quad (1)$$

where $\mu_t = \frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W} C_t(x,y)$ denotes the global spatial confidence of frame t, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition is met and 0 otherwise. The variables τ_{pixel} and τ_{frame} represent the pixel-level and frame-level thresholds, respectively. A pixel is incorporated into the subsequent reconstruction only when $\mathcal{V}_t(u,v) = 1$.

❷ Point Cloud Mapping & Density Filtering. Qualifying 2D pixels are back-projected into the world coordinate system via the camera intrinsic matrix K, generating a global point cloud $\mathcal{P} = \{(\mathbf{p}_i, s_i, \mathbf{rgb}_i)\}_{i=1}^{M}$ carrying semantic labels $s_i \in \mathcal{C}$. Addressing the boundary outliers inherent in depth estimation, we propose a core region extraction strategy: for each point, we calculate its K-nearest neighbor local density $\rho_i = \frac{1}{K}\sum_{j\in\mathcal{N}_K(i)} \|\mathbf{p}_i - \mathbf{p}_j\|^{-1}$, and eliminate low-density "tail" points based on density percentiles. This yields an exceptionally pure geometry-semantic substrate.

2.2. Allocentric Cognitive Mapping

Inspired by the spatial mapping mechanisms of BI, we distill the unstructured point cloud into two highly abstract cognitive maps, enabling the model to proactively acquire spatial knowledge on demand via tool invocation.

❶ Landmark Cognitive Mapping. Traditional methods rely on ambiguous relative relations or simplified grid representation [18, 34]. To overcome this, we formally define the ***Allocentric-Spatial Tree (AST)***, which reorganizes spatial entities as a directed acyclic graph within an absolute coordinate system. Specifically, we perform adaptive DBSCAN clustering on each semantic category within the point cloud to separate distinct instances. For each instance node, the AST discards traditional bounding boxes and instead fits a minimum bounding ellipse in the top-down view (X-Z plane), extracting the centroid (x_c, z_c), major and minor axes a and b, and rotation angle θ. This parameterization: ***1)*** significantly enhances robustness against reconstruction boundary noise; ***2)*** perfectly aligns with the fuzzy probability nature of human spatial footprint perception. Output as dense structured text (e.g., YAML), the AST explicitly encodes hierarchical containment relationships among entities along with multi-dimensional geometric attributes.

❷ Route Cognitive Mapping. For navigation-oriented tasks, World2Mind also enables extracting the masks of traversable categories (e.g., floors), back-projects and voxelizes them, and subsequently partitions them into an $N\times N$ grid map on the top-down plane. Combined with the mapping of the camera trajectory sequence $\{T_t\}$, this route map provides the model with explicit priors regarding passability and the human observer's motion trajectory.

2.3. Geometry-Semantics Interwoven Reasoning

In physical scenarios, 2D visual observations are susceptible to occlusions and adverse viewpoints, while 3D reconstruction information may contain local errors. To resolve potential contradictions between these two modalities, we design a rigorous three-stage interwoven reasoning chain.

Stage 1: Tool Invocation Judgement. To reduce unnecessary computational overhead, the model must first evaluate the spatial relevance of the query. The model should proactively invoke World2Mind only when the task explicitly involves spatial reasoning, such as occlusion inference, distance estimation, or path planning.

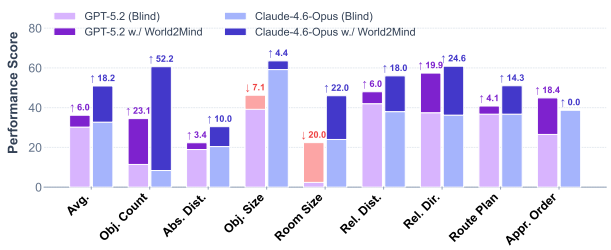

Figure 2. **Performance comparison under the text-only model ("blind") setting.** we report the performance gap on the VSI-Bench (Tiny subset) between foundation models relying solely on commonsense reasoning and those leveraging world2mind to acquire structured spatial knowledge for allocentric reasoning.

Stage 2: Modality-Decoupled Cue Collection. We force the model to extract information independently to prevent early modality bias. The model must simultaneously gather corroborating evidence from three independent sources: egocentric vision, the AST text returned by World2Mind, and optional 2D top-down map visualizations.

Stage 3: Conflict Resolution and Cross-Validation. This is the crux of the reasoning chain. The model needs to proactively coordinate evidence across different modalities and identify cross-modal conflicts, such as missing objects due to visual truncation or coordinate drift caused by depth errors. By cross-validating the objective geometric parameters of the AST against subjective visual appearances, the model can dynamically weigh the credibility between visual illusions and reconstruction artifacts, ultimately outputting highly reliable and logically interpretable decisions.

3. Experiment

3.1. Experimental Setup

We conduct evaluations on two challenging spatial reasoning benchmarks: VSI-Bench [34], which focuses on video-based reasoning in real-world physical scenes, and MindCube [18], which emphasizes multi-view cognitive mapping and mental simulation. Our evaluation primarily targets the most frontier multimodal foundation models, including GPT, Claude, and Gemini, given their exceptional proficiency in tool invocation and instruction-following.

3.2. Main Results and Analysis

Evaluation on VSI-Bench. As shown in Tab. 1, the seamless integration of World2Mind yields a substantial leap in the average performance (Avg.) across all frontier models. Specifically, GPT-5.2 achieves a 7.3% improvement, while Claude-4.6-Opus achieves a staggering 17.7% improvement. In-depth analysis of the subtasks reveals that the performance gains are most striking in task categories that rely heavily on allocentric priors. For instance, on Relative Direction, Route Planning, and Relative Distance, the performance of Claude-4.6-Opus skyrockets by 15.6%, 30.6%, and 24.0%, respectively. This compellingly demonstrates the critical role of the allocentric spatial knowledge provided by World2Mind in bridging the spatial reasoning gap.

Figure 3. Complete reasoning trace under World2Mind.

Evaluation on MindCube. The results in Tab. 1 further corroborate the universality and robustness of our framework in sparse multi-view inputs. Even for Gemini-3-Pro, whose native spatial reasoning capability is already top-tier (with a baseline Avg. of 75.1%), World2Mind successfully shatters its performance ceiling, pushing its average accuracy to 81.6% (+6.5%). Notably, in tasks like "Rotation" that severely test 3D spatial imagination, the model achieves a remarkable performance breakthrough (GPT-5.2 improves by 19.5%) due to its ability to perform logical deduction grounded in the AST.

3.3. Ablation and Case Study

To explore the limits of how structured text of AST empowers the spatial cognition of large language models, we follow [34] to conduct ablation studies under the text-only ("blind") setting (see Fig. 2). When visual image inputs are completely stripped away, foundation models that rely solely on commonsense priors degrade to near-random guessing on spatial tasks. Astonishingly, however, when equipped with World2Mind, both GPT-5.2 and Claude-4.6-Opus exhibit a remarkable performance rebound in the "blind" state. On core reasoning tasks such as Object Size and Route Planning, their scores closely approach those achieved with full visual inputs. This profound finding ***indicates that pure, high-quality allocentric geometric priors are entirely sufficient to ignite powerful 3D mental reconstruction and simulation capabilities of foundation models under the text-based reasoning.*** Furthermore, we visualize the complete reasoning traces powered by World2Mind in Fig. 3, which clearly demonstrate that the interwoven reasoning process exhibits exceptional robustness and logical interpretability when resolving cross-modal conflicts.

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

[2] Jacob LS Bellmund, Peter Gärdenfors, Edvard I Moser, and Christian F Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018. 1

[3] Nicola J Broadbent, Larry R Squire, and Robert E Clark. Spatial memory, recognition memory, and the hippocampus. *Proceedings of the National Academy of Sciences*, 101(40): 14515–14520, 2004. 1

[4] Neil Burgess. Spatial memory: how egocentric and allocentric combine. *Trends in cognitive sciences*, 10(12):551–557, 2006. 1

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. 1, 2

[6] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1

[7] Pingyi Chen, Yujing Lou, Shen Cao, Jinhui Guo, Lubin Fan, Yue Wu, Lin Yang, Lizhuang Ma, and Jieping Ye. Sd-vlm: Spatial measuring and understanding with depth-encoded vision-language models. *arXiv preprint arXiv:2509.17664*, 2025. 1

[8] Siyi Chen, Mikaela Angelina Uy, Chan Hee Song, Faisal Ladhak, Adithyavairavan Murali, Qing Qu, Stan Birchfield, Valts Blukis, and Jonathan Tremblay. Spacetools: Tool-augmented spatial reasoning via double interactive rl. *arXiv preprint arXiv:2512.04069*, 2025. 1

[9] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 1

[10] Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, et al. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408, 2025. 1

[11] Howard Eichenbaum. The role of the hippocampus in navigation is memory. *Journal of neurophysiology*, 117(4):1785–1796, 2017. 1

[12] Google. Gemini 3.1 pro: Best for complex tasks and bringing creative concepts to life. https://deepmind.google/models/gemini/pro/, 2026. 1

[13] Torkel Hafting, Marianne Fyhn, Sturla Molden, May-Britt Moser, and Edvard I Moser. Microstructure of a spatial map in the entorhinal cortex. *Nature*, 436(7052):801–806, 2005. 1

[14] Jiaxin Huang, Ziwen Li, Hanlve Zhang, Runnan Chen, Xiao He, Yandong Guo, Wenping Wang, Tongliang Liu, and Mingming Gong. Surprise3d: A dataset for spatial understanding and reasoning in complex 3d scenes. *arXiv preprint arXiv:2507.07781*, 2025. 1

[15] Shuai Huang, Wenxuan Zhao, and Jun Gao. Si-bench: Benchmarking social intelligence of large language models in human-to-human conversations. *arXiv preprint arXiv:2510.23182*, 2025. 1

[16] Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025. 1, 2

[17] Jingli Lin, Runsen Xu, Shaohao Zhu, Sihan Yang, Peizhou Cao, Yunlong Ran, Miao Hu, Chenming Zhu, Yiman Xie, Yilin Long, et al. Mmsi-video-bench: A holistic benchmark for video-based spatial intelligence. *arXiv preprint arXiv:2512.10863*, 2025. 1

[18] Fangzheng Liu, Don Derek Haddad, and Joe Paradiso. Mindcube: an interactive device for gauging emotions. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–2, 2024. 1, 3, 4

[19] Zhanpeng Luo, Ce Zhang, Silong Yong, Cunxi Dai, Qianwei Wang, Haoxi Ran, Guanya Shi, Katia Sycara, and Yaqi Xie. pyspatial: Generating 3d visual programs for zero-shot spatial reasoning. *arXiv preprint arXiv:2603.00905*, 2026. 1

[20] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*, 2025. 1

[21] Zhenhua Ning, Zhuotao Tian, Shaoshuai Shi, Guangming Lu, Daojing He, Wenjie Pei, and Li Jiang. Enhancing spatial reasoning in multimodal large language models through reasoning-based segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7851–7860, 2025. 1

[22] 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. 1

[23] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978. 1

[24] OpenAI. Gpt-4v(ision) system card. https://cdn.openai.com/papers/GPTV_System_Card.pdf, 2023. 1

[25] Jianing Qi, Jiawei Liu, Hao Tang, and Zhigang Zhu. Beyond semantics: Rediscovering spatial awareness in vision-language models. *arXiv preprint arXiv:2503.17349*, 2025. 1

[26] Santhosh Kumar Ramakrishnan, Erik Wijmans, Philipp Kraehenbuehl, and Vladlen Koltun. Does spatial cognition emerge in frontier models? *arXiv preprint arXiv:2410.06468*, 2024. 1

[27] Daniela Schiller, Howard Eichenbaum, Elizabeth A Buffalo, Lila Davachi, David J Foster, Stefan Leutgeb, and Charan Ranganath. Memory and space: towards an understanding of the cognitive map. *Journal of Neuroscience*, 35(41):13904–13911, 2015. 1

[28] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 1

[29] Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025. 1

[30] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1

[31] Yuxin Wang, Lei Ke, Boqiang Zhang, Tianyuan Qu, Hanxun Yu, Zhenpeng Huang, Meng Yu, Dan Xu, and Dong Yu. N3d-vlm: Native 3d grounding enables accurate spatial reasoning in vision-language models. *arXiv preprint arXiv:2512.16561*, 2025. 1

[32] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 1

[33] Mingrui Wu, Zhaozhi Wang, Fangjinhua Wang, Jiaolong Yang, Marc Pollefeys, and Tong Zhang. From indoor to open world: Revealing the spatial reasoning gap in mllms. *arXiv preprint arXiv:2512.19683*, 2025. 1

[34] 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*, pages 10632–10643, 2025. 1, 3, 4

[35] Weichen Zhang, Ruiying Peng, Chen Gao, Jianjie Fang, Xin Zeng, Kaiyuan Li, Ziyou Wang, Jinqiang Cui, Xin Wang, Xinlei Chen, et al. The point, the vision and the text: Does point cloud boost spatial reasoning of large language models? *arXiv preprint arXiv:2504.04540*, 2025. 1

[36] Zaibin Zhang, Yuhan Wu, Lianjie Jia, Yifan Wang, Zhongbo Zhang, Yijiang Li, Binghao Ran, Fuxi Zhang, Zhuohan Sun, Zhenfei Yin, et al. Think3d: Thinking with space for spatial reasoning. *arXiv preprint arXiv:2601.13029*, 2026. 1