OpenSpatial: A Principled Data Engine for Empowering Spatial Intelligence

Jianhui Liu[1,4,*], Haoze Sun[4,*], Wenbo Li[4,†,✉], Yanbing Zhang[4], Rui Yang[1], Zhiliang Zhu[4], Yijun Yang[2,4], Shenghe Zheng[2,4], Nan Jiang[4], Jiaxiu Jiang[4], Haoyang Huang[4], Tien-Tsin Wong[3], Nan Duan[4], and Xiaojuan Qi[1,✉]

[1] The University of Hong Kong
[2] The Hong Kong University of Science and Technology
[3] Monash University
[4] Joy Future Academy, JD

Abstract. Spatial understanding is a fundamental cornerstone of human-level intelligence. Nonetheless, current research predominantly focuses on domain-specific data production, leaving a critical void: the absence of a principled, open-source engine capable of fully unleashing the potential of high-quality spatial data. To bridge this gap, we elucidate the design principles of a robust data generation system and introduce OpenSpatial—an open-source data engine engineered for high quality, extensive scalability, broad task diversity, and optimized efficiency. OpenSpatial adopts 3D bounding boxes as the fundamental primitive to construct a comprehensive data hierarchy across five foundational tasks: Spatial Measurement (SM), Spatial Relationship (SR), Camera Perception (CP), Multi-view Consistency (MC), and Scene-Aware Reasoning (SAR). Leveraging this scalable infrastructure, we curate OpenSpatial-3M, a large-scale dataset comprising 3 million high-fidelity samples. Extensive evaluations demonstrate that versatile models trained on our dataset achieve state-of-the-art performance across a wide spectrum of spatial reasoning benchmarks. Notably, the best-performing model exhibits a substantial average improvement of 19%, relatively. Furthermore, we provide a systematic analysis of how data attributes influence spatial perception. By open-sourcing both the engine and the 3M-scale dataset, we provide a robust foundation to accelerate future research in spatial intelligence.

Code: github.com/VINHYU/OpenSpatial

Dataset: huggingface.co/datasets/JoyAI-Image-OpenSpatial

Keywords: Spatial Intelligence · Vision-language Model

1 Introduction

Multi-modal large language models (MLLMs) [1,3,14,21,36,40,41,49] have progressed from image-text alignment to instruction-following systems capable of

* Equal Contribution, † Project Lead, ✉ Corresponding Authors

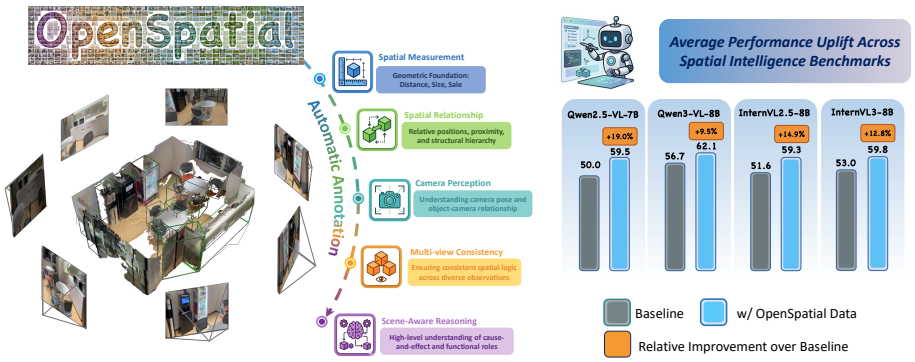


Fig. 1: Overview of OpenSpatial. The left panel provides a high-level schematic of the OpenSpatial pipeline. The right panel demonstrates that models trained with OpenSpatial-generated data exhibit significant improvements in spatial intelligence. Notably, the evaluation benchmarks are consistent with those reported in Tab. 1.

visual intelligence. Yet their spatial competence still trails their semantic expressiveness: models can produce convincing descriptions but often fail to perceive accurate distance, maintain multi-view consistency, or construct spatial cognitive maps – capabilities central to embodied decision-making [24, 29] and robotics [5, 20, 37]. This gap has motivated "spatialized" VLMs [11, 18, 34, 53, 54] and dedicated benchmarks [19, 52, 55, 56], yielding measurable improvements. However, these gains remain uneven across tasks and scenes, suggesting that the bottleneck is not architecture alone but the foundations of spatial generalization.

A major foundation is **data**: which spatial signals are present, how they are synthesized, and what distributions they cover. Current data-centric efforts, however, face two systemic obstacles. First, limited diversity of spatial data constrains the robustness of state-of-the-art models. This scarcity leads to "spatial myopia", where models exhibit high benchmark scores but lack the versatility required for real-world environments. Second, and more critically, much spatial data is generated by closed, under-specified pipelines. Existing works [7, 53, 54] release only fixed, preprocessed datasets (sometimes in limited subsets) while keeping their generative engines proprietary, making it difficult to run controlled ablations, scale data consistently, or study which design choices actually drive spatial capability. This black-box ecosystem fragments progress into disconnected silos and raises the barrier to reproducible, systematic advancement.

We argue that advancing spatial intelligence requires moving beyond static dataset releases toward open, reusable data infrastructure. We therefore introduce **OpenSpatial**, as shown in Fig. 1, an open-source data engine that synthesizes high-quality, scalable, and task-diverse supervision for spatial understanding. OpenSpatial is built on three key designs. **(1) 3D box–centric grounding for quality**: by anchoring supervision in object-aligned 3D boxes rather than 2D projections [22], it yields high-fidelity, viewpoint-consistent labels that capture true 3D structure and support metric reasoning. **(2) 3D lifting for scalability**: it automatically elevates sparse cues into high-quality 3D box priors, enabling

data generation to extend beyond curated sets to in-the-wild sources. **(3) Scene-graph-driven synthesis for diversity**: it programmatically enumerates objects, attributes, and relations to generate balanced QA across measurement, relations, camera changes, multi-view consistency, and scene-level reasoning, mitigating "spatial myopia". OpenSpatial is further engineered for throughput via parallel execution and feature reuse, enabling rapid large-scale annotations. Together, these choices make spatial supervision transparent and controllable, supporting principled ablations, reliable scaling, and improved generalization.

Built on this infrastructure, we curate the OpenSpatial-3M dataset, a 3-million-sample training suite spanning five core capabilities– Spatial Measurement, Spatial Relationship, Camera Perception, Multi-view Consistency, and Scene-Aware Reasoning, organized as a progressive curriculum that bridges egocentric observations with stable world-coordinate understanding. Models fine-tuned on this data achieve state-of-the-art performance on challenging spatial benchmarks (*e.g.*, BLINK, AllAngles, MMSI), consistently surpassing strong open-source baselines. As shown in Fig. 1, OpenSpatial-3M consistently enhances performance across various architectures, yielding a **14.1%** average improvement and a maximum gain of **19%** over the baseline. Beyond performance, OpenSpatial's modular pipeline enables controlled analyses of which design choices drive improvements (*e.g.*, box-centric grounding and data filtering), supporting reproducible, data-driven scaling of spatial perception across architectures.

In summary, this work advances the frontier of spatial intelligence through three key contributions.

- We introduce **OpenSpatial**, an open-source, controllable data engine for synthesizing high-quality, scalable, and task-diverse spatial supervision.
- We release **OpenSpatial-3M**, a large-scale curriculum-style training suite covering five foundational spatial capabilities.
- We demonstrate strong empirical gains and provide diagnostic analyses enabled by the engine's modularity, clarifying how specific data designs improve spatial generalization and offering a reproducible foundation for future work.

2 Related Work

Large Vision-Language Models. The rapid evolution of Large Vision-Language Models (LVLMs) has revolutionized multimodal intelligence by bridging high-dimensional visual perception with complex linguistic reasoning, a progress fundamentally anchored in the advancement of Visual Instruction Tuning. This paradigm, pioneered by LLaVA [30], demonstrated that projecting deep visual features into the LLM's embedding space via a lightweight interface enables the model to follow complex multimodal prompts with sophisticated cross-modal logic. Building upon this foundation, the Qwen series [2, 3, 28, 43] introduced critical architectural refinements—specifically the Naive Dynamic Resolution mechanism for processing visual inputs of arbitrary aspect ratios and the transition toward a unified multimodal backbone—significantly bolstering fine-grained

comprehension and complex scene analysis. Modern LVLMs typically employ a hierarchical training pipeline, progressing from an initial feature alignment phase that synchronizes visual tokens with linguistic semantics to extensive Supervised Fine-Tuning (SFT) on diverse, instruction-rich datasets. As exemplified by the scaling strategies of InternVL [12, 45, 61] and the query-based alignment approach in InstructBLIP [16], this trajectory has shifted the field from rudimentary image-text matching toward robust and generalizable visual intelligence.

Large Vision-Language Models for Spatial Reasoning Despite the significant strides in general image and video understanding, existing LVLMs still struggle with sophisticated spatial reasoning tasks that necessitate the interpretation of intricate geometric transformations and spatial configurations. To enhance the spatial intelligence of Large Vision-Language Models (LVLMs), existing research has diverged into architectural augmentation, large-scale dataset curation [17,50,51], and advanced training paradigms [32]. Architecturally, models such as Spatial-MLLM [46], VLM-3R [18], and 3DThinker [11] incorporate geometric priors via external 3D encoders, while SpatialBot [6] and VILASR [47] utilize external tools for depth estimation and ground perception. Simultaneously, the field has transitioned toward data-driven scaling; SpatialVLM [9] and SpatialRGPT [13] pioneered the synthesis of massive spatial VQA datasets, a trajectory further extended by VST [53] and SenseNova-SI [7], and they bolster the model's comprehension capabilities by scaling up both the volume and the diversity of the training data. To refine reasoning capabilities, multi-stage frameworks like SpatialLadder [26] and Cambrian-S [54] employ progressive SFT, whereas SpaceR [34] and MindCube [58] integrate cognitive maps with reinforcement learning to optimize reasoning traces. While these methodologies have introduced various specialized datasets, they often lack a granular analysis from a data-centric perspective on how specific construction strategies fundamentally enhance a model's spatial intelligence. To bridge this gap, we systematically investigate the principles of spatial data synthesis and introduce a more curated, large-scale dataset that specifically addresses previously overlooked perspective-taking tasks, thereby fostering a more holistic spatial reasoning framework.

3 Implementation Principles of OpenSpatial

To enable reproducible progress in spatial intelligence, we introduce **OpenSpatial**, an open-source data engine that generates high-quality spatial supervision from a unified *3D box–centric* representation (Fig. 2). Unlike static dataset releases, OpenSpatial exposes the full data-production pipeline and supports two complementary annotation modes: (i) *human annotation* for maximal accuracy, and (ii) *automated 3D lifting* for scalable expansion to in-the-wild web data and open-source assets. Both modes output the same canonical format– a scene mesh with object-aligned 3D boxes– which then drives consistent attribute extraction and QA synthesis. Built on this engine, we curate **OpenSpatial-3M**, a 3M-entry training suite covering five foundational capabilities (SM/SR/CP/MC/SAR in

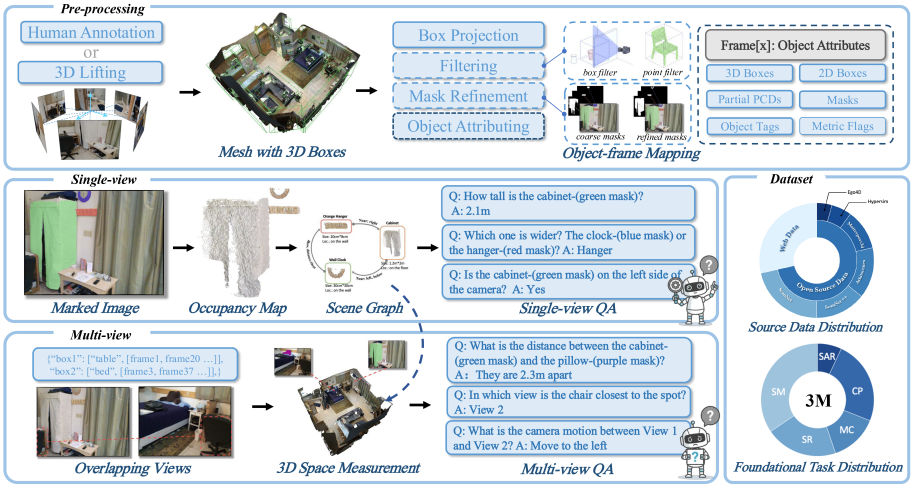


Fig. 2: Illustration of the data engine. The left panel of the figure illustrates the data processing and annotation pipeline, while the right panel presents the detailed statistics of the dataset, including source data distribution and task distribution.

Fig. 3), further divided into 19 diverse sub-tasks, forming a scalable and extensible foundation for general-purpose spatial understanding.

3.1 Data Pipeline

OpenSpatial turns raw multi-view images (or video keyframes) into spatial question-answer pairs through a staged pipeline (Fig. 2). It starts by producing scene-level 3D oriented bounding boxes (OBBs) for all visible objects, either via manual labeling or via an automated lifting procedure. These scene-level boxes are then converted into frame-level object attributes through projection, visibility filtering, and mask refinement, resulting in a consistent object–frame index (3D/2D boxes, masks, partial point clouds, tags, and metric flags). This shared representation supports two downstream annotation branches: single-view QA, generated from per-frame scene graphs with explicit visual anchors, and multi-view QA, generated by sampling overlapping views and using the viewpoint-invariant 3D boxes to enforce cross-view correspondence and consistency.

3D Box-Centric Design: Spatial understanding requires a stable 3D scene state: object position, size, orientation, and relations under viewpoint changes and occlusion. OpenSpatial uses Oriented Bounding Boxes (OBBs) as the core representation because they offer a compact, scalable middle ground between weak 2D labels and expensive dense 3D reconstructions. OBBs are world-coordinate and viewpoint-invariant, giving each object a single consistent 3D anchor across frames, which makes cross-view association and supervision straightforward. They also encode the minimum 3D structure needed for spatial reasoning (depth, extent, orientation), supporting metric, topological, and directional relations. Finally, OBBs act as a canonical anchor for downstream processing, enabling

consistent projection, visibility/occlusion filtering, and box-conditioned mask refinement to synchronize 3D geometry with precise 2D visual grounding. Concretely, each object is parameterized as an OBB $(x, y, z, x_l, y_l, z_l, r, p, y)$: (x, y, z) is the box center in world coordinates, (x_l, y_l, z_l) are the side length along the X/Y/Z axes, and (r, p, y) are Roll/Pitch/Yaw. All boxes are defined in a global world coordinate system with a Z-up convention, providing a consistent geometric anchor shared across frames and camera trajectories.

Scene-level Bounding Box Annotation: Given a posed video sequence, our first step is to obtain an oriented 3D bounding box for each object. OpenSpatial supports two complementary annotation modes. *Manual annotation*, following the EmbodiedScan protocol [44], leverages human effort to label objects in 3D and yields high-precision boxes, but is time-consuming and difficult to scale. To extend beyond curated datasets to web data and open-source assets without fine-grained labels, we additionally provide an *automated 3D lifting* pipeline. Starting from video keyframes or multi-view images, we perform per-view object recognition with Gemini [14] and instance mask extraction with SAM [23], then associate and merge instances in 3D space and fit a convex hull to produce the final oriented boxes. Qualitative visualizations are reported in the experiments (Fig. 4). After this step, each scene is represented as a reconstructed 3D mesh together with a set of object-aligned oriented 3D bounding boxes (Fig. 2).

Attribute-Centric Object–Frame Mapping. After obtaining a scene mesh with object-aligned 3D boxes, the next step is to convert these *scene-level* annotations into *frame-level* attributes that reliably support downstream scene-graph construction and QA synthesis across diverse task types. To this end, our engine extracts a comprehensive set of object attributes from the initial 3D bounding boxes (top-right part of Fig. 2). We first project each 3D box onto individual frames, then apply two filters to ensure data integrity: **(1)** boxes outside the current camera frustum are discarded; **(2)** to handle occlusions where an object projects into the frame but is heavily truncated, we perform depth-based validation. Specifically, pixels within the projected 2D box are back-projected into world coordinates using the depth map and camera intrinsics/extrinsics to form a local point cloud, and we compute the volumetric occupancy of these points inside the 3D box. Boxes with occupancy below a threshold are removed.

For boxes that pass filtering, the validated point-cloud pixels define coarse masks, which are further refined by SAM [23] to obtain fine-grained 2D instance masks that tightly align 3D geometry with visual appearance. Because automated labeling may contain duplicate objects with similar semantics, these masks serve as robust instance indicators and can be further converted into bounding boxes or keypoints as precise spatial prompts. Finally, we consolidate all extracted attributes– including masks, 2D/3D boxes, partial point clouds, and object tags– into a structured indexing system across frames. Each object is also assigned a *metric flag* indicating whether its box reflects real-world scale; if False, measurement-related QA generation is skipped to avoid noisy supervision. This object–frame indexing forms a unified and reliable foundation for all subsequent single-view and multi-view spatial understanding tasks.

Scene-Graph-Driven QA Synthesis for Diverse Spatial Supervision Given the object-frame indexing produced in the previous step, OpenSpatial synthesizes spatial QA in two complementary settings: single-view and multi-view– with an explicit emphasis on task diversity. The engine programmatically enumerates objects, their attributes, and inter-object relations (via scene graphs) to generate a balanced collection of questions spanning measurement, spatial relations, camera/viewpoint changes, multi-view consistency, and scene-level reasoning, thereby mitigating the narrow coverage that often leads to "spatial myopia". Concretely, we generate QA in the following two types:

- *Single-view annotation.* For each frame, we construct a structured scene graph from the indexed objects and their attributes (e.g., 2D/3D boxes, masks, tags). To prevent referential ambiguity when multiple instances share similar semantics, we render a visually marked image that highlights the queried object(s) as explicit anchors. Conditioned on the marked image and the scene graph, we generate diverse single-view QA that probes object–object and object–environment relationships, including relational queries (*e.g.*, left, right, front, behind), attribute comparisons (*e.g.*, size, relative depth), and context-dependent reasoning grounded in the current view.
- *Multi-view annotation.* Multi-view QA targets cross-view spatial reasoning, requiring consistent correspondence and geometry under viewpoint changes. The key challenge is selecting view pairs with sufficient overlap to enable inference while still introducing meaningful viewpoint variation. Our 3D box–centric representation provides a principled solution: because 3D boxes are anchored in a global world coordinate system, they serve as viewpoint-invariant references for linking instances across frames. We therefore sample view pairs that share a subset of 3D boxes, ensuring both contextual overlap and diversity. For each pair, we build a unified multi-view scene graph that merges object instances across views and generate cross-view QA– such as re-identification under perspective shifts, reasoning about camera changes, and consistency or measurement checks when permitted– encouraging the model to form a coherent 3D representation that generalizes across viewpoints.

3.2 Description of OpenSpatial-3M Dataset

Data Source: Following VST [53], we leverage the meticulously annotated 3D bounding boxes from EmbodiedScan [44] as the foundational data for our pipeline, which aggregates indoor scenes from ScanNet [15], Matterport3D [8], ARKitScenes [4], and SUN-RGBD [38]. Notably, we excluded SUN-RGBD due to its relatively lower annotation fidelity. To further enhance environmental diversity, we incorporated pre-processed data from ScanNet++ [57] and Hypersim [35] as supplementary sources. Additionally, we collected a set of in-the-wild web data to further broaden the diversity of our data sources. By including these real-world images/videos, we improve the dataset's coverage of various challenging scenarios, ensuring better generalization across different environments.

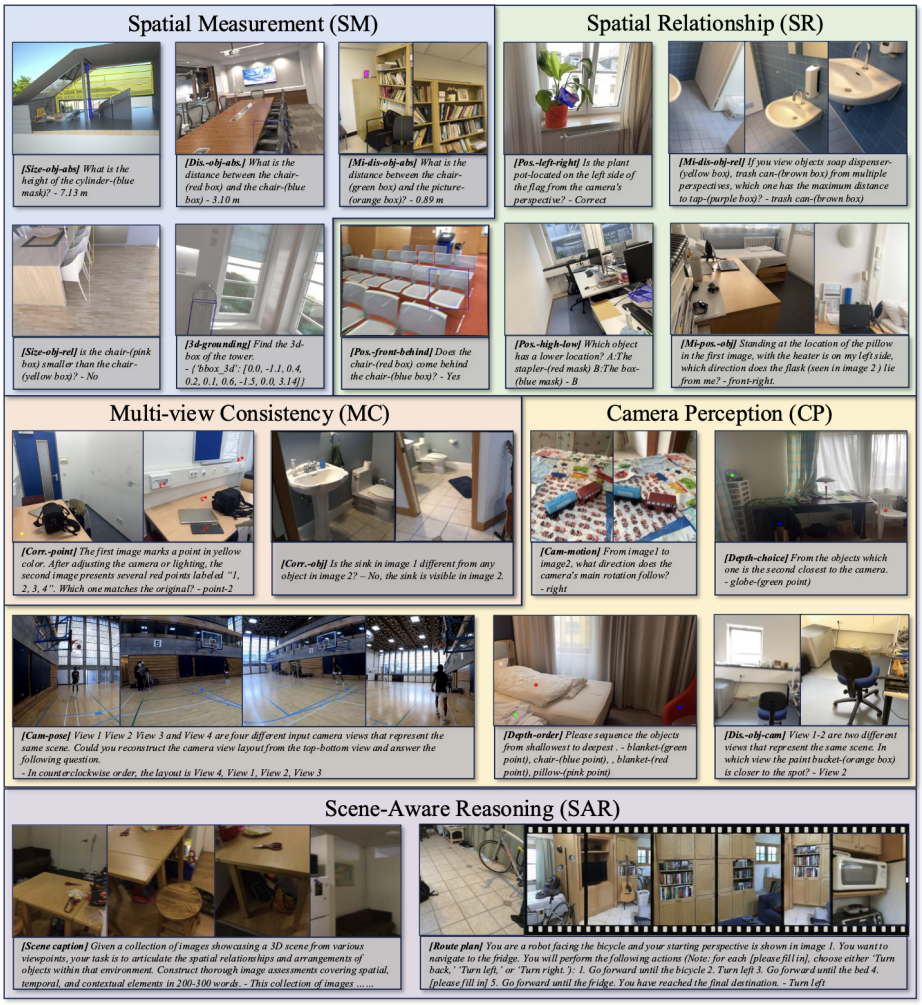


Fig. 3: Overview of the OpenSpatial dataset. OpenSpatial-3M comprises 3M high-quality samples for spatial understanding, encompassing five primary categories: Spatial Measurement (SM), Spatial Relationship (SR), Camera Perception (CP), Multi-view Consistency (MC), and Scene-Aware Reasoning (SAR). Representative cases for each category are curated and illustrated above. For brevity, the displayed QA pairs have been condensed; please refer to the Appendix for comprehensive details.

Task Taxonomy: We decompose spatial understanding into five core capabilities, as shown in Fig. 3, each further categorized into a diverse set of sub-tasks. A detailed characterization of these dimensions is provided in the following:

– Spatial Measurement (SM): Spatial measurement quantifies geometric metrics of objects and their configurations in a 3D coordinate system, estimating

absolute physical scales such as length, width, height, and distance, serving as foundational building blocks of spatial understanding.
- Spatial Relationship (SR): Spatial Relationship characterizes the 3D spatial arrangement between entities, focusing on relative localization and inter-object dependencies. It provides the qualitative framework necessary to describe the layout of a scene beyond individual object coordinates.
- Camera Perception (CP): Camera Perception helps model estimate camera poses and relative object-camera relationship. This sensor-aware intelligence serves as a foundational prior for implicit 3D reconstruction, enabling the translation of 2D observations into structured 3D coordinate systems.
- Multi-view Consistency (MC): Multi-view Consistency serves as the cornerstone for scene-level understanding. It establishes robust spatial correspondences across diverse viewpoints by identifying shared objects and contexts. By correlating physical entities from different perspectives, this capability requires the model to maintain a persistent 3D representation, ensuring spatial reasoning remains coherent despite camera pose changes.
- Scene-Aware Reasoning (SAR): Scene-Aware Reasoning focuses on holistic scene-level understanding and long-range spatial logic. It empowers the model with the ability to perceive spatial layouts and perform planning or navigation. By synthesizing the spatial configuration of obstacles and open spaces, the model develops the high-level reasoning required to determine traversability and optimal movement within a complex 3D environment.

4 Experiments

4.1 Implementation Details

Training Setting: We perform supervised fine-tuning (SFT) on several representative open-source VLMs to evaluate the effectiveness of our data engine. Adhering to the training protocol established in VST [53], each model is trained for a single epoch using 32 NVIDIA GPUs with a global batch size of 128, unless specified otherwise. We employ the AdamW optimizer for parameter updates, setting the base learning rate to 5×10^{-5}. We apply a decoupled learning rate strategy, setting the vision encoder's learning rate to a smaller 5×10^{-6}.

Training Data: In our experiments, we primarily utilize the OpenSpatial-3M dataset as our primary source. To further extend our coverage and incorporate more diverse spatial scenarios, we integrate the high-quality open-source dataset SenseNova-800K into our training mixture, supplementing spatial reasoning dimensions not fully addressed by our primary corpus. Furthermore, to effectively maintain the models' general multimodal capabilities while enhancing their spatial intelligence, we employ a balanced 1:1 ratio of general multi-modal data from LLaVA-OneVision [25] and spatial reasoning data from OpenSpatial-3M.

Table 1: Performance comparison of representative VLMs on spatial reasoning and general multimodal benchmarks. Dark blue and light blue backgrounds denote the best and second-best results, respectively. All models are evaluated under the same codebase to ensure a fair comparison. **Note**: 1)VSI is tested with 16 frames. 2)'±' denotes the performance gain or loss relative to the baseline model.

Models	3D-Avg	BLINK	AllAngles	ERQA	VSI	3DSR	MMSI	CV-3D	RealWorldQA	MMStar	MMB	MMU
Proprietary Models												
Gemini-2.5-Pro [14]	62.4	70.6	61.3	55.8	48.4	57.6	36.9	91.3	77.3	77.5	90.1	81.7
Open-source General Models												
InternVL2.5-4B [61]	47.3	50.8	45.1	41.0	28.3	44.0	28.5	76.4	64.2	58.5	78.5	50.0
InternVL2.5-8B [61]	51.6	54.9	48.9	40.8	39.3	51.0	28.6	79.9	69.4	62.6	82.3	53.3
InternVL3-2B [61]	47.9	52.8	48.6	36.3	30.3	46.4	25.9	77.3	65.5	61.5	77.7	45.9
InternVL3-8B [61]	53.2	55.7	50.5	40.5	38.7	52.7	30.9	86.0	70.6	68.5	82.0	57.7
Qwen2.5-VL-3B-Instruct [43]	45.6	49.0	42.8	40.8	32.0	45.2	25.0	64.8	65.2	56.6	77.2	48.4
Qwen2.5-VL-7B-Instruct [43]	50.0	55.3	50.1	41.0	36.0	49.0	26.5	73.8	68.1	65.3	82.3	55.2
Qwen3-VL-4B-Instruct [28]	56.2	62.6	49.1	40.2	53.6	52.5	28.0	92.3	71.4	67.5	82.8	57.7
Qwen3-VL-8B-Instruct [28]	56.7	66.1	49.5	40.1	55.6	52.8	28.1	90.8	70.7	70.1	83.8	60.2
Deepseek-VL2-27B-A4.5 [48]	-	54.3	46.2	40.8	-	50.1	29.0	79.1	70.2	62.3	81.3	51.3
MIMO-VL-7B-SFT [27]	54.6	59.7	52.9	41.0	37.5	56.1	29.3	86.9	73.5	71.1	80.9	65.9
Open-source SI Models												
SpaceR-7B [34]	50.8	54.3	49.8	40.5	44.4	47.5	29.4	76.3	64.2	63.9	81.2	54.3
SenseNova-SI-1.1-Qwen2.5-VL-7B [7]	51.8	55.0	47.7	39.0	55.2	46.7	32.5	77.4	60.5	58.9	80.4	48.0
SenseNova-SI-1.1-Qwen3-VL-8B [7]	55.5	56.4	47.3	41.8	58.8	51.8	34.6	89.2	63.8	65.0	80.6	59.8
VST-7B-SFT [53]	57.9	62.1	49.5	43.8	55.3	53.3	33.3	94.8	71.5	63.1	80.8	50.6
Ours												
OpenSpatial-InternVL2.5-8B	59.3 (+7.7)	63.5(+8.6)	58.3(+9.4)	43.0(+2.2)	56.7(+17.6)	52.0(+1.0)	38.7(+10.1)	93.8(+13.9)	68.6(-0.8)	57.4	78.5	48.0
OpenSpatial-InternVL3-8B	59.8(+6.6)	66.0(+10.3)	58.3 (+7.8)	44.5(+4.0)	57.4(+18.7)	53.5(+0.8)	38.6(+7.7)	93.7(+7.7)	66.6(-3.9)	62.4	82.0	51.2
OpenSpatial-Qwen2.5-VL-7B	59.5(+9.5)	65.9(+10.6)	58.4(+8.3)	41.8(+0.8)	56.7 (+20.7)	53.2(+4.2)	39.6(+13.1)	92.5(+18.4)	68.3(+0.2)	62.2	80.9	49.4
OpenSpatial-Qwen3-VL-8B	62.1(+5.4)	68.2(+2.1)	59.8(+10.3)	44.2(+4.1)	61.6(+6.0)	56.2(+3.4)	41.9(+13.8)	94.0(+3.2)	71.0(+0.3)	63.7	82.1	56.8

Evaluation: We evaluate the spatial reasoning capability of the VLMs across several representative benchmarks: BLINK [19], AllAngles [56], ERQA [39], VSI [52], 3D-SR [33], MMSI [55], CVBench-3D [42], and RealWorldQA [49]. For general multimodal capabilities, we extend our evaluation to the MMStar [10], MMBench [31], and MMMU [59] benchmarks. All models are assessed under identical settings using their respective native system prompts to ensure a fair comparison and minimize the impact of prompt variability.

4.2 Quality Evaluation

Main Results: As illustrated in Tab. 1, the OpenSpatial model series not only sets a new state-of-the-art on specialized spatial benchmarks but also reliably preserves its versatility on general-purpose benchmarks without catastrophic forgetting. Upon integrating the data synthesized by our engine, we observe a substantial performance surge across all spatial reasoning tasks compared to the respective baseline, with improvements typically ranging from **5.4 to 9.5 points**. These consistent gains across diverse metrics strongly suggest that OpenSpatial fosters a holistic understanding of 3D geometry rather than merely overfitting to specific spatial patterns. Particularly noteworthy are the results on BLINK, AllAngles, and MMSI, where our models achieve remarkable leaps **exceeding 10 points**. This significant margin allows us to outstrip existing spatial intelligence models by a wide gap, **serving as a strong testament to the unparalleled quality and diversity of our generated data**. From a model-centric perspective, we observe that Qwen3-VL-8B exhibits superior compatibility with our training data. This can be largely attributed to the integration of SigLIP [60] as the vision encoder, which significantly bolsters the model's visual perception and allows for a more nuanced spatial understanding of complex 3D scenes. How-

ever, we also identify a clear performance bottleneck in certain desktop-level and outdoor scenarios. This marginal improvement is primarily attributed to the current data distribution skew, and we intend to actively broaden our data coverage to encompass these complex environments in future iterations.

Table 2: Comparison with open-source datasets for spatial reasoning. '−' indicates the difference from the best result. MAD: Mean Deviation; Std. Dev.: Standard Deviation. **Note**: The subset of OpenSpatial is constructed independently and does not include any data from SenseNova-SI.

Data source	Data size	MAD	Std. Dev.	BLINK	AllAngles	ERQA	VSI	3DSR	MMSI	CV-3D	RealWorldQA
Cambrain-S [54]	590k	-6.0	5.4	54.1(-10.1)	48.6 (-5.3)	39.2(-3.0)	57.0	49.7(-4.4)	29.2(-7.1)	75.3(-17.9)	68.1(-2.6)
SenseNova-SI [7]	800k	-6.5	7.0	59.7 (-4.5)	47.5(-6.4)	38.0(-4.2)	55.0(-2.0)	48.9 (-5.2)	36.3	69.0(-24.2)	65.1(-5.6)
VST [53]	500k	-2.8	3.9	61.4(-2.8)	50.7(-3.2)	42.2	44.6(-12.4)	54.1	32.4(-3.9)	93.2	70.6(-0.1)
OpenSpatial(subset)	500k	-2.5	4.4	64.2	53.9	41.3(-0.9)	43.0(-14.0)	52.0(-2.1)	34.7(-1.6)	91.8(-1.4)	70.7

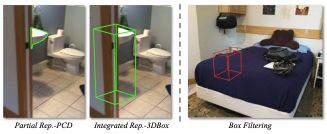
Comparative Study with Open-source Data: For a fair quality assessment, we benchmarked a scale-matched subset of OpenSpatial against open-source datasets. We specifically excluded SenseNova-SI-800k to focus on the intrinsic quality of our generated data. Following a unified training protocol based on Qwen2.5-VL, the comparative results are reported in Tab. 2. Statistical analysis reveals that OpenSpatial and VST both emerge as versatile and well-rounded datasets, achieving competitive stability and the narrowest mean gaps (−2.5 and −2.8, respectively) across all benchmarks. In contrast, while Cambrian-S and SenseNova-SI show larger overall fluctuations (Mean Deviation: −6.0 to −6.5 and Standard Deviation: 5.4 to 7.0), they exhibit specialized proficiency in niche domains, such as the VSI and MMSI benchmarks. This performance pattern highlights a clear data complementarity: while our engine prioritizes a broad and consistent spatial understanding. SenseNova-SI-800k exhibits localized strengths on specific benchmarks that complement our broad coverage, thus we integrate it in OpenSpatial to further bolster performance in specialized scenarios and achieve a more versatile spatial understanding

Module Reasonability: In this section, we evaluate the effectiveness of individual modules within our data engine. To ensure consistency, we reproduce the ablation datasets using ScanNet as the source, with each set maintained at approximately 200k samples. As shown in Tab. 3, shifting from a box-centric to a point-cloud-centric representation leads to varying degrees of performance degradation. This is further illustrated by qualitative examples on the right side of the table: partial point clouds fail to represent the complete geometry of objects, leading to inaccurate data generation, particularly for spatial measurement tasks. Furthermore, the filtering mechanism is crucial for box-centric design. Due to visual occlusion (exemplified by the red boxes), failing to filter such cases introduces hallucinations and significantly undermines model performance.

Quantitative Validation of 3D Lifting: We evaluate 3D lifting quality on 30 randomly sampled ScanNet scenes using EmbodiedScan annotations as reference (Table 4). After filtering, precision reaches 80.2%. We prioritize precision over

Table 3: Rationality Analysis of Module Design.

Setting: data size(200k)	BLINK	AllAngles	VSI	CV-3D
Qwen2.5-VL-7B	55.3	50.1	36.0	73.8
+Point Cloud Centric	57.2	49.7	37.2	83.7
+3D-Box Centric	60.3	53.2	41.7	89.9
+3D-Box Centric (no filter)	56.6	47.0	32.1	78.2



recall to ensure downstream VLM training reliability. Table 5 ablates all QA data from 3D lifting on Qwen2.5-VL, showing consistent gains.

Table 4: 3D Lifting Quality.

Config. (ScanNet)	Precision. (%)	Recall (%)
Before Filter	76.3	**67.9**
After Filter	**80.2**	67.5

Table 5: Ablation: Lifting Data.

Config.	BLINK	AllAngles	ERQA	VSI
w/o	64.3	56.2	**41.8**	55.2
w	**65.9**	**58.4**	**41.8**	**56.7**

Table 6: Ablation on data scaling. **Color Map**: worst to best.

Data Size	3D-Avg	BLINK	AllAngles	ERQA	VSI	3DSR	MMSI	CV-3D	RealWorldQA
20%	57.6	64.9	55.0	41.0	51.8	52.0	34.7	91.8	69.5
40%	58.5	66.2	56.8	42.1	52.4	52.9	35.8	91.0	71.1
60%	58.6	66.9	56.7	42.0	53.2	52.3	36.0	92.1	70.7
80%	59.2	66.2	59.7	41.2	53.5	53.0	37.6	91.5	70.5
Full	59.7	65.9	58.4	41.8	56.7	54.3	39.6	92.5	68.3

4.3 Scalability Evaluation

Data Scaling: We perform category-wise downsampling on both spatial reasoning data and general data, utilizing Qwen2.5-VL as our base model. The results, as summarized in Tab. 6, reveal that while individual benchmarks may not exhibit a strictly monotonic increase with data scaling, the 3D-Avg metric shows a consistent positive correlation. This trend suggests that **increasing data volume systematically enhances the model's comprehensive spatial reasoning capabilities**. However, we also observe that the rate of performance gain diminishes as the data scale grows, indicating that **further improvements in spatial intelligence require exponentially larger datasets**.

Data Source Scaling: Existing 3D datasets are constrained by a limited number of scenes and a heavy bias towards indoor environments. To further scale the data volume and enhance scene diversity, we develop a robust 3D lifting pipeline capable of reconstructing 3D scenes from in-the-wild video data, while simultaneously generating comprehensive annotations including semantic tags, masks, and 3D bounding boxes. Fig. 4 illustrates our annotation results on an uncurated outdoor video. As shown, our pipeline not only recovers the scene geometry (point clouds) with high fidelity but also produces accurate tags and boxes, effectively meeting the stringent requirements for 3D spatial understanding data production. Leveraging this pipeline, we are able to significantly scale

Table 7: Ablation on model scaling. Dark blue : best result. Light blue : second best.

Model Size	3D-Avg	BLINK	AllAngles	ERQA	VSI	3DSR	MMSI	CV-3D	RealWorldQA
OpenSpatial-Qwen2.5-VL-3B	56.1	61.0	53.0	40.0	55.2	47.8	32.0	94.0	66.0
OpenSpatial-Qwen2.5-VL-7B	59.7	65.9	58.4	41.8	56.7	54.3	39.6	92.5	68.3
OpenSpatial-Qwen2.5-VL-32B	61.3	68.2	63.3	44.0	57.3	55.3	39.8	93.4	69.3

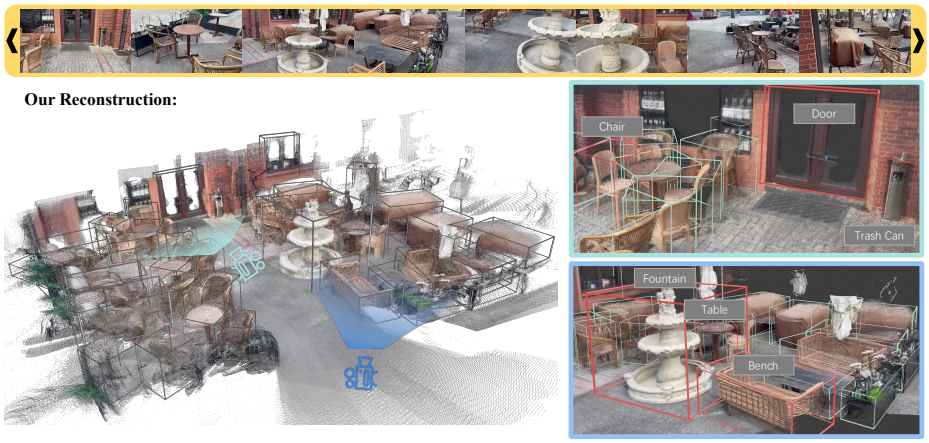


Fig. 4: Visualization of 3D lifting results from in-the-wild outdoor web data.

our dataset to unprecedented diversity and volume. Tab. 8 demonstrates the effectiveness of the spatial understanding data produced solely from web-sourced data via our 3D lifting pipeline.

Model Scaling: Investigating model size is equally crucial for spatial understanding, as models with larger parameter scales possess greater representative capacity to internalize and organize complex spatial knowledge. As illustrated in Tab. 6, we evaluate the performance of Qwen2.5-VL across various scales, including 3B, 7B, and 32B parameters, under identical data configurations. It is evident that nearly all evaluation metrics exhibit a consistent upward trend as the model size increases, highlighting the positive impact of model capacity on spatial reasoning tasks. These results further validate the significance of a scalable data engine. By consistently translating increased data volume into performance gains, our engine demonstrates its ability to provide the high-quality supervision necessary for advancing the spatial intelligence of VLMs.

4.4 Diversity Evaluation

To provide a granular understanding of how task diversity steers the evolution of spatial reasoning, we conducted a dual-faceted evaluation as illustrated in Fig. 5. **Task-Specific Contributions and Complementarity**: As depicted in the left heatmap of Fig. 5, individual tasks exhibit heterogeneous performance footprints across diverse benchmarks. For instance, tasks focused on Spatial Measurement (SM) yield substantial gains in metric-heavy evaluations, whereas Camera Perception (CP) tasks predominantly bolster the model's ability to decode extrinsic

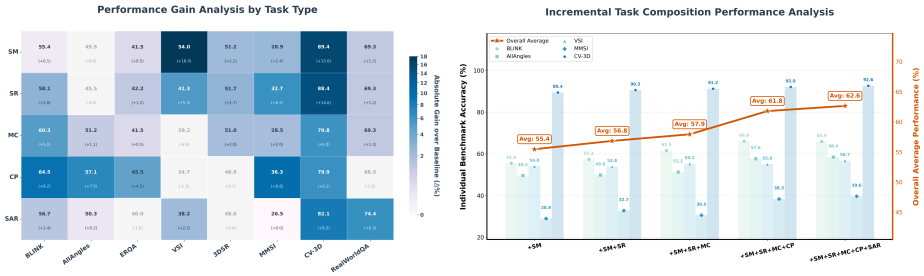


Fig. 5: Impact of diverse tasks on spatial intelligence. Best zoomed in.

parameters and ego-motion, leading to significant improvements in benchmarks requiring precise viewpoint awareness. This divergence underscores that our task library is not redundant but rather possesses a strong complementary architecture, where each task targets a unique dimension of spatial cognition.

Incremental Synergy and Compositional Trends: Moving to the right panel of Fig. 5, we investigate the cumulative impact of incremental task integration. The results reveal a compelling compositional synergy: with the increment of the task diversity, the model's comprehensive capabilities scale accordingly. We observe occasional, localized performance plateaus or slight "dips" upon the introduction of certain task combinations; these are likely attributed to data distribution shifts or the interference of gradient directions during multi-task optimization. However, the overarching trajectory remains unmistakably positive. The orange "Overall Average" curve maintains a steady and robust ascent, signifying that increased task diversity effectively mitigates the limitations of single-task learning and fosters a more holistic and resilient spatial intelligence.

Table 8: Effectiveness of data source scaling. We independently validate the effectiveness of our 3D lifting data.

Setting: 200k	BLINK	3DSR	CV-3D	RealWorldQA
Qwen2.5-VL-7B	55.3	49.0	73.8	68.1
+3D Lifting	62.2	54.3	87.9	71.8

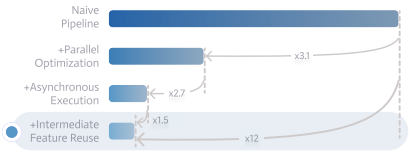


Fig. 6: Efficiency breakdown.

4.5 Efficiency Evaluation

To enhance the efficiency of our data production pipeline, we implemented a series of systematic optimizations. First, parallel processing was applied across most components to maximize throughput. Second, we leveraged message queues to enable asynchronous execution between consecutive stages; this pipelining strategy allows the current stage to perform inference on a batch while the preceding stage simultaneously processes the next. Finally, for tasks sharing common intermediate features, we developed an automatic reuse mechanism to avoid redundant computations and further streamline the workflow. The performance gains resulting from these efficiency optimizations are illustrated in Fig. 6.

5 Conclusion

In this work, we introduce OpenSpatial, a principled data engine that shifts the focus from static datasets to a transparent, scalable infrastructure for spatial intelligence. By establishing a 3D box-centric paradigm, the OpenSpatial engine serves as a vital bridge between sparse 2D visual cues and intrinsic 3D metric properties, providing a viewpoint-invariant foundation that was previously confined to closed-source pipelines. This engine not only enables the synthesis of our OpenSpatial-3M dataset—which achieves state-of-the-art performance across diverse MLLM architectures—but more importantly, it establishes a sustainable foundation for producing diverse spatial data across multiple sources organized into five hierarchical task categories, allowing for continuous expansion and refinement of spatial understanding. By open-sourcing OpenSpatial, we aim to democratize the creation of high-quality data, serving as a robust cornerstone for the community to advance embodied AI and robotics.

Acknowledgements

This work was primarily supported by Joy Future Academy and JD.com for research funding and computing resources. It was also supported in part by the Hong Kong Research Grant Council - General Research Fund (Grant No. 17202422, 17212923, 17215025), Theme-based Research Scheme (Grant No. T45-701/22-R), and Strategic Topics Grant (Grant No. STG3/E-605/25-N). Part of the described research work was conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization. Text Reading, and Beyond **2**(1), 1 (2023)
3. 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)
4. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021)
5. Brohan, A., Brown, N., Carbajal, J., Chebotar, Y., Dabis, J., Finn, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Hsu, J., et al.: Rt-1: Robotics transformer for real-world control at scale. arXiv preprint arXiv:2212.06817 (2022)

6. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 9490–9498. IEEE (2025)
7. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., et al.: Scaling spatial intelligence with multimodal foundation models. arXiv preprint arXiv:2511.13719 (2025)
8. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
9. 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 (CVPR). pp. 14455–14465 (June 2024)
10. Chen, L., Li, J., Dong, X., Zhang, P., Zang, Y., Chen, Z., Duan, H., Wang, J., Qiao, Y., Lin, D., et al.: Are we on the right way for evaluating large vision-language models? Advances in Neural Information Processing Systems **37**, 27056–27087 (2024)
11. Chen, Z., Zhang, M., Yu, X., Luo, X., Sun, M., Pan, Z., Feng, Y., Pei, P., Cai, X., Huang, R.: Think with 3d: Geometric imagination grounded spatial reasoning from limited views. arXiv preprint arXiv:2510.18632 (2025)
12. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: 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. pp. 24185–24198 (2024)
13. Cheng, A.C., Yin, H., Fu, Y., Guo, Q., Yang, R., Kautz, J., Wang, X., Liu, S.: Spatialrgpt: Grounded spatial reasoning in vision-language models. Advances in Neural Information Processing Systems **37**, 135062–135093 (2024)
14. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
15. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
16. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
17. Daxberger, E., Wenzel, N., Griffiths, D., Gang, H., Lazarow, J., Kohavi, G., Kang, K., Eichner, M., Yang, Y., Dehghan, A., et al.: Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7395–7408 (2025)
18. Fan, Z., Zhang, J., Li, R., Zhang, J., Chen, R., Hu, H., Wang, K., Qu, H., Wang, D., Yan, Z., et al.: Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. arXiv preprint arXiv:2505.20279 (2025)
19. Fu, X., Hu, Y., Li, B., Feng, Y., Wang, H., Lin, X., Roth, D., Smith, N.A., Ma, W.C., Krishna, R.: Blink: Multimodal large language models can see but not perceive. In: European Conference on Computer Vision. pp. 148–166. Springer (2024)
20. Goyal, A., Xu, J., Guo, Y., Blukis, V., Chao, Y.W., Fox, D.: Rvt: Robotic view transformer for 3d object manipulation. In: Conference on Robot Learning. pp. 694–710. PMLR (2023)

21. Guo, D., Wu, F., Zhu, F., Leng, F., Shi, G., Chen, H., Fan, H., Wang, J., Jiang, J., Wang, J., et al.: Seed1. 5-vl technical report. arXiv preprint arXiv:2505.07062 (2025)
22. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
23. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
24. Lee, S., Park, S., Kim, H.: Dynscene: Scalable generation of dynamic robotic manipulation scenes for embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12166–12175 (2025)
25. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
26. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Spatialladder: Progressive training for spatial reasoning in vision-language models. arXiv preprint arXiv:2510.08531 (2025)
27. Li, J., Chen, J., Qu, Y., Xu, S., Lin, Z., Zhu, J., Xu, B., Tan, W., Fu, P., Ju, J., et al.: Xiaomi mimo-vl-miloco technical report. arXiv preprint arXiv:2512.17436 (2025)
28. Li, M., Zhang, Y., Long, D., Chen, K., Song, S., Bai, S., Yang, Z., Xie, P., Yang, A., Liu, D., et al.: Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. arXiv preprint arXiv:2601.04720 (2026)
29. Lin, X., Lin, T., Huang, L., Xie, H., Su, Z.: Bip3d: Bridging 2d images and 3d perception for embodied intelligence. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9007–9016 (2025)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
31. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
32. Long, Y., Yang, Y., Wei, H., Chen, W., Zhang, T., Liu, C., Jiang, K., Chen, J., Tang, K., Wen, B., et al.: Spatialreward: Bridging the perception gap in online rl for image editing via explicit spatial reasoning. arXiv preprint arXiv:2602.07458 (2026)
33. Ma, W., Chen, H., Zhang, G., Chou, Y.C., Chen, J., de Melo, C., Yuille, A.: 3dsr-bench: A comprehensive 3d spatial reasoning benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6924–6934 (2025)
34. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv preprint arXiv:2504.01805 (2025)
35. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
36. Seed, B.: Seed1. 8 model card: Towards generalized real-world agency. Tech. rep., Technical report (model card), December 2025. URL https://lf3-static ... (2025)
37. Shridhar, M., Manuelli, L., Fox, D.: Perceiver-actor: A multi-task transformer for robotic manipulation. In: Conference on Robot Learning. pp. 785–799. PMLR (2023)

38. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
39. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
40. Team, K., Bai, Y., Bao, Y., Chen, G., Chen, J., Chen, N., Chen, R., Chen, Y., Chen, Y., Chen, Y., et al.: Kimi k2: Open agentic intelligence. arXiv preprint arXiv:2507.20534 (2025)
41. Team, K., Du, A., Yin, B., Xing, B., Qu, B., Wang, B., Chen, C., Zhang, C., Du, C., Wei, C., et al.: Kimi-vl technical report. arXiv preprint arXiv:2504.07491 (2025)
42. Tong, P., Brown, E., Wu, P., Woo, S., Iyer, A.J.V., Akula, S.C., Yang, S., Yang, J., Middepogu, M., Wang, Z., et al.: Cambrian-1: A fully open, vision-centric exploration of multimodal llms. Advances in Neural Information Processing Systems **37**, 87310–87356 (2024)
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
44. Wang, T., Mao, X., Zhu, C., Xu, R., Lyu, R., Li, P., Chen, X., Zhang, W., Chen, K., Xue, T., et al.: Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19757–19767 (2024)
45. 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)
46. 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)
47. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv preprint arXiv:2506.09965 (2025)
48. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
49. xAI: Grok-1.5v preview. https://x.ai/news/grok-1.5v (2024)
50. Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T.J., Cheng, Z., Shin, D., Lei, F., et al.: Osworld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems **37**, 52040–52094 (2024)
51. Xu, R., Wang, W., Tang, H., Chen, X., Wang, X., Chu, F.J., Lin, D., Feiszli, M., Liang, K.J.: Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. arXiv preprint arXiv:2505.17015 (2025)
52. 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)
53. 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)
54. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)

55. Yang, S., Xu, R., Xie, Y., Yang, S., Li, M., Lin, J., Zhu, C., Chen, X., Duan, H., Yue, X., et al.: Mmsi-bench: A benchmark for multi-image spatial intelligence. arXiv preprint arXiv:2505.23764 (2025)
56. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
57. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12–22 (2023)
58. 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)
59. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
60. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
61. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)