SPACEMIND++: TOWARD ALLOCENTRIC COGNITIVE MAPS FOR SPATIALLY GROUNDED VIDEO MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent multimodal large language models (MLLMs) have made remarkable progress in visual understanding and language-based reasoning, yet they lack a persistent world-centered representation for spatially consistent reasoning in 3D environments. Inspired by the mammalian dual-stream system, where semantic and spatial cues are processed separately and integrated into an allocentric cognitive map, we propose *SpaceMind*++, a video MLLM architecture that explicitly builds a voxelized cognitive map from RGB videos. This map reorganizes fragmented egocentric observations into a shared 3D metric representation, enabling the model to preserve object permanence and spatial topology across changing viewpoints. To make this allocentric representation usable by a pretrained video MLLM without disrupting its native visual-token interface, we introduce Coordinate-Guided Deep Iterative Fusion, a new mechanism that relays map-level spatial knowledge back into the original 2D visual features. This fusion is explicitly guided by coordinate embeddings and 3D Rotary Positional Encoding, which ground semantic interactions in metric 3D space, resembling the entorhinal binding of sensory features to metric space. Extensive experiments show that SpaceMind++ achieves new **state-of-the-art** performance on VSI-Bench. Furthermore, it demonstrates strong **cross-benchmark generalization** on SPBench, SITE-Bench, and SPAR-Bench across diverse 3D environments.

1 INTRODUCTION

Spatial intelligence requires models to reason about the 3D structure of environments beyond isolated visual observations. Despite recent advances that have endowed Multimodal Large Language Models (MLLMs) with strong visual understanding and reasoning capabilitiesBai et al. (2025); Li et al. (2024); Lin et al. (2024); Zhu et al. (2025b), a fundamental gap persists between language-level reasoning and 3D spatial organizationWang et al. (2025c); Zhang et al. (2025). This disconnect limits the ability of MLLMs to ground their predictions in physical space, leading to unstable spatial reasoning, such as inconsistent distance estimation and object-level hallucinationsYang et al. (2024). This limitation stems from a lack of spatial constancy: visual observations are egocentric, partial, and transient, with viewpoint-dependent distances, occlusions, and fragmented frame-level evidence of the underlying 3D scene.

Recent efforts to build 3D-aware MLLMs attempt to address this issue by incorporating feed-forward geometry encoders alongside general-purpose visual encoders. These methods commonly fuse 2D visual features and 3D geometric cues through MLPsWu et al. (2025a), cross-attentionFan et al. (2026), or camera-guided integrationZhao et al. (2025). However, they largely treat geometry as an auxiliary modality aligned with egocentric visual tokens, enhancing spatial cues only locally without forming an explicit and persistent 3D structure. Rather, reliable spatial intelligence requires more than extracting geometric cues from individual observations; it requires a stable world-centered representation that preserves object permanence, metric consistency, and global scene layout across changing viewpoints.

Human spatial cognition offers a blueprint for spatially grounded reasoning. Neuroscience studies highlight two key mechanisms for spatial awareness: the dual visual stream Goodale and Milner (1992); Ungerleider and Mishkin (1982) and the cognitive map Hafting et al. (2005); O'Keefe and Nadel (1978); Tolman (1948). The former separates visual processing into ventral "what" and dorsal

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

"where" pathways, while the latter provides an allocentric representation for reasoning about spatial relationships beyond the current field of view. Although cognitive maps have inspired computational models such as the Tolman-Eichenbaum Machine Whittington et al. (2020; 2021; 2022) and visual-language navigation systems Gupta et al. (2017); Ruan et al. (2025), their use as an architectural principle for large-scale MLLMs remains under-explored.

Inspired by principles of human spatial cognition, we present **SpaceMind++**, a video MLLM that introduces an allocentric cognitive map as a first-class architectural component for spatial reasoning. In contrast to prior 3D-aware MLLMs that treat 3D geometry as an auxiliary feature stream, SpaceMind++ projects frame-level 2D semantic features into a voxelized map organized in 3D space. This transformation converts fragmented egocentric observations into a structured world-centered representation, allowing the model to maintain spatial topology across viewpoints.

To make this map useful for language reasoning, we further introduce **Coordinate-Guided Deep Iterative Fusion (CDIF)**, which relays spatial knowledge from the map back to the original visual tokens through iterative map self-attention and map-to-visual cross-attention. The LLM therefore receives geometry-enhanced visual tokens while preserving the pretrained visual-language interface. A key requirement for such fusion is that semantic interactions must be governed by metric 3D relationships rather than by arbitrary sequence order. Inspired by hippocampal–entorhinal spatial coding, which supports metric and allocentric representations of space Hafting et al. (2005); Rolls (1999); Rolls et al. (1997), we make attention coordinate-aware by injecting explicit 3D coordinate embeddings and applying 3D Continuous Rotary Position Embeddings. This design binds visual semantics to spatial locations during both attention steps, enabling SpaceMind++ to reason over object relations within a 3D metric structure. Experiments on VSI-Bench and other benchmarks demonstrate that SpaceMind++ achieves state-of-the-art performance across a wide range of vision-based spatial understanding and reasoning tasks.

In summary, our main contributions are:

- We introduce a brain-inspired design perspective for spatially grounded video MLLMs, drawing motivation from the mammalian dual visual stream and cognitive map to separate semantic recognition, spatial localization, and allocentric scene organization.
- We propose SpaceMind++, a novel video MLLM architecture that builds a voxelized allocentric cognitive map from RGB videos, effectively preserving spatial topology and object permanence across diverse viewpoints.
- We introduce Coordinate-Guided Deep Iterative Fusion (CDIF), a coordinate-aware mechanism that performs iterative map reasoning and map-to-visual reading with explicit 3D coordinate embeddings and 3D RoPE, while keeping the pretrained MLLM interface intact.
- We curate SpaceMind-900K for spatial instruction tuning and validate SpaceMind++ across multiple benchmarks, achieving new state-of-the-art performance on VSI-Bench and strong generalization on SPBench, SITE-Bench, and SPAR-Bench.

2 Related Work

2.1 Spatial Reasoning in Multimodal Large Language Models

Multimodal large language models (MLLMs) have advanced visual question answering, image captioning, and video understandingAlayrac et al. (2022); Bai et al. (2023); Chen et al. (2024b); Li et al. (2023); Lin et al. (2024); Radford et al. (2021); Zhu et al. (2025b), yet remain unreliable on metric estimation, spatial relationship judgment, and navigation-oriented reasoningStogiannidis et al. (2025); Yang et al. (2024); Yu et al. (2025). Spatial MLLMs broadly follow two directions. The first aligns language models with explicit 3D inputs, such as point cloudsFu et al. (2025); Hong et al. (2023); Mao et al. (2025) or depth mapsCai et al. (2025a); Cheng et al. (2024); Zhou et al. (2025), but often depends on reconstruction pipelines that limit scalability to unconstrained RGB videos. The second augments VLMs with geometry encoders or internal spatial representations, as in Spatial-MLLMWu et al. (2025a), VLM-3RFan et al. (2026), and SpaceMindZhao et al. (2025). These methods typically fuse geometry at the token level, strengthening local cues without explic-

itly reorganizing observations into a persistent world-centered structure. Our work instead uses 3D geometry to organize an allocentric cognitive map for spatial reasoning across changing viewpoints.

2.2 Feed-forward Visual Geometry for 3D Scene Understanding

Classical 3D reconstruction pipelines, such as Structure-from-Motion Schönberger and Frahm (2016) and Multi-View Stereo Schönberger et al. (2016); Wang et al. (2021); Yao et al. (2018), recover camera poses and dense geometry from images. However, their reliance on multi-stage optimization and sufficient view overlap makes them difficult to integrate into end-to-end multimodal reasoning systems. Recent feed-forward visual geometry models, including DUSt3RWang et al. (2024), MASt3RLeroy et al. (2024), CUT3RWang et al. (2025b), and VGGTWang et al. (2025a), alleviate this limitation by directly predicting dense correspondences, point maps, depth, and camera parameters from RGB images or short sequences. These models provide scalable geometric priors for in-the-wild inputs and have been increasingly adopted by geometry-aware MLLMsFan et al. (2026); Wu et al. (2025a). Nevertheless, their outputs are primarily designed for reconstruction rather than language-conditioned spatial reasoning, and existing MLLMs typically use them as auxiliary token-level features through projection layers or cross-attentionFan et al. (2026); Wu et al. (2025a). In contrast, our work converts feed-forward geometry into a voxelized cognitive map, where semantic features are aggregated in a persistent 3D metric space to support cross-view and allocentric spatial reasoning.

2.3 Brain-Inspired Spatial Reasoning for MLLMs

Brain-inspired spatial reasoning views cognitive maps as structured environmental representations that integrate local observations into coherent spatial memory Epstein et al. (2017); Kuipers (2000); Zhang et al. (2026). Existing MLLMs mainly use this idea as an auxiliary reasoning scaffold. Thinking-in-Space employs a lightweight grid-based cognitive map for spatial layout understanding Yang et al. (2024); SpaceR encourages 2D object-layout map imagination during reinforcement learning Ouyang et al. (2025); and Video2Layout replaces coarse grids with metric-grounded BEV layouts for finer spatial computation Huang et al. (2025). These methods demonstrate the utility of explicit layouts, although their representations remain largely planar and object-level. Complementary 3D approaches inject positional cues into visual patches (LLaVA-3D Zhu et al. (2025a)), maintain long-term spatiotemporal memory across extended trajectories (3DLLM-Mem Hu et al. (2025)), construct an object-centric Metric-CogMap for deterministic Cog-CoT reasoning (Map2Thought Gao et al. (2026)), or build compact 3D memory and replace visual tokens with map tokens (Cog3DMap Gwak et al. (2026)). In contrast, SpaceMind++ treats the map as an intermediate coordinate-grounded neural workspace: it performs spatial reasoning in the map and then returns the resulting knowledge to the original visual-token interface.

3 Method

3.1 Overall Architecture

The goal of SpaceMind++ is to empower Multimodal Large Language Models (MLLMs) with a 3D cognitive structure, enabling precise spatial reasoning in physical environments. Given a sequence of RGB video frames $\mathcal{S} = \{I_i\}_{i=1}^{N}$, $I_i \in \mathbb{R}^{3\times H\times W}$ and a textual prompt T, our model performs a multi-stage transformation that grounds language reasoning in a structured 3D metric space. SpaceMind++ follows the brain-inspired dual-stream design (left in Figure 1) and instantiates it as a three-stage video MLLM architecture (right in Figure 1): dual-stream feature extraction, allocentric cognitive map construction, and map-to-visual fusion.

Dual-stream feature extraction. Inspired by the biological dual-pathway process, we first decompose the input video into semantic and geometric streams.

Ventral stream A pretrained 2D visual encoder E_v (InternViT) extracts semantic tokens $f_v \in \mathbb{R}^{N\times M_v\times D_v}$ from the video frames, capturing *what* is present in the scene.

Dorsal stream Simultaneously, a geometry-aware spatial encoder E_s (e.g., VGGTWang et al. (2025a)) processes the same input to produce the geometric latent features $f_s \in \mathbb{R}^{N\times M_v\times D_s}$, per-

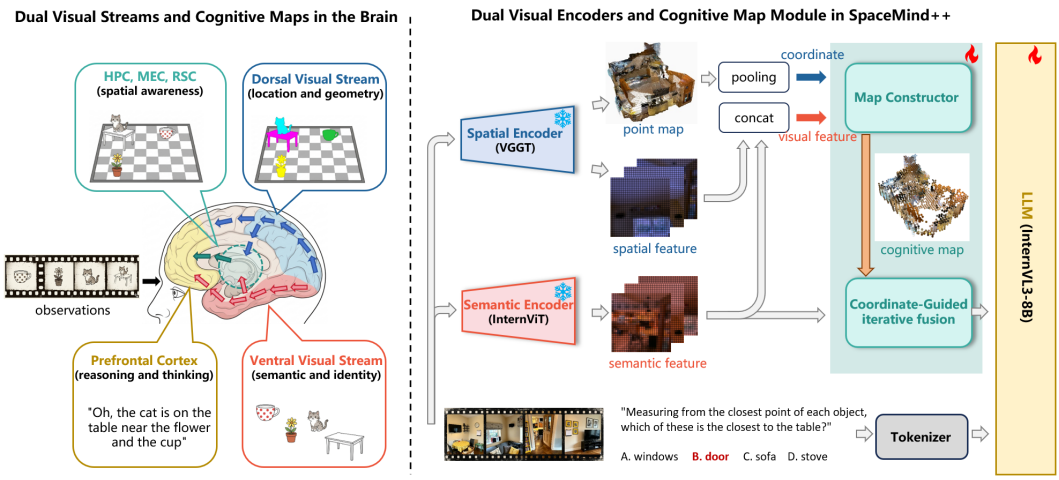

Figure 1: **Left**: biological motivation. Mammalian spatial cognition separates semantic identity and geometric localization through ventral (red area) and dorsal (blue area) visual streams, and integrates them into an allocentric cognitive map (cyan area) for spatial awareness and reasoning. **Right**: model architecture. SpaceMind++ extracts semantic and spatial features from video, organizes them into a voxelized allocentric cognitive map, and uses CDIF to relay map-level spatial knowledge back to visual tokens before language decoding (downward yellow arrow).

pixel 3D coordinates $P \in \mathbb{R}^{N\times H\times W\times 3}$ and confidence scores $C \in \mathbb{R}^{N\times H\times W\times 1}$, representing *where* objects are located.

$$f_v = E_v(\mathcal{S}), \quad \{f_s, P, C\} = E_s(\mathcal{S}). \tag{1}$$

Allocentric cognitive map fusion. The core of SpaceMind++ is the Cognitive Map module, which integrates the dual-stream features in an explicit 3D space. It scatters egocentric visual tokens into a unified voxel grid according to their 3D coordinates, producing an allocentric cognitive map $\mathcal{M}$:

$$\mathcal{M} = \text{Voxelize}(f_v, f_s, P), \quad f_{\text{fused}} = \mathcal{F}(f_v, \mathcal{M}, P). \tag{2}$$

To make the map accessible to the pretrained MLLM without changing its visual-token interface, $\mathcal{F}$ alternates between *spatial reasoning* over map tokens and *map reading* from map tokens back to visual tokens (the coordinate-guided iterative fusion module in Figure 1), progressively injecting 3D spatial knowledge into the original visual stream as f_{fused}.

Language decoding. The resulting geometry-enhanced visual tokens are fed into the LLM backbone G together with the text prompt T to generate a response R.

3.2 Allocentric Cognitive Map Construction

As illustrated in the left part of Figure 2, the map constructor converts patch-level visual features and 3D coordinates into a compact voxelized cognitive map through coordinate quantization, topology-preserving aggregation, and stochastic sampling.

3.2.1 Dynamic Coordinate Quantization

We first align the dense geometric outputs of the spatial encoder with the patch-level visual tokens. Specifically, the dense point map and confidence map are average-pooled to the visual-token resolution, producing patch-aligned coordinates $P \in \mathbb{R}^{L\times 3}$ and confidence scores $C \in \mathbb{R}^{L\times 1}$, where $L = N \times M_v$. To reduce the effect of noisy geometry, we retain only high-confidence coordinates and compute a dynamic scene center $\mathbf{c} \in \mathbb{R}^3$ as their centroid:

$$\mathbf{c} = \frac{1}{|\mathcal{V}_{\text{valid}}|} \sum_{i\in\mathcal{V}_{\text{valid}}} P_i, \quad \text{where} \quad \mathcal{V}_{\text{valid}} = \{i \mid C_i > \tau_{\text{conf}}\}. \tag{3}$$

Here, τ_{conf} is the confidence threshold. We then recenter each coordinate by the dynamic scene center, $\tilde{P}_i = P_i - \mathbf{c}$, and discretize the recentered 3D space into a voxel grid with extent D (we set $D = 100$ for all experiments) and scale-free resolution r. Since the spatial encoder predicts geometry in a relative metric space, r is defined on this normalized scale rather than in absolute physical units. The continuous coordinates are quantized into integer voxel indices $\mathbf{u}_i = (u_{x,i}, u_{y,i}, u_{z,i}) \in \mathbb{Z}^3$:

$$\mathbf{u}_i = \left\lfloor \frac{\tilde{P}_i}{r} \right\rfloor + \frac{D}{2}. \tag{4}$$

This step corresponds to the feature voxelization panel in Figure 2, where continuous 3D coordinates are mapped into discrete voxel indices so that tokens from nearby locations can be grouped together. We then convert each 3D voxel index into a unique 1D spatial hash key h_i for efficient grouping:

$$h_i = u_{x,i}D^2 + u_{y,i}D + u_{z,i}. \tag{5}$$

Because the linearized spatial index assumes a predefined grid boundary, tokens whose computed 3D indices fall outside the grid extent $[0, D-1]^3$ are discarded.

3.2.2 Topology-Preserving Feature Aggregation

In video streams, the same region may be observed repeatedly, causing substantial token redundancy. We therefore use voxelization as a geometry-aware bottleneck and group tokens by their 1D spatial hash keys h. Let $\mathcal{H}_j = \{i \mid h_i = j\}$ denote the set of tokens assigned to the j-th voxel.

As shown in Figure 2, tokens within each voxel are aggregated into one compact map token after filtering inconsistent observations. We first compute a mean feature prototype $\boldsymbol{\mu}_j$ by averaging the concatenated features $x_i = [f_{v,i}; f_{s,i}]$ in the bin, which represents the consensus semantic-geometric state of that region. We then remove outlier observations whose cosine similarity to $\boldsymbol{\mu}_j$ is below τ_{sim}:

$$\tilde{\mathcal{H}}_j = \left\{ i \in \mathcal{H}_j \;\middle|\; \frac{x_i \cdot \boldsymbol{\mu}_j}{\|x_i\| \, \|\boldsymbol{\mu}_j\|} > \tau_{\text{sim}} \right\}, \quad \boldsymbol{\mu}_j = \frac{1}{|\mathcal{H}_j|} \sum_{i \in \mathcal{H}_j} x_i. \tag{6}$$

The final representation for the j-th voxel is obtained by applying scatter-mean pooling over the refined set $\tilde{\mathcal{H}}_j$:

$$\mathbf{v}_j = \frac{1}{M_j} \sum_{i \in \tilde{\mathcal{H}}_j} x_i, \quad \mathbf{p}_j = \frac{1}{M_j} \sum_{i \in \tilde{\mathcal{H}}_j} \tilde{P}_i, \quad t_j = \min_{i \in \tilde{\mathcal{H}}_j} t_i, \quad M_j = |\tilde{\mathcal{H}}_j|. \tag{7}$$

Here, $\mathbf{v}_j$, $\mathbf{p}_j$, and t_j denote the voxel feature, spatial coordinate, and first-observed timestamp of the j-th voxel, respectively. By retaining the minimum timestamp among all assigned tokens, each voxel records when its corresponding physical region first appears in the video, providing a lightweight cue for appearance-order reasoning.

3.2.3 Global Stochastic Sampling for Scale Invariance

For large scenes, voxelization can yield dense maps. When the number of voxels M exceeds $M_{\max}$, we uniformly sample $M_{\max}$ valid voxels instead of truncating the sequence, reducing redundancy while preserving the global spatial layout. We set $M_{\max} = 4000$ in all experiments.

3.3 Coordinate-Guided Deep Iterative Fusion

After constructing the voxelized cognitive map, SpaceMind++ uses Coordinate-Guided Deep Iterative Fusion (CDIF) to transfer map-level spatial knowledge back to visual tokens while preserving the original visual-token interface. As shown in Figure 2, CDIF consists of L layers that alternate between Map Reasoning and Map Reading.

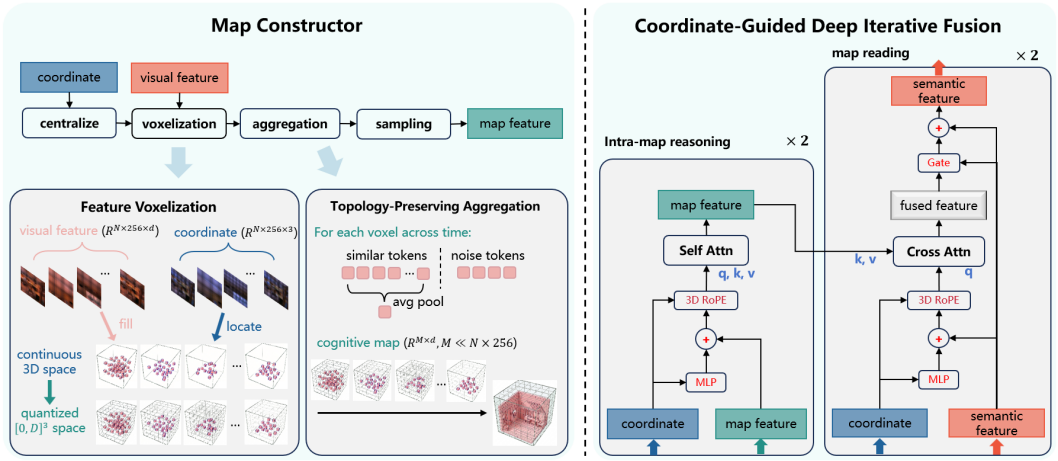

Figure 2: Detailed components of SpaceMind++. The map constructor transforms patch-level visual features and 3D coordinates into a voxelized cognitive map. CDIF then alternates between intra-map reasoning and map reading to inject spatial knowledge back into semantic tokens. Both stages are guided by coordinate embeddings and 3D RoPE to preserve metric 3D relationships.

3.3.1 Map Reasoning via Self-Attention

Within each fusion layer l, CDIF first performs intra-map self-attention to propagate semantic and geometric information across voxels. We encode each voxel coordinate $\mathbf{p}_j$ with an MLP and add it to the voxel feature $\mathbf{v}_j^{(l)}$, then apply 3D Continuous Rotary Positional Embedding (3D RoPE) Su et al. (2021) to make the queries and keys coordinate-aware. This enables attention to model relative spatial relationships in metric 3D space:

$$\tilde{\mathbf{v}}_j^{(l)} = \mathbf{v}_j^{(l)} + \mathrm{MLP}(\mathbf{p}_j), \quad \hat{\mathbf{q}}_j^{(l)}, \hat{\mathbf{k}}_j^{(l)} = \mathrm{RoPE}(\mathrm{Linear}(\tilde{\mathbf{v}}_j^{(l)}), \mathbf{p}_j). \tag{8}$$

The voxel map is then updated via self-attention:

$$\mathcal{V}^{(l+1)} = \mathrm{SelfAttn}(\hat{\mathbf{Q}}, \hat{\mathbf{K}}, \mathbf{V}(\mathcal{V}^{(l)})). \tag{9}$$

This step enables the cognitive map to perform global geometric inference, such as modeling object-to-object spatial relationships, within a unified 3D coordinate frame.

3.3.2 Map Reading via Semantic-Aware Gated Attention

After the cognitive map is refined, each 2D visual token reads spatial context from the map. For a visual token $\mathbf{f}_{v,i}^{(l)}$ with 3D coordinate $\mathbf{p}_i$, we first inject its coordinate information through an MLP-based embedding. The coordinate-enhanced token is then projected into query and key modulated by 3D RoPE, making the map-reading attention aware of its metric location:

$$\tilde{\mathbf{f}}_{v,i}^{(l)} = \mathbf{f}_{v,i}^{(l)} + \mathrm{MLP}(\mathbf{p}_i), \quad \mathbf{q}_i^{(l)} = \mathrm{RoPE}(\mathrm{Linear}(\tilde{\mathbf{f}}_{v,i}^{(l)}), \mathbf{p}_i). \tag{10}$$

The model then performs cross-attention over the refined voxel map $\mathcal{V}^{(l+1)}$ to retrieve the fused spatial feature $\mathbf{f}_{\mathrm{fused},i}^{(l)}$:

$$\mathbf{f}_{\mathrm{fused},i}^{(l)} = \mathrm{CrossAttn}(\mathbf{q}_i^{(l)}, \mathbf{K}(\mathcal{V}^{(l+1)}), \mathbf{V}(\mathcal{V}^{(l+1)})). \tag{11}$$

Gated information integration. To prevent the visual stream from absorbing irrelevant geometric background, we use its semantic content to compute a gate that modulates the retrieved spatial feature before the residual update:

$$\mathbf{g}_i^{(l)} = \sigma\left(\mathrm{MLP}(\mathbf{f}_{v,i}^{(l)})\right), \qquad \mathbf{f}_{v,i}^{(l+1)} = \mathbf{f}_{v,i}^{(l)} + \mathbf{g}_i^{(l)} \odot \mathrm{FFN}\left(\mathbf{f}_{\mathrm{fused},i}^{(l)}\right). \tag{12}$$

This design ensures that the 2D visual stream remains the primary carrier of semantic identity while selectively integrating 3D spatial knowledge only when the visual content, such as a specific object or landmark, requires geometric grounding for downstream reasoning.

Table 1: VSI-Bench evaluation. Best and second-best results within each model group are shown in bold and underlined, respectively.

Methods	Avg.	Numerical Question				Multiple-Choice Question			
		Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order
Proprietary Models (API)									
GPT-5 Singh et al. (2026)	55.0	**53.3**	34.4	**73.3**	47.5	**63.7**	48.6	50.2	**68.9**
Gemini-3 Pro Google DeepMind (2025)	**56.0**	49.0	**42.8**	71.5	41.8	56.6	**57.5**	**61.9**	60.0
Grok-4 xAI (2025)	47.9	37.1	32.9	60.8	45.4	53.1	39.6	47.4	66.8
Open-source VLMs									
InternVL3-78B Zhu et al. (2025b)	48.5	**71.2**	**53.7**	44.4	39.5	55.9	39.5	28.9	54.5
LLaVA-NeXT-Video-72B Liu et al. (2024)	40.9	48.9	22.8	57.4	35.3	42.4	36.7	**35.0**	48.6
Qwen3VL-8B-Instruct Yang et al. (2025a)	**57.9**	67.5	47.0	**76.3**	**61.9**	**58.0**	**50.9**	**35.0**	**66.3**
LLaVA-OneVision-72B Li et al. (2024)	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
Specialized Spatial Reasoning Models									
Spatial-MLLM Wu et al. (2025a)	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VST-7B Yang et al. (2025b)	60.6	72.0	44.4	74.3	68.3	59.7	55.8	44.9	65.2
VLM-3R Fan et al. (2026)	60.9	70.2	49.4	69.2	67.1	65.4	80.5	45.4	40.1
SenseNova-SI-1.2 Cai et al. (2025b)	69.6	72.7	56.0	77.1	75.7	70.4	81.7	42.8	79.9
SpaceMind Zhao et al. (2025)	70.2	73.9	61.5	77.6	74.8	67.7	88.6	46.9	70.7
SpaceMind++ (Ours)	**73.2**	**74.5**	**62.4**	**77.9**	**76.9**	**73.5**	**89.7**	**48.2**	**84.1**

4 Implementation Details

Training data. We train SpaceMind++ on a spatial instruction-tuning corpus containing approximately 900K QA samples. The corpus integrates four spatial-reasoning data sources: ViCA-322KFeng (2025), VLM-3R-dataFan et al. (2026), SQA3D-trainMa et al. (2023), and VSI-590K-VideoYang et al. (2025c). As shown in Figure 3, the dataset covers a broad spectrum of spatial reasoning skills, including object counting, relative and absolute distance estimation, size estimation, object ordering, situated 3D perception, complex spatial reasoning, and route planning. This mixture encourages the model to learn both local metric grounding and global spatial consistency, which are essential for reasoning on an allocentric cognitive map.

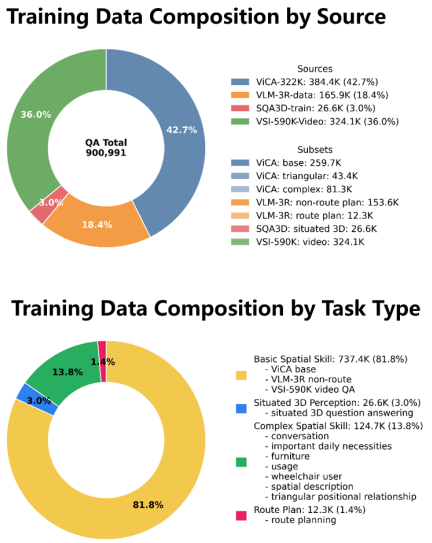

Figure 3: Statistics of the SpaceMind-900K dataset

Training setup. We use InternVL3-8B Zhu et al. (2025b) as the base MLLM and keep both the ViT backbone and VGGT encoder frozen during training. We train the CDIF layers and the associated cross-modal projectors, while applying LoRA Hu et al. (2021) to the language model with rank $r = 256$ and scaling factor $\alpha = 512$. The trainable parameters are optimized with AdamW Loshchilov and Hutter (2019) using a cosine annealing schedule and a peak learning rate of 2×10^{-5}. Training is conducted on 64 NVIDIA H200 GPUs for approximately 41 hours.

5 Evaluation

We evaluate SpaceMind++ on four spatial reasoning benchmarks. VSI-Bench is used as the primary benchmark for video-based spatial intelligence, while SPBench, SITE-Bench, and SPAR-Bench are used to assess generalization across multi-view and video spatial reasoning settings.

5.1 VSI-Bench

As shown in Table 1, SpaceMind++ achieves a new state-of-the-art average score of 73.2, improving upon the strong SpaceMind baseline on VSI-Bench. Although SpaceMind and SenseNova-SI are

Table 2: Generalization evaluation on SPBench and SITE-Bench. Results cover the single-image and multi-view settings of SPBench and the video-only setting of SITE-Bench.

Methods	SPBench-SI			SPBench-MV			SPBench	SITE-Bench
	NQ	MCQ	Avg.	NQ	MCQ	Avg.		
Proprietary Models								
GPT-4o OpenAI et al. (2024)	24.5	60.3	42.4	40.7	59.4	50.1	46.2	–
Gemini-2.0-Flash Comanici et al. (2025)	49.0	60.4	54.7	51.9	50.7	51.3	53.0	–
Open-Source Models								
InternVL-2.5-8B Chen et al. (2024a)	28.3	56.3	42.3	37.3	47.5	42.4	42.3	–
Kimi-VL-A3B Moonshot AI et al. (2025)	25.7	44.9	35.3	23.3	57.6	40.5	37.9	–
LLaVA-OneVision-7B Li et al. (2024)	25.4	41.0	33.2	20.6	49.6	35.1	34.2	–
Qwen2.5-VL-7B Bai et al. (2025)	36.3	60.5	48.4	28.9	49.8	39.3	43.9	53.7
Spatial Reasoning Models								
Video-R1 Feng et al. (2025)	27.7	62.0	44.9	32.5	53.0	42.8	43.8	–
SpaceR-7B Ouyang et al. (2025)	35.7	61.5	48.6	63.2	53.7	58.5	53.5	56.5
VILASR-7B Wu et al. (2025b)	36.6	63.7	50.2	56.2	59.6	57.9	54.0	56.1
Spatial-MLLM-4B Wu et al. (2025a)	38.1	49.3	43.7	63.7	58.9	61.3	52.5	44.0
SpaceMind Zhao et al. (2025)	**66.3**	53.2	59.7	76.2	70.5	73.8	67.3	–
EgoMind Chen et al. (2026)	–	–	–	–	–	–	55.0	58.0
SpaceMind++ (Ours)	56.7	**65.6**	**61.1**	**76.5**	**82.3**	**78.9**	**70.0**	**61.0**

Table 3: Generalization evaluation on SPAR-Bench high-level tasks. Results are reported for distance (Dist), object-relation (ObjRel), and spatial-imagination (SpImag) tasks, including multi-view (MV) settings.

Methods	High Avg.	Dist OO	Dist OO-MV	ObjRel OC-MV	ObjRel OO	ObjRel OO-MV	SpImag OC	SpImag OC-MV	SpImag OO	SpImag OO-MV
Closed-source Models										
GPT-4o OpenAI et al. (2024)	43.80	65.00	64.88	44.75	50.82	43.21	29.84	32.56	27.81	35.29
GPT-4.1 OpenAI et al. (2023)	42.93	71.76	67.26	46.25	54.95	41.00	30.38	29.65	20.20	24.93
Doubao-1.5-vision-pro ByteDance Seed et al. (2025)	49.49	**74.71**	**69.64**	35.75	**70.33**	47.65	34.95	33.14	**35.76**	42.86
Open-source Models										
InternVL2.5-38B Chen et al. (2024a)	44.13	69.12	66.67	43.75	64.29	37.67	25.27	31.98	31.79	26.61
Qwen2.5-VL-72B Bai et al. (2025)	43.80	58.82	61.90	40.75	53.57	45.98	26.88	35.17	34.11	36.97
LLaVA-v1.6-7B Liu et al. (2024)	20.18	51.76	7.74	6.25	32.14	6.37	39.52	10.47	21.52	5.88
SpaceR-7B Ouyang et al. (2025)	45.61	62.35	61.61	52.25	51.92	46.81	37.90	36.05	24.83	34.17
SpaceMind++ (Ours)	**51.28**	39.78	33.00	**70.00**	48.63	**53.46**	**51.61**	**53.78**	27.15	**45.66**

already highly competitive spatial reasoning models, SpaceMind++ further improves several key subcategories that reflect the benefits of allocentric cognitive map construction.

Specifically, SpaceMind++ improves relative distance estimation from 67.7 to 73.5 and appearance ordering from 70.7 to 84.1. Relative distance estimation requires metric grounding between objects in 3D space, while appearance ordering further requires preserving temporal object events and spatial relationships across changing viewpoints. We also observe competitive performance on object count, room size, and relative direction estimation. Overall, the performance suggests that SpaceMind++ is most effective when spatial reasoning requires persistent object-level memory and metric consistency across viewpoints.

5.2 Generalization to Spatial Reasoning Benchmarks

SPBench and SITE-Bench. SPBench evaluates spatial reasoning under both single-image and multi-view (eight images) settingsLi et al. (2025). SpaceMind++ remains competitive on SPBench-SI and reaches 78.9 on SPBench-MV, producing a 70.0 overall score (Table 2). Its larger multi-view gain is consistent with the cognitive map aggregating cross-view evidence and preserving object relationships. SITE-Bench evaluates multiple-choice spatial VQA across single-image, multi-image, and video inputsWang et al. (2025c). Because its image subset contains only 1–5 frames, we report the video-only EgoMind-style average as an auxiliary reference and do not claim full SITE-Bench superiority; SpaceMind++ nevertheless remains competitive on this setting.

SPAR-Bench. SPAR-Bench provides a difficulty-stratified evaluation of spatial understandingZhang et al. (2025). We focus on the high-level split, which emphasizes multi-object relational reasoning and spatial imagination across views; the omitted low- and medium-level splits are domi-

nated by direct depth, distance, and view-alignment perception. As shown in Table 3, SpaceMind++ achieves the best average among open-source models, with clear advantages on multi-view object-relation and spatial-imagination tasks. These results support the value of a stable allocentric map for cross-view object association and hypothetical-view reasoning.

5.3 Ablation Studies

We conduct controlled ablations along three axes: data scaling versus architectural improvements, the contributions of cognitive-map construction and coordinate-guided fusion, and sensitivity to the voxel-token budget and spatial resolution. All component and map-parameter variants use the same 900K corpus to isolate each factor.

To disentangle data scaling from architectural improvements, Table 4 compares SpaceMind and SpaceMind++ under matched training sets. Increasing the corpus from 626K to 900K improves SpaceMind from 69.6 to 72.1 (+2.5) and SpaceMind++ from 71.2 to 73.2 (+2.0). Under matched data, SpaceMind++ outperforms SpaceMind by 1.6 points on 626K and 1.1 points on 900K.

Table 4: Controlled ablations on VSI-Bench. Component and map variants use 900K. DIF omits 3D positional guidance.

Data–architecture control Model	626K	900K	Data gain
SpaceMind	69.6	72.1	+2.5
SpaceMind++	71.2	**73.2**	+2.0
Architecture gain	+1.6	+1.1	–

Component ablations Model variant	Training data	Score
InternVL3-8B baseline	900K	70.8
SpaceMind baseline	900K	72.1
VGGT + multilayer cross-attention	900K	71.3
VGGT + random map + DIF	900K	70.6
VGGT + cognitive map + DIF	900K	71.7
VGGT + cognitive map + CDIF (Ours)	900K	**73.2**

Cognitive-map parameters $M_{\max}$	Voxel size	Training data	Score
1000	0.05	900K	73.0
4000	0.10	900K	72.4
4000	0.05	900K	**73.2**

We then isolate the model components using the same 900K corpus. The InternVL3-8B baseline obtains 70.8, while directly fusing VGGT features through multilayer cross-attention reaches 71.3, a 0.5-point improvement over the InternVL3-8B baseline. This result shows that access to raw geometric features alone is insufficient. Random-map context with Deep Iterative Fusion (DIF), the coordinate-agnostic counterpart of CDIF that omits 3D positional guidance, obtains 70.6, indicating that additional context tokens and attention alone do not explain the gain. Replacing the random context with the voxelized cognitive map improves DIF to 71.7, and restoring coordinate-guided CDIF further raises performance to 73.2. These respective gains support both spatial organization through the cognitive map and coordinate-guided transfer of map information back into the visual stream.

Finally, reducing $M_{\max}$ from 4000 to 1000 at a voxel size of 0.05 changes performance only from 73.2 to 73.0, demonstrating robustness to a substantial reduction in voxel tokens. In contrast, increasing the voxel size from 0.05 to 0.10 at $M_{\max} = 4000$ lowers the score to 72.4. Coarser voxelization merges more spatially and semantically heterogeneous observations.

6 Conclusion

SpaceMind++ achieves state-of-the-art performance on VSI-Bench and strong cross-benchmark transfer, particularly on tasks requiring object memory, metric consistency, and cross-view integration. By organizing geometry as a persistent voxelized allocentric map and relaying its spatial knowledge through CDIF, SpaceMind++ provides a promising direction for spatially grounded video MLLMs.

Limitations and Future Work. SpaceMind++ depends on the feed-forward spatial encoder, so inaccurate geometry or sparse observations can weaken its cognitive map. Map reasoning and map-to-visual fusion also add computation, while a fixed voxel budget may omit fine-grained details. Future work will explore uncertainty-aware mapping, adaptive or hierarchical voxel allocation, dynamic scenes, embodied agents, and long-horizon spatial memory.

AI USE STATEMENT

Large language models were used to assist with programming and figure preparation, as well as grammar and expression polishing. They did not contribute to research ideation, methodology design, or experiments.

ETHICS STATEMENT

This work develops a video MLLM for spatial reasoning and does not involve human-subject experiments. Training and evaluation use existing research datasets, and we do not collect or release new personally identifiable information. The model may inherit dataset biases and can produce inaccurate spatial estimates when the underlying geometry is unreliable. Spatial understanding is also dual-use and may enable privacy-invasive video analysis or unauthorized environment mapping. SpaceMind++ should therefore not be deployed in safety-critical settings without task-specific validation and human oversight, and its computational and environmental costs should also be considered.

REPRODUCIBILITY STATEMENT

We have made effort to ensure the reproducibility of our work. The proposed model, implementation details are described in detail in the main paper. The source code and data will be made publicly available.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and others. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A frontier large vision-language model with versatile abilities. *arXiv*, abs/2308.12966, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, and others. Qwen2.5-VL technical report. *arXiv*, abs/2502.13923, 2025.

ByteDance Seed et al. Seed1.5-VL technical report. *arXiv*, abs/2505.07062, 2025.

Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. SpatialBot: Precise spatial understanding with vision language models. In *IEEE International Conference on Robotics and Automation*, 2025a.

Zhongang Cai, Ruisi Wang, Chenyang Gu, Fanyi Pu, Junxiang Xu, Yubo Wang, Wanqi Yin, Zhitao Yang, Chen Wei, Qingping Sun, Tongxi Zhou, Jiaqi Li, Hui En Pang, Oscar Qian, Yukun Wei, Zhiqian Lin, Xuanke Shi, Kewang Deng, Xiaoyang Han, Zukai Chen, Xiangyu Fan, Hanming Deng, Lewei Lu, Liang Pan, Bo Li, Ziwei Liu, Quan Wang, Dahua Lin, and Lei Yang. Scaling spatial intelligence with multimodal foundation models. *arXiv*, abs/2511.13719, 2025b.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv*, abs/2412.05271, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Zhenghao Chen, Huiqun Wang, and Di Huang. EgoMind: Activating spatial cognition through linguistic reasoning in MLLMs. *arXiv*, abs/2604.03318, 2026.

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. In *Advances in Neural Information Processing Systems*, 2024.

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

Russell A. Epstein, Eva Zita Patai, Joshua B. Julian, and Hugo J. Spiers. The cognitive map in humans: Spatial navigation and beyond. *Nature Neuroscience*, 20(11):1504–1513, 2017. doi: 10.1038/nn.4656.

Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, Hongyu Xu, Justin Theiss, Tianlong Chen, Jiachen Li, Zhengzhong Tu, Zhangyang Wang, and Rakesh Ranjan. VLM-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. *arXiv*, abs/2503.21776, 2025.

Qing Feng. Towards visuospatial cognition via hierarchical fusion of semantic and spatial representations. *arXiv*, abs/2505.12363, 2025.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-LLM: Extending language model for 3d visual understanding and reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2025.

Xiangjun Gao, Zhensong Zhang, Dave Zhenyu Chen, Songcen Xu, Long Quan, Eduardo Pérez-Pellitero, and Youngkyoon Jang. Map2thought: Explicit 3d spatial reasoning via metric cognitive maps. *arXiv*, abs/2601.11442, 2026.

Melvyn A. Goodale and A. David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, 1992. doi: 10.1016/0166-2236(92)90344-8.

Google DeepMind. Gemini 3 pro: The frontier of vision AI, 2025. Accessed: 2026-03-21.

Saurabh Gupta, Varun Tolani, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2017.

Chanyoung Gwak, Yoonwoo Jeong, Byungwoo Jeon, Hyunseok Lee, Jinwoo Shin, and Minsu Cho. Cog3dmap: Multi-view vision-language reasoning with 3d cognitive maps. *arXiv*, abs/2603.23023, 2026.

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. doi: 10.1038/nature03721.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-LLM: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems*, 2023.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021.

Wenbo Hu, Yining Hong, Yanjun Wang, Leison Gao, Zibu Wei, Xingcheng Yao, Nanyun Peng, Yonatan Bitton, Idan Szpektor, and Kai-Wei Chang. 3dllm-mem: Long-term spatial-temporal memory for embodied 3d large language model. *arXiv*, abs/2505.22657, 2025.

Yibin Huang, Wang Xu, Wanyue Zhang, Helu Zhi, Jingjing Huang, Yangbin Xu, Yangang Sun, Conghui Zhu, and Tiejun Zhao. Video2layout: Recall and reconstruct metric-grounded cognitive map for spatial reasoning. *arXiv*, abs/2511.16160, 2025.

Benjamin Kuipers. The spatial semantic hierarchy. *Artificial Intelligence*, 119(1–2):191–233, 2000.

Vincent Leroy, Yohann Cabon, and Jerome Revaud. MASt3r: Grounding image matching in 3d. *arXiv*, abs/2406.09756, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Li, Yanwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv*, abs/2408.03326, 2024.

Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. SpatialLadder: Progressive training for spatial reasoning in vision-language models. *arXiv*, abs/2510.08531, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, volume 202, pages 19730–19742, 2023.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv*, abs/2311.10122, 2024.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. `https://llava-vl.github.io/blog/2024-01-30-llava-next/`, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. SQA3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023.

Yongsen Mao, Junhao Zhong, Chuan Fang, Jia Zheng, Rui Tang, Hao Zhu, Ping Tan, and Zihan Zhou. SpatialLM: Training large language models for structured indoor modeling. *arXiv*, abs/2506.07491, 2025.

Moonshot AI et al. Kimi-VL technical report. *arXiv*, abs/2504.07491, 2025.

John O'Keefe and Lynn Nadel. *The Hippocampus as a Cognitive Map*. Clarendon Press, Oxford, 1978.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, et al. GPT-4 technical report. *arXiv*, abs/2303.08774, 2023.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, et al. GPT-4o system card. *arXiv*, abs/2410.21276, 2024.

Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. SpaceR: Reinforcing MLLMs in video spatial reasoning. *arXiv*, abs/2504.01805, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

Edmund T. Rolls. Spatial view cells and the representation of place in the primate hippocampus. *Hippocampus*, 9(4):467–480, 1999.

Edmund T. Rolls, Richard G. Robertson, and Philippe Georges-Francois. Spatial view cells in the primate hippocampus. *European Journal of Neuroscience*, 9(8):1789–1794, 1997. doi: 10.1111/j.1460-9568.1997.tb01538.x.

Shouwei Ruan, Liyuan Wang, Caixin Kang, Qihui Zhu, Songming Liu, Xingxing Wei, and Hang Su. From reactive to cognitive: Brain-inspired spatial intelligence for embodied agents. *arXiv*, 2025.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

Johannes L. Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 501–518, 2016.

Aman Singh et al. OpenAI GPT-5 system card. *arXiv*, abs/2601.03267, 2026.

Ilias Stogiannidis, Steven McDonagh, and Sotirios A. Tsaftaris. Mind the gap: Benchmarking spatial reasoning in vision-language models. *arXiv*, abs/2503.19707, 2025.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *arXiv*, abs/2104.09864, 2021. doi: 10.48550/arXiv.2104.09864.

Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948. doi: 10.1037/h0061626.

Leslie G. Ungerleider and Mortimer Mishkin. Two cortical visual systems. In David J. Ingle, Melvyn A. Goodale, and Richard J. W. Mansfield, editors, *Analysis of Visual Behavior*, pages 549–586. MIT Press, Cambridge, MA, 1982.

Fangjinhua Wang, Silvano Galliani, Christoph Vogel, Pablo Speciale, and Marc Pollefeys. PatchmatchNet: Learned multi-view patchmatch stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14194–14203, 2021.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotný. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5294–5306, 2025a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. CUT3r: Continuous 3d perception model with persistent state. *arXiv*, abs/2501.12387, 2025b.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. SITE: Towards spatial intelligence thorough evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025c.

James C. R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E. J. Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020. doi: 10.1016/j.cell.2020.10.024.

James C. R. Whittington, Joseph Warren, and Timothy E. J. Behrens. Relating transformers to models and neural representations of the hippocampal formation. *arXiv*, abs/2112.04035, 2021.

James C. R. Whittington, David McCaffary, Jacob J. W. Bakermans, and Timothy E. J. Behrens. How to build a cognitive map: Insights from models of the hippocampal formation. *Nature Neuroscience*, 25(10):1257–1272, 2022. doi: 10.1038/s41593-022-01153-y.

Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-MLLM: Boosting MLLM capabilities in visual-based spatial intelligence. In *Advances in Neural Information Processing Systems*, 2025a.

Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv*, abs/2506.09965, 2025b.

xAI. Grok 4 model card. `https://data.x.ai/2025-08-20-grok-4-model-card.pdf`, 2025. Accessed: 2026-05-06.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, and others. Qwen3 technical report. *arXiv*, abs/2505.09388, 2025a.

Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. *arXiv*, abs/2412.14171, 2024.

Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv*, abs/2511.05491, 2025b.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, Daohan Lu, Rob Fergus, Yann LeCun, Li Fei-Fei, and Saining Xie. Cambrian-s: Towards spatial supersensing in video. *arXiv*, abs/2511.04670, 2025c.

Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision*, pages 767–783, 2018.

Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, Zhedong Zheng, Zhipeng Zhang, Yifan Wang, Lin Song, Lijun Wang, Yanwei Li, Ying Shan, and Huchuan Lu. How far are VLMs from visual spatial intelligence? a benchmark-driven perspective. *arXiv*, abs/2509.18905, 2025.

Jiahui Zhang, Yurui Chen, Yanpeng Zhou, Yueming Xu, Ze Huang, Jilin Mei, Junhui Chen, Yu-Jie Yuan, Xinyue Cai, Guowei Huang, Xingyue Quan, Hang Xu, and Li Zhang. From flatland to space: Teaching vision-language models to perceive and reason in 3d. *arXiv*, abs/2503.22976, 2025.

Pingyue Zhang, Zihan Huang, Yue Wang, Jieyu Zhang, Letian Xue, Zihan Wang, Qineng Wang, Keshigeyan Chandrasegaran, Ruohan Zhang, Yejin Choi, Ranjay Krishna, Jiajun Wu, Li Fei-Fei, and Manling Li. Theory of space: Can foundation models construct spatial beliefs through active exploration? *arXiv*, abs/2602.07055, 2026.

Ruosen Zhao, Zhikang Zhang, Jialei Xu, Jiahao Chang, Dong Chen, Lingyun Li, Weijian Sun, and Zizhuang Wei. SpaceMind: Camera-guided modality fusion for spatial reasoning in vision-language models. *arXiv*, abs/2511.23075, 2025.

Enshen Zhou, Jingkun An, Cheng Chi, Yi Han, Shanyu Rong, Chi Zhang, Pengwei Wang, Zhongyuan Wang, Tiejun Huang, Lu Sheng, and Shanghang Zhang. RoboRefer: Towards spatial referring with reasoning in vision-language models for robotics. *arXiv*, abs/2506.04308, 2025.

Chenming Zhu, Tai Wang, Wenwei Zhang, Jiangmiao Pang, and Xihui Liu. LLaVA-3d: A simple yet effective pathway to empowering LMMs with 3d awareness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv*, abs/2504.10479, 2025b.