VoxAnchor: Explicit Voxel-Semantic Grounding for Spatial Understanding in Videos

Xinglin Li[1*], TingTing Long[2*], Jingzhi Zhou[1], Chuxuan Zeng[3], Jiajing Chen[4], Jian Yang[1], and Jin Xie[5†]

[1] College of Computer Science, Nankai University, Tianjin, China
[2] Computer Center, Peking University, Beijing, China
[3] China Unicom Greater Bay Area Innovation Institute, Guangzhou, China
[4] Amazon, USA
[5] School of Intelligence Science and Technology, Nanjing University, China
xinglinli@mail.nankai.edu.cn, l.tingting@pku.edu.cn,
jingzhi.zhou@mail.nankai.edu.cn, zengchuxuan@chinaunicom.cn
cjiajing@amazon.com, csjyang@nankai.edu.cn, csjxie@nju.edu.cn

Abstract. Although Multimodal Large Language Models (MLLMs) have made impressive progress in understanding videos—such as recognizing objects, actions, and events—they still struggle with spatial reasoning. In particular, they have difficulty forming a consistent 3D understanding of a scene from separate 2D video frames. Unlike humans, who can naturally infer depth, distance, object size, and their own movement through space, these models often lack the ability to accurately reconstruct a coherent 3D environment. This limitation makes it challenging for them to estimate real-world measurements, such as how far an object is or how much the camera has moved, when relying only on 2D image sequences. To bridge these gaps, we propose VoxAnchor, a framework for explicit spatial-semantic grounding. Specifically, we reify the 2D video sequence into a spatiotemporally continuous representation by unprojecting visual tokens, geometric features, and view-consistent semantics into a unified 3D voxel grid. To facilitate precise spatial reasoning, we propose a Voxel-Aware Attention mechanism, which constrains feature interactions within physically consistent volumetric units. This approach enables the model to effectively aggregate fragmented observations and infer metric-aware spatial relationships from monocular video. VoxAnchor sets a new state-of-the-art on VSI (66.8%) and VSTI (66.3%) benchmarks, outperforming significantly more parameter-intensive models and confirming that explicit geometric grounding effectively bridges the 2D-to-3D gap. Code will be available at https://github.com/Embrace-Arch/VoxAnchor.

Keywords: Video Spatial Reasoning · 3D Grounding · Multimodal Large Language Models

1 Introduction

Video Spatial Reasoning serves as the fundamental cornerstone of Embodied Spatial Intelligence [38,41], equipping MLLMs with the capacity to actively per-

* Equal contribution. † Corresponding author.

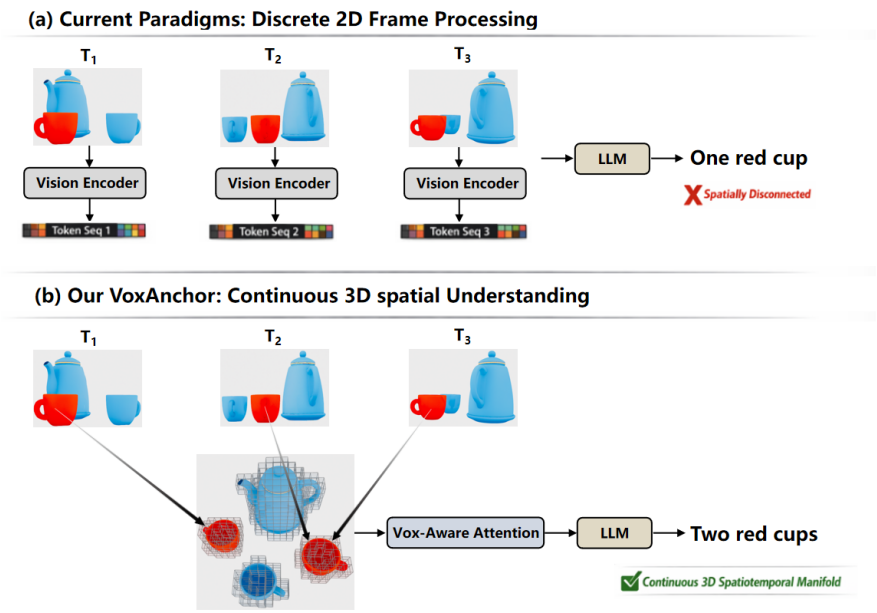

Fig. 1: Conceptual comparison of spatial reasoning paradigms in Video MLLMs. (a) Current Paradigms: Standard models process video streams as discrete sequences of 2D frames, extracting spatially disconnected visual tokens. Lacking an explicit 3D anchor, the language model frequently suffers from *spatial conflation*. It struggles to maintain object persistence under viewpoint shifts, often erroneously conflating spatially distinct entities (e.g., misidentifying two different red cups as a single object). **(b) Our VoxAnchor:** We actively unproject temporal 2D observations into a continuous 3D spatiotemporal manifold. By discretizing the scene into a metric-aware voxel grid and constraining feature interactions via Voxel-Aware Attention, our framework better preserves physical structural integrity. This explicit voxel-semantic grounding enables the model to correctly deduce the spatial layout (e.g., recognizing two distinct cups), bridging the gap between 2D pixels and 3D physical reality.

ceive and reason within dynamic 3D environments from an egocentric perspective [3, 4]. Fundamentally, this capability requires a rigorous structural grounding of the physical world, extending beyond basic visual perception to the coherent interpretation of complex spatial layouts, dense object geometries, and the precise localization of interactive entities. Unlike static imagery where absolute scale recovery remains inherently ambiguous due to the lack of temporal constraints [8], continuous video streams provide rich spatiotemporal dynamics such as motion parallax [13, 52]. These dynamic cues act as indispensable geometric constraints, enabling artificial agents to construct a robust and persistent internal representation of the physical world [42, 46]. Exploiting this temporal continuity is therefore paramount for explicitly decoding the agent's spatial trajectory and absolute physical scale, facilitating the critical representational shift from projective 2D recognition to a rigorous, metric-aware cognition of 3D space.

Although video-based spatial reasoning shows great potential, current methods still face major structural limitations. One research direction tries to improve spatial understanding by using generative world models to create new viewpoints of a scene [3]. The idea is that if a model can imagine how a scene looks from different angles, it may better understand its 3D structure. However, because these generative models rely on probabilistic sampling, they often introduce hallucinated objects or subtle inconsistencies. Such probabilistic inconsistencies fundamentally disrupt spatiotemporal continuity, severely compromising the model's capacity for precise spatial reasoning tasks that demand rigorous representational persistence, such as tracking when an object first appears or understanding the exact sequence of events.

Another common approach incorporates geometric information from pretrained 3D foundation models to provide implicit 3D priors [10, 25, 38, 50]. This strategy is more computationally practical than training MLLMs entirely from scratch on large-scale 3D datasets [15]. However, it still faces two fundamental challenges. First, most current MLLM architectures process video as a series of independent frames [1, 2, 20, 22, 24, 36, 48, 49] rather than as a continuous 3D manifold. By failing to explicitly enforce cross-frame instance consistency, the model cannot reliably resolve identity persistence under ego-motion, thereby precluding the formation of a stable, metric-aware internal environment. Second, the latent spaces of vision-language models are overwhelmingly dominated by semantic inductive biases, which inherently overshadow fine-grained geometric reasoning. This representational imbalance significantly degrades the model's capacity to preserve strict geometric consistency across disparate viewpoints, ultimately limiting its ability to maintain stable 3D spatial relationships across continuous temporal sequences.

As illustrated in Fig. 1(a), current video MLLMs struggle with spatial conflation. We argue that these limitations stem from a fundamental neglect of the physical essence of video data. Rather than a discrete sequence of isolated 2D frames, a video stream is intrinsically a continuous, high-bandwidth projection of an evolving 3D environment onto a 2D temporal manifold [12, 27, 43]. In human cognitive perception, geometric structure and semantic representation are inextricably coupled, serving as an inherent inductive bias that governs coherent scene comprehension [9, 21, 26]. However, contemporary MLLMs structurally fail to invert this projection from 2D pixels to 3D-consistent cognition. This representational chasm arises precisely because these architectures lack an explicit spatial anchoring mechanism to systematically unify fragmented temporal observations into a coherent 3D physical manifold.

To fundamentally rectify these structural limitations, we propose VoxAnchor, a framework designed to explicitly reify the latent 3D physical world by anchoring multimodal representations within a rigorous volumetric space, as illustrated in Fig. 1(b). Rather than relying on implicit 2D heuristics that inevitably lead to spatial conflation, we perform geometry-guided inverse projection, lifting 2D visual tokens and geometric priors into a unified, metric-aware 3D voxel grid. To govern this multimodal integration, we introduce two core designs: a Voxel-

Aware Attention mechanism that strictly constrains feature interactions within physically consistent volumetric units, and a Geo-Semantic Adapter that ensures semantic entities maintain their geometric integrity across temporal streams. Ultimately, this explicit structural grounding transforms the MLLM from a passive discrete frame processor into a metric-aware cognitive agent, capable of robustly decoding absolute self-motion and dynamic object proximity.

- We propose VoxAnchor, a novel framework predicated on explicit voxel-semantic grounding. By unprojecting 2D video observations into a spatiotemporally continuous 3D manifold, our method enables the model to resolve complex spatial dynamics, including continuous camera ego-motion and physical object configurations, directly from monocular visual streams.
- We introduce a Voxel-Aware Attention mechanism that anchors visual tokens and dense geometric priors within a metric-aware 3D volumetric grid. By strictly constraining feature interactions to physically consistent voxel boundaries, this design effectively stitches fragmented 2D frames into a coherent, 3D-grounded representation.
- We develop a Geo-Semantic Adapter that explicitly bridges the representational gap between low-level geometry and high-level semantics. This module enforces strict cross-frame instance consistency, significantly enhancing the model's capacity to reason about evolving spatial relationships over time.
- Empirically, VoxAnchor achieves state-of-the-art average performance on the VSI and VSTI benchmarks. By explicitly constraining multimodal features within a unified 3D volumetric space, our efficient 4B-parameter model excels in tasks demanding strict metric awareness and cross-frame object persistence, demonstrating that rigorous structural grounding profoundly unlocks the spatial capabilities of foundational models.

2 Related Work

2.1 Spatial Foundation Models

The paradigm of 3D reconstruction has transitioned from traditional Structure-from-Motion (SfM) pipelines, such as COLMAP [31], which recover sparse geometry via feature matching, to modern foundation models capable of end-to-end spatial regression. While 3D Gaussian Splatting (3DGS) [19] introduced efficient volumetric representations for high-fidelity novel view synthesis, Scene Representation Transformers [30] moved toward latent scene tokens to reduce dependency on camera calibration. This trajectory has recently shifted toward direct geometric regression; DUSt3R [35] pioneered dense point map estimation from unposed imagery, a strategy scaled by VGGT [33] to achieve joint camera and structural prediction across multiple frames. Building upon these advances, π^3 [37] introduces a scalable permutation-equivariant framework for robust multi-view geometry, while Depth Anything 3 (DA3) [23] establishes a unified system for recovering consistent 3D visual spaces from unconstrained viewpoints. Crucially, IGGT [21] further bridges the gap between reconstruction

and understanding by unifying low-level geometry with instance-level semantics into a 3D-consistent representation. Building on this inspiration, our work leverages such 3D-consistent features to anchor MLLM representations, shifting the focus from static scene analysis to dynamic, metric-aware spatial reasoning in video streams.

2.2 Video-based Spatial Reasoning

Deriving spatial understanding from video streams is fundamental for effective perception and interaction within the physical world. To achieve this, recent frameworks have focused on leveraging pretrained 3D foundation models to extract spatial priors directly from 2D inputs. Specifically, Spatial-MLLM [38] and VG-LLM [50] utilize VGGT [33] as an auxiliary encoder to extract explicit geometric features, while VLM-3R [10] integrates implicit 3D tokens from CUT3R [34]. Broadening this scope, G2VLM [15] enhances perception by explicitly predicting depth to ground spatial reasoning, whereas VST [42] leverages large-scale data training to achieve 3D spatial reasoning capabilities directly on 2D footage. Moving a step further into explicit reconstruction, GS-Reasoner [5] utilizes VGGT-SLAM to construct 3D point clouds in real-time and subsequently employs PTv3 [40] to extract geometric features. Diverging from these direct feature injection strategies, 3DRS [16] adopts a knowledge distillation paradigm, treating VGGT as a teacher model to guide alignment via feature similarity loss. However, these approaches remain dependent on specific geometric foundation models; to transcend this limitation, Cambrian-S [43] pioneers a predictive sensing perspective, driving internal world modeling through a self-supervised next latent-frame prediction objective, thereby achieving robust spatiotemporal perception solely from 2D visual signals without requiring any auxiliary geometric encoders.

2.3 Instance Persistence and 3D-Consistent Representations

Achieving robust instance persistence across temporal sequences requires bridging the gap between 2D visual tracking and 3D geometric consistency. Traditional approaches largely rely on 2D segmentation and optical flow foundation models, such as SAM 2 [29] and CoTracker [18], to maintain identity across frames. However, these methods operate in the projective plane, often failing to preserve metric consistency under significant viewpoint changes or occlusions. Conversely, dynamic 3D reconstruction techniques, such as 4D Gaussian Splatting [39], excel at modeling temporal geometry. While recent advancements (e.g., Gaussian-Grouping [44], Feature 3DGS [51]) have introduced semantic fields to enable discrete entity tracking, they fundamentally rely on computationally prohibitive, per-scene offline optimization. This renders them structurally incompatible with MLLMs, which demand instantaneous, feed-forward spatial reasoning directly from dynamic visual streams. To enable feed-forward 3D consistency, recent models like IGGT [21] elegantly lift 2D masks into structural representations.

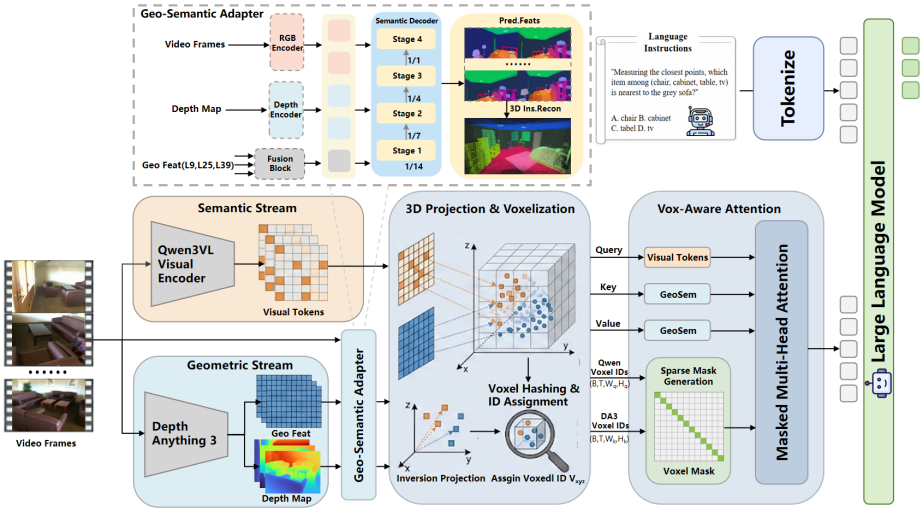

Fig. 2: Overall architecture of our proposed framework. The pipeline processes video frames through a dual-stream encoding design. The **Semantic Stream** utilizes a Qwen3-VL visual encoder to extract visual tokens, while the **Geometric Stream** leverages Depth Anything 3 (DA3) to predict depth maps, camera parameters, and geometric features. A dedicated **Geo-Semantic Adapter** then fuses these inputs to generate 3D-consistent GeoSem representations. In the **3D Projection & Voxelization** module, we unproject both visual tokens and GeoSem features into a unified 3D space, which is subsequently discretized into voxels with unique IDs assigned via Voxel Hashing. Finally, a Voxel-Aware Attention module constructs a sparse Voxel Mask based on spatial ID overlap to guide a Masked Multi-Head Attention—where unprojected visual tokens (Queries) and GeoSem features (Keys/Values) yield spatially-grounded tokens for LLM reasoning.

However, being built atop the VGGT [33] architecture, these representations inherently focus on relative geometric consistency, meaning a representational gap still exists when reasoning about absolute real-world metric scales.

3 Methodology

We introduce VoxAnchor, a framework tailored to reify 2D visual observations into a physically consistent 3D world space via explicit voxel-semantic grounding. As shown in Fig. 2, our methodology systematically integrates three core components: i) Dual-stream encoding (Sec. 3.3), which employs a decoupled architecture to extract high-level semantic tokens and dense metric geometric features in parallel; ii) Geo-Semantic Adapter (Sec. 3.4) to generate 3D-consistent GeoSem representations; and iii) Voxel-Aware Attention (Sec. 3.5) to enforce 3D locality for robust spatial reasoning. By executing this hierarchical unprojection, VoxAnchor effectively transforms fragmented 2D projections into a unified, metric-aware 3D manifold.

3.1 Problem Formulation

We study the problem of *Video Spatial Reasoning*, which aims to equip MLLMs with the capability to perform metric-aware spatial cognition of dynamic 3D environments strictly from 2D egocentric observations. Formally, let $V = \{I_1, I_2, \ldots, I_T\}$ denote a temporal sequence of monocular video frames, where each frame $I_t \in \mathbb{R}^{H \times W \times 3}$. Given a natural language query Q that probes explicit spatial configurations (e.g., metric distances, object scale, or ego-motion), our objective is to learn a parameterized multimodal mapping $f_\theta : (V, Q) \to A$, where A represents the accurate textual response.

Fundamentally, this paradigm diverges from conventional VideoQA—which primarily targets temporal semantic summarization or action recognition—by demanding rigorous *geometric grounding*. The core challenge lies in solving an implicitly ill-posed inverse projection problem: deriving absolute 3D geometric properties purely from a lower-dimensional 2D projective manifold over time. Overcoming this representational gap is a critical prerequisite for Embodied AI, as it strictly dictates an artificial agent's ability to transition from passive visual perception to active, physically consistent interaction within unconstrained 3D environments.

3.2 Motivation

As intuitively illustrated in Fig. 1, the fundamental bottleneck in video spatial reasoning is projective collapse. Standard multimodal paradigms compress continuous 3D dynamics into flattened 2D projections, compelling attention mechanisms to aggregate context based on spurious 2D visual contiguity rather than authentic physical proximity. Even when augmented with implicit 3D priors, existing MLLMs remain structurally bound to these 2D positional biases [5, 10, 25, 38, 50], inevitably leading to the spatial conflation of distinct entities under ego-motion.

To fundamentally resolve this representational mismatch, VoxAnchor introduces a physically-grounded attention routing mechanism. Conceptually akin to 2D-to-3D feature lifting in autonomous driving [28], we explicitly unproject 2D visual tokens into a unified 3D coordinate system using camera intrinsics, extrinsics, and predicted depth. By restricting cross-attention interactions strictly to physically consistent voxel boundaries, we replace deceptive 2D pixel adjacency with metric-aware 3D proximity, thereby effectively preserving stable instance identity and absolute metric scale.

3.3 Dual-Stream Spatio-Semantic Encoding

As shown in Fig. 2, VoxAnchor adopts a dual-stream encoding pipeline to separately model visual semantics and physical structure. Given the input video sequence V, we independently process the observations through two specialized pathways: a Semantic Stream and a Geometric Stream. This decoupled design

prevents premature representational entanglement, ensuring that high-level semantic reasoning remains orthogonal to the precision of metric structural priors.

Semantic Stream To capture high-level contextual and categorical semantics, we map the input sequence through a pre-trained Vision-Language foundation encoder (Qwen3-VL). This mapping yields a sequence of 2D visual tokens, formally denoted as $F_{sem} \in \mathbb{R}^{T \times \frac{H}{16} \times \frac{W}{16} \times C_s}$. While these embeddings encapsulate dense semantic priors, they reside strictly within a projective 2D space, structurally devoid of absolute metric depth and consistent 3D configurations.

Geometric Stream. Concurrently, to endow the model with a rigorous metric hypothesis space, we introduce a parallel geometric stream instantiated by a 3D foundation model (Depth Anything 3). This stream acts as a deterministic inverse-rendering function. Benefiting from DA3's unified architecture, it directly regresses the unconstrained 2D sequence into a set of explicit 3D structural parameters within a single forward pass: dense metric depth maps $D \in \mathbb{R}^{T \times H \times W}$, camera intrinsic matrices K, and extrinsic ego-motion trajectories E. To preserve multi-scale structural fidelity, we additionally extract hierarchical geometric abstractions $F_{geo} \in \mathbb{R}^{T \times \frac{H}{14} \times \frac{W}{14} \times C_g}$ from the intermediate layers (e.g., Layers 9, 25, and 39) of the backbone. This stream ensures that metric-aware features remain representationally independent of semantic tokens, facilitating a more principled spatial-semantic fusion in the subsequent voxelization stage.

3.4 Geo-Semantic Adapter

The foundational challenge in aligning multimodal priors lies in their inherent representation gap: visual tokens (F_{sem}), extracted by the Qwen3-VL visual encoder, encapsulate dense semantic abstractions but lack spatial structure, whereas geometric embeddings (F_{geo}), output by the DA3 backbone, provide rich 3D structural priors but remain semantically agnostic. To systematically bridge this chasm, we propose the Geo-Semantic Adapter (see Fig. 2, top left), designed to reconcile multi-scale geometric features and align cross-resolution spatial representations.

Multi-Scale Structural Refinement. As shown in Fig. 2, the Geo-Semantic Adapter refines semantic visual tokens using predicted depth maps and RGB cues. Specifically, spatial modulators are scale-specific feature maps generated by progressive upsampling blocks, which inject structural cues through gated skip connections at full, 1/4, and 1/7 resolutions. This bridges the coarse DA3 grid (1/14) and image resolution for finer object-level grounding.

- **Stage 1: Bottleneck Alignment.** The hierarchical geometric features (F_{geo}) from the DA3 backbone are aggregated via a Fusion Block. This forms a structural bottleneck aligned with the semantic stream, serving as the starting point for subsequent multi-scale feature refinement.
- **Stages 2-4: Gated Structural Injection.** During the progressive decoding stages (Stage 2, 3, and 4 in Fig. 2), scale-matched modulators at 1/7, 1/4, and 1/1 resolutions are injected into the semantic stream via gated skip connections. This mechanism adaptively integrates RGB texture and depth

geometry to enforce physical consistency, thereby mitigating the spatial entanglement of distinct object entities.
- **Final Output.** The final fused representation is processed through a 1×1 convolution and L_2 normalization. This stage yields a continuous, metric-aware spatial feature field at full resolution, providing a high-fidelity foundation for the subsequent voxelization. Crucially, this representation ensures cross-frame 3D object consistency, enabling semantic entities to maintain their identity and geometric integrity across complex temporal sequences.

Instance-Guided Representation Regularization. To guarantee that the fused representations exhibit strict instance-level consistency under complex temporal ego-motion, the adapter undergoes a parameter-efficient pre-training phase. By bottlenecking dense pixel interactions through sparse instance proxies, this design facilitates the processing of high-resolution spatial features while significantly reducing memory overhead. Specifically, by treating per-frame instance masks as a physical partitioning of the scene, we dynamically aggregate latent prototypes p_k for each distinct object entity k by calculating the mean embedding of all pixels assigned to the corresponding instance mask. We optimize the embedding space via a proxy-based InfoNCE objective, which projects pixel-level representations f_i onto a normalized hypersphere and maximizes their mutual information with the target dynamic prototype p_{y_i}:

$$L_{info} = -\frac{1}{N_{fg}} \sum_{i=1}^{N_{fg}} \log \frac{\exp(f_i \cdot p_{y_i}/\tau)}{\sum_{j=1}^{K} \exp(f_i \cdot p_j/\tau)} \tag{1}$$

where N_{fg} represents the foreground pixel cardinality, K denotes the dynamically inferred instance count, and τ is the temperature scaling factor. This objective is further augmented with intra-instance compactness and inter-instance margin penalties. Together, they serve as a rigorous geometric regularization that explicitly disentangles distinct objects within the latent metric space, ensuring robust downstream spatial reasoning.

3.5 Voxel-Aware Spatial Anchoring and Interaction

To bridge the representational gap between projective pixels and 3D physical reality, we introduce an explicit geometric grounding mechanism. Unlike conventional MLLMs that passively process implicit geometric embeddings, we actively leverage dense depth priors and camera extrinsics to construct a structurally consistent metric environment. Our approach explicitly unprojects visual tokens into a unified 3D coordinate system. By treating the predicted depth as a first-order geometric constraint, we ensure that subsequent multimodal interactions are grounded within a coherent 3D physical space.

3D Projection and Voxelization. As illustrated in the 3D Projection & Voxelization module of Fig. 2, we define an analytic inverse projection operator to map each 2D visual token from the projective plane back to its absolute 3D coordinate. Specifically, the unstructured visual tokens (shown as orange

and blue grids in the Semantic Stream) and the 3D-consistent GeoSem features from the adapter are processed through an Inversion Projection. For a token anchored at pixel coordinate (u, v) with its corresponding predicted depth d, intrinsic matrix K, and extrinsic ego-motion matrix E, the transformation to the world coordinate $p_{world} \in \mathbb{R}^3$ is formulated as:

$$p_{world} = E^{-1}(d \cdot K^{-1}[u, v, 1]^T) \tag{2}$$

To render this continuous 3D space computationally tractable for the LLM, we partition the environment into a regular volumetric grid with a voxel size of $s = 0.5\,\text{m}$. We then apply a spatial hashing function to map these dense coordinates into a set of discrete, globally unique Voxel IDs (H). This process assigns a unique physical address to both the unstructured visual tokens and the dual-stream Geo-Semantic embeddings, effectively anchoring multimodal features within a structured 3D grid.

Voxel-Aware Attention. To enforce strict spatial locality, we fundamentally restructure the LLM's attention routing as shown in the Voxel-Aware Attention module of Fig. 2. We define the 2D visual tokens as Queries (Q), while the dual-stream GeoSem features (derived from the geometric stream and refined by the Geo-Semantic Adapter) serve as Keys (K) and Values (V). Guided by the overlapping Voxel IDs (e.g., matching Qwen Voxel IDs with DA3 Voxel IDs in Sparse Mask Generation), we dynamically generate a Sparse Voxel Mask (M). The interaction between a query token i and a key token j is strictly gated by their physical co-location:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} + M\right)V \tag{3}$$

where the spatial penalty term $M_{i,j} = 0$ if $H_i = H_j$ (indicating physical co-location), and $M_{i,j} = -\infty$ otherwise. In practice, this operation is extended to a multi-head mechanism, where the spatial regularization M is applied across all attention heads to ensure consistent 3D locality. This physically motivated regularization prunes fallacious non-local interactions, compelling the model to synthesize spatial intelligence strictly from geometrically valid 3D neighborhoods rather than deceptive 2D adjacencies.

4 Experiments

4.1 Datasets and Experiment setup

To rigorously assess the model's capacity for explicit geometric grounding and spatiotemporal reasoning, we evaluate our framework on two complementary benchmarks. First, to evaluate representational persistence under dynamic ego-motion, we adopt VSTI-Bench [10]. Featuring approximately 6,000 QA pairs, it explicitly challenges the model's ability to resolve continuous temporal dynamics, including camera displacement, view-consistent object relations, and spatial

shifts. Furthermore, to assess explicit 3D cognition and metric awareness, we utilize VSI-Bench [41], comprising over 5,000 egocentric QA pairs sourced from ScanNet [6], ScanNet++ [45], and ARKitScenes [7]. This dataset serves as a rigorous testbed, probing both spatial relational understanding through configurational tasks (e.g., object counting, relative positioning) and scale-aware metric regression via measurement estimation tasks (e.g., absolute distance, object/room sizing). Following standard protocols, we report absolute Accuracy [11, 14, 47] for discrete Multiple-Choice (MCA) queries and Mean Relative Accuracy [41] for continuous Numerical Answer (NC) tasks.

Baselines. We compare our model with a diverse set of proprietary and open-source VLMs (e.g., GPT-4o [17], Gemini-1.5 Pro [32], LLaVA-NeXT-Video [49], Qwen2.5-VL [2], Qwen3-VL [1]). We also compare the performance with recent spatial-enhanced models for spatial reasoning, including VG-LLM [50], Spatial-MLLM [38], GS-Reasoner [5] and VLM-3R [10].

Implementation Details. Our framework uses Qwen3-VL-4B [1] as the base MLLM and DA3 [23] for geometric perception. Training follows two stages. First, the Geo-Semantic Adapter is contrastively pre-trained on InsScene-15K [21] using DA3 features from Layers 9, 25, and 39 on a single NVIDIA L40 GPU. Second, the model is fine-tuned on VSI590K [43] and the VSTI-Bench [10] training split for two epochs with AdamW, a learning rate of 1×10^{-4}, and a global batch size of 64. During fine-tuning, the Qwen3-VL visual encoder, DA3 backbone, and pre-trained Geo-Semantic Adapter are frozen; we update the Voxel-Aware Attention module and LoRA parameters of the language model with rank $r = 128$ and scaling factor $\alpha = 256$. Multimodal fine-tuning is parallelized across 4 NVIDIA A800 GPUs.

Evaluation on VSTI-Bench. Table 1 details the empirical evaluations on VSTI-Bench, which introduces complex spatiotemporal dynamics such as continuous camera ego-motion and inter-object relational shifts. VoxAnchor effectively mitigates representational drift, achieving a leading average score of 66.3%. Standard 2D-based models experience significant performance degradation as temporal dynamics increase, primarily because they struggle to maintain instance-level consistency under complex camera ego-motion. In contrast, our Voxel-Aware Attention mechanism ensures that semantic entities preserve their geometric integrity across disparate viewpoints. By strictly constraining feature interactions within physically consistent volumetric boundaries, VoxAnchor surpasses the strong VLM-3R-7B baseline by a 7.5% margin. These quantitative gains validate that explicit 3D anchoring is a critical prerequisite for reliable spatiotemporal reasoning, allowing the model to successfully synthesize fragmented temporal observations into a coherent, 3D-aware representation.

Evaluation on VSI-Bench. As presented in Table 2, VoxAnchor establishes a new state-of-the-art with an average score of 66.8%, significantly narrowing the representational gap between 2D projective observations and 3D physical reality. We observe a particularly pronounced empirical advantage in Numerical Answer tasks that demand strict metric awareness, such as Room Size (71.0%) and Object Size (72.5%) estimation. In these regimes, conventional MLLMs exhibit sys-

Table 1: Evaluations on VSTI-Bench for 3D spatiotemporal reasoning tasks. We compare VoxAnchor against proprietary models, open-sourced VLMs, and spatial-enhanced baselines, where **bold** indicates the best performance in each category.

Methods	Avg.	Numerical Question		Multiple-Choice Question		
		Cam-Obj Abs Dist	Camera Displace	Camera Mov Dir	Obj-Obj Rel Pos	Cam-Obj Rel Dist
Proprietary Models (API)						
GPT-4o	38.2	29.5	23.4	37.3	58.1	42.5
Gemini-1.5 Flash	32.1	28.5	20.9	24.4	52.6	33.9
Open-sourced VLMs						
LongVA-7B	32.3	13.5	5.1	43.7	57.9	41.2
InternVL2-8B	43.5	32.9	13.5	48.0	68.0	55.0
InternVL2-40B	43.2	11.9	34.9	33.3	63.8	72.2
LongVILA-8B	30.5	20.0	11.6	35.4	52.3	33.4
VILA-1.5-8B	37.3	30.1	27.3	42.2	50.4	36.7
VILA-1.5-40B	38.2	28.2	15.7	28.8	65.4	53.0
Qwen2.5-VL-7B	38.2	22.9	4.9	47.4	65.9	49.9
Qwen2.5-VL-72B	40.3	18.0	10.0	41.0	74.2	58.4
Qwen3-VL-4B	39.7	29.5	32.9	46.0	68.6	25.2
LLaVA-OneVision-7B	41.7	29.9	19.3	47.5	62.1	49.8
LLaVA-NeXT-Video-7B	40.0	28.2	1.8	49.8	64.7	55.6
LLaVA-NeXT-Video-72B	44.0	32.3	10.5	48.1	78.3	50.9
Spatial-Enhanced Models						
VLM-3R-7B	58.8	39.4	39.6	60.6	86.5	68.6
VoxAnchor-4B (Ours)	**66.3**	**40.5**	**46.1**	**83.6**	**89.5**	**71.9**

tematic failure due to their structural reliance on scale-invariant 2D projective priors. For instance, while proprietary models like GPT-4o yield suboptimal performance (a mere 5.3%) on Absolute Distance estimation, VoxAnchor achieves a remarkable 52.6%. This nearly ten-fold improvement empirically validates that geometric grounding is not an emergent property of scale, but rather a consequence of explicit structural inductive biases. By anchoring visual features within a metric-aware 3D environment, our 4B-parameter model effectively recovers absolute physical scales, overcoming the spatial conflation prevalent in traditional 2D attention mechanisms.

4.2 Ablation Study

To rigorously isolate the efficacy of our architectural constraints, we conduct comprehensive ablations on VSTI-Bench. These experiments systematically dissect the impact of different 3D foundation models and the incremental contributions of our core structural components, demonstrating how explicit grounding unlocks spatial intelligence even within a compact 4B-parameter footprint.

Impact of Geometric Priors. We first investigate the representational influence of distinct 3D foundation models (Table 3). While augmenting the baseline (54.4%) with feed-forward 3D spatial models (e.g., VGGT or IGGT)

Table 2: Comparative results on VSI-Bench [41]. VoxAnchor is evaluated against general-purpose and spatial-enhanced MLLMs across numerical and multiple-choice spatial queries. **Bold** denotes top performance.

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-4o	34.0	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5
Gemini-1.5 Flash	42.1	49.8	30.8	53.5	54.4	37.7	41.0	31.5	37.8
Gemini-1.5 Pro	45.4	56.2	30.9	64.1	43.6	51.3	46.3	36.0	34.6
Open-sourced VLMs									
LongVA-7B	29.2	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7
InternVL2-8B	34.6	23.1	28.7	48.2	39.8	36.7	30.7	29.9	39.6
InternVL2-40B	36.0	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6
LongVILA-8B	21.6	29.1	9.1	16.7	0.0	29.6	30.7	32.5	25.5
VILA-1.5-8B	28.9	17.4	21.8	50.3	18.8	32.1	34.8	31.0	24.8
VILA-1.5-40B	31.2	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9
Qwen2.5-VL-7B	33.0	40.9	14.8	43.4	10.7	38.6	38.5	33.0	29.8
Qwen2.5-VL-72B	37.0	25.1	29.3	54.5	38.8	38.2	37.0	34.0	28.9
Qwen3-VL-4B	54.8	62.6	44.5	70.3	63.9	56.4	46.3	32.0	62.3
LLaVA-OneVision-7B	32.4	47.7	20.2	47.4	12.3	42.5	35.2	29.4	24.4
LLaVA-OneVision-72B	40.2	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6
LLaVA-NeXT-Video-7B	35.6	48.5	14.0	47.8	24.2	43.5	42.4	34.0	30.6
LLaVA-NeXT-Video-72B	40.9	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6
Spatial-Enhanced Models									
VG-LLM-4B	47.3	66.0	37.8	55.2	59.2	44.6	45.6	33.5	36.4
VG-LLM-8B	50.7	67.9	37.7	58.6	62.0	46.6	40.7	32.4	59.2
Spatial-MLLM-4B	48.4	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3
VLM-3R-7B	60.9	**70.2**	49.4	69.2	67.1	65.4	80.5	45.4	40.1
GS-Reasoner-7B	64.7	69.1	**61.9**	70.0	65.7	65.4	**88.9**	44.3	52.3
VoxAnchor-4B (Ours)	**66.8**	66.3	52.6	**72.5**	**71.0**	**67.3**	80.0	**46.4**	**78.0**

yields notable improvements, their structural paradigms inherently bottleneck metric reasoning. Specifically, although IGGT attempts to bridge the 2D-to-3D gap by appending an instance-level semantic head to the VGGT backbone, it inevitably inherits the scale ambiguity of its relative projective reconstruction, leaving the generated 3D features misaligned with absolute physical metrics. In contrast, substituting these with Depth Anything 3 (DA3) achieves a peak average accuracy of 63.0%. This performance disparity highlights a critical insight: by leveraging a dedicated metric regression head to explicitly scale the predicted depth manifold, DA3 successfully recovers the absolute real-world scale. This empirical result validates our architectural choice: compared to scale-ambiguous implicit encodings, DA3 provides a mathematically rigorous, metric-aware structural prior, fundamentally establishing a reliable 3D physical reference frame for downstream spatiotemporal anchoring.

Efficacy of Explicit 3D Grounding. Table 4 systematically dissects the gains from our core structural modules while keeping DA3 as the geometric backbone. The first row uses DA3 geometric features fused with visual tokens through standard cross-attention, without voxel masking or the Geo-Semantic Adapter, achieving 63.0%. Replacing standard cross-attention with Voxel-Aware Attention improves the average score to 64.5% (+1.5), showing that voxel-constrained routing provides additional gains beyond simply injecting DA3 features. By explicitly suppressing spurious 2D spatial conflation, this mechanism delivers sub-

Table 3: Ablation of 3D foundation model. We compare different 3D foundation models (VGGT, IGGT, DA3) for spatial reasoning tasks on VSTI-Bench.

Methods	Avg.	Numerical Question		Multiple-Choice Question		
		Cam-Obj Abs. Dist.	Cam. Displace	Cam. Mov. Dir.	Obj-Obj Rel. Pos.	Cam-Obj Rel. Dist.
Baseline						
Qwen3-VL-SFT	54.4	36.0	32.7	53.6	79.1	70.6
3D Foundation Model						
Baseline + VGGT	59.5	38.6	40.5	62.3	86.4	70.0
Baseline + IGGT	60.7	38.0	41.1	65.8	87.3	71.3
Baseline + DA3	63.0	40.0	43.9	73.2	87.1	71.0

Table 4: Ablation of structural components on VSTI-Bench. We keep the geometric backbone fixed and only vary Voxel-Aware Attention and the Geo-Semantic Adapter. **Bold** denotes the best performance in each metric.

Voxel-Aware Attention	Geo-Semantic Adapter	Avg.	Numerical Question		Multiple-Choice Question		
			Cam-Obj Abs. Dist.	Camera Displace	Camera Mov. Dir.	Obj-Obj Rel. Pos.	Cam-Obj Rel. Dist.
✗	✗	63.0	40.0	43.9	73.2	87.1	71.0
✓	✗	64.5	39.4	43.0	79.2	**89.9**	**73.0**
✓	✓	**66.3**	**40.5**	**46.1**	**83.6**	89.5	71.9

stantial gains in directional and relative reasoning, such as Camera Mov. Dir. (+6.0%) and Obj-Obj Rel. Pos. (+2.8%). Adding the Geo-Semantic Adapter further raises the average score to 66.3% (+1.8), suggesting that semantic-aware geometric refinement improves cross-frame consistency for spatiotemporal reasoning. The gains are especially visible on Camera Mov. Dir. (+4.4%) and Camera Displace (+3.1%) over the Voxel-Aware baseline. Overall, these ablations show that Voxel-Aware Attention and the Geo-Semantic Adapter provide complementary improvements on top of DA3-based geometric features. Additional analyses on voxel-size sensitivity, robustness to noisy depth, and inference efficiency are provided in the supplementary material.

5 Conclusion

In this work, we introduce VoxAnchor to mitigate spatial conflation of video MLLMs via explicit voxel-semantic grounding. By unprojecting 2D observations into a metric-aware 3D manifold, we constrain attention mechanisms using voxel-based physical locality, effectively bridging the 2D-to-3D representational gap. Empirically, our compact 4B-parameter model achieves state-of-the-art spatial reasoning performance, outperforming significantly larger baselines in tasks demanding metric precision and object persistence. These results underscore that explicit 3D structural priors are indispensable for advancing embodied spatial intelligence.

Acknowledgements

We thank the reviewers and the area chair for their critical and constructive comments. This work was supported by the National Key R&D Program of China No. 2024YFC3015801, National Science Fund of China under Grant Nos. U24A20330, 62361166670, and 62276144.

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025) 3, 11
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. ArXiv **abs/2502.13923** (2025), https://api.semanticscholar.org/CorpusID:276449796 3, 11
3. Cao, M., Li, X., Liu, X., Reid, I., Liang, X.: Spatialdreamer: Incentivizing spatial reasoning via active mental imagery. arXiv preprint arXiv:2512.07733 (2025) 2, 3
4. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14455–14465 (2024) 2
5. Chen, Y., Qi, Z., Zhang, W., Jin, X., Zhang, L., Liu, P.: Reasoning in space via grounding in the world (2025), https://arxiv.org/abs/2510.13800 5, 7, 11
6. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T.A., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2432–2443 (2017), https://api.semanticscholar.org/CorpusID:7684883 11
7. Dehghan, A., Baruch, G., Chen, Z., Feigin, Y., Fu, P., Gebauer, T., Kurz, D., Dimry, T., Joffe, B., Schwartz, A., Shulman, E.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. In: NeurIPS Datasets and Benchmarks (2021), https://api.semanticscholar.org/CorpusID:237277377 11
8. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Neural Information Processing Systems (2014), https://api.semanticscholar.org/CorpusID:2255738 2
9. Epstein, R.A., Baker, C.I.: Scene perception in the human brain. Annual review of vision science (2019), https://api.semanticscholar.org/CorpusID:195260816 3
10. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., Xu, H., Theiss, J., Chen, T., Li, J., Tu, Z., Wang, Z., Ranjan, R.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. ArXiv **abs/2505.20279** (2025), https://api.semanticscholar.org/CorpusID:279449925 3, 5, 7, 10, 11
11. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Ji, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. 2025 IEEE/CVF Conference

on Computer Vision and Pattern Recognition (CVPR) pp. 24108–24118 (2024), https://api.semanticscholar.org/CorpusID:270199408 11
12. Gibson, J.J.: The ecological approach to visual perception: Classic edition (2014), https://api.semanticscholar.org/CorpusID:69960436 3
13. Godard, C., Aodha, O.M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 3827–3837 (2018), https://api.semanticscholar.org/CorpusID:46936649 2
14. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D.X., Steinhardt, J.: Measuring massive multitask language understanding. ArXiv **abs/2009.03300** (2020), https://api.semanticscholar.org/CorpusID:221516475 11
15. Hu, W., Lin, J., Long, Y., Ran, Y., Jiang, L., Wang, Y., Zhu, C., Xu, R., Wang, T., Pang, J.: G2vlm: Geometry grounded vision language model with unified 3d reconstruction and spatial reasoning. ArXiv **abs/2511.21688** (2025), https://api.semanticscholar.org/CorpusID:283262189 3, 5
16. Huang, X., Wu, J., Xie, Q., Han, K.: Mllms need 3d-aware representation supervision for scene understanding. ArXiv **abs/2506.01946** (2025), https://api.semanticscholar.org/CorpusID:279119197 5
17. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 11
18. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXII. Lecture Notes in Computer Science, vol. 15120, pp. 18–35 (2024) 5
19. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics (TOG) **42**, 1 – 14 (2023), https://api.semanticscholar.org/CorpusID:259267917 4
20. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. ArXiv **abs/2408.03326** (2024), https://api.semanticscholar.org/CorpusID:271719914 3
21. Li, H., Zou, Z., Liu, F., Zhang, X., Hong, F., Cao, Y., Lan, Y., Zhang, M., Yu, G., Zhang, D., Liu, Z.: Iggt: Instance-grounded geometry transformer for semantic 3d reconstruction (2025), https://arxiv.org/abs/2510.22706 3, 4, 5, 11
22. Lin, B., Zhu, B., Ye, Y., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. ArXiv **abs/2311.10122** (2023), https://api.semanticscholar.org/CorpusID:265281544 3
23. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. ArXiv **abs/2511.10647** (2025), https://api.semanticscholar.org/CorpusID:282992334 4, 11
24. Lin, J., Yin, H., Ping, W., Lu, Y., Molchanov, P., Tao, A., Mao, H., Kautz, J., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 26679–26689 (2023), https://api.semanticscholar.org/CorpusID:266174746 3
25. Liu, Z., Du, Y., Fu, T., Su, S., Ho, C., Wang, C.: Vision-language memory for spatial reasoning. ArXiv **abs/2511.20644** (2025), https://api.semanticscholar.org/CorpusID:283251433 3, 7
26. Malcolm, G.L., Groen, I.I.A., Baker, C.I.: Making sense of real-world scenes. Trends in cognitive sciences **20 11**, 843–856 (2016), https://api.semanticscholar.org/CorpusID:3682956 3

27. Marr, D.: Vision: A computational investigation into the human representation and processing of visual information (1983), https://api.semanticscholar.org/CorpusID:267799236 3
28. Philion, J., Fidler, S.: Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In: European Conference on Computer Vision (2020), https://api.semanticscholar.org/CorpusID:221112099 7
29. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R.B., Dollár, P., Feichtenhofer, C.: SAM 2: Segment anything in images and videos. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025 (2025) 5
30. Sajjadi, M.S.M., Meyer, H., Pot, E., Bergmann, U.M., Greff, K., Radwan, N., Vora, S., Lucic, M., Duckworth, D., Dosovitskiy, A., Uszkoreit, J., Funkhouser, T.A., Tagliasacchi, A.: Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6219–6228 (2021), https://api.semanticscholar.org/CorpusID:244709599 4
31. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 4104–4113 (2016), https://api.semanticscholar.org/CorpusID:1728538 4
32. Team, G., Georgiev, P., Lei, V.I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., et al.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530 (2024) 11
33. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotný, D.: Vggt: Visual geometry grounded transformer. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5294–5306 (2025), https://api.semanticscholar.org/CorpusID:277043968 4, 5, 6
34. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10510–10522 (2025), https://api.semanticscholar.org/CorpusID:275789153 5
35. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 20697–20709 (2023), https://api.semanticscholar.org/CorpusID:266436038 4
36. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 3
37. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning (2025), https://arxiv.org/abs/2507.13347 4
38. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025) 1, 3, 5, 7, 11
39. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 20310–20320 (2023), https://api.semanticscholar.org/CorpusID:263908793 5
40. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler, faster, stronger. 2024 IEEE/CVF Conference on

Computer Vision and Pattern Recognition (CVPR) pp. 4840–4851 (2023), https://api.semanticscholar.org/CorpusID:266335894 5
41. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: 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) 1, 11, 13
42. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025) 2, 5
43. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Lu, D., Fergus, R., LeCun, Y., Li, F.F., Xie, S.: Cambrian-s: Towards spatial supersensing in video. ArXiv **abs/2511.04670** (2025), https://api.semanticscholar.org/CorpusID:282812799 3, 5, 11
44. Ye, M., Danelljan, M., Yu, F., Ke, L.: Gaussian grouping: Segment and edit anything in 3d scenes. In: European Conference on Computer Vision (2023), https://api.semanticscholar.org/CorpusID:265551523 5
45. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) pp. 12–22 (2023), https://api.semanticscholar.org/CorpusID:261064784 11
46. Yin, B., Wang, Q., Zhang, P., Zhang, J., Wang, K., Wang, Z., Zhang, J., Chandrasegaran, K., Liu, H., Krishna, R., et al.: Spatial mental modeling from limited views. In: Structural Priors for Vision Workshop at ICCV'25 (2025) 2
47. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 9556–9567 (2023), https://api.semanticscholar.org/CorpusID:265466525 11
48. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Conference on Empirical Methods in Natural Language Processing (2023), https://api.semanticscholar.org/CorpusID:259075356 3
49. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), https://llava-vl.github.io/blog/2024-04-30-llava-next-video/ 3, 11
50. Zheng, D., Huang, S., Li, Y., Wang, L.: Learning from videos for 3d world: Enhancing mllms with 3d vision geometry priors. ArXiv **abs/2505.24625** (2025), https://api.semanticscholar.org/CorpusID:279070749 3, 5, 7, 11
51. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 21676–21685 (2023), https://api.semanticscholar.org/CorpusID:265722936 5
52. Zhou, T., Brown, M.A., Snavely, N., Lowe, D.G.: Unsupervised learning of depth and ego-motion from video. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6612–6619 (2017), https://api.semanticscholar.org/CorpusID:11977588 2