Scaling Spatial Intelligence with Multimodal Foundation Models

Zhongang Cai*,[1], Ruisi Wang*,[1], Chenyang Gu*,[1], Fanyi Pu*,[1,2], Junxiang Xu*,[1], Yubo Wang*,[1], Wanqi Yin*,[1], Zhitao Yang*,[1], Chen Wei*,[1], Qingping Sun*,[1], Tongxi Zhou*,[1], Jiaqi Li*,[1], Hui En Pang*,[2], Oscar Qian*,[1,2], Yukun Wei[1], Zhiqian Lin[1], Xuanke Shi[1], Kewang Deng[1], Xiaoyang Han[1], Zukai Chen[1], Xiangyu Fan[1], Hanming Deng[1], Lewei Lu[1], Liang Pan[1], Bo Li[2], Ziwei Liu✉,[2], Quan Wang✉,[1], Dahua Lin✉,[1], Lei Yang*,✉,[1]
[1]SenseTime Research, [2]Nanyang Technological University
* Core Contributors, ✉ Corresponding Authors

https://github.com/OpenSenseNova/SenseNova-SI

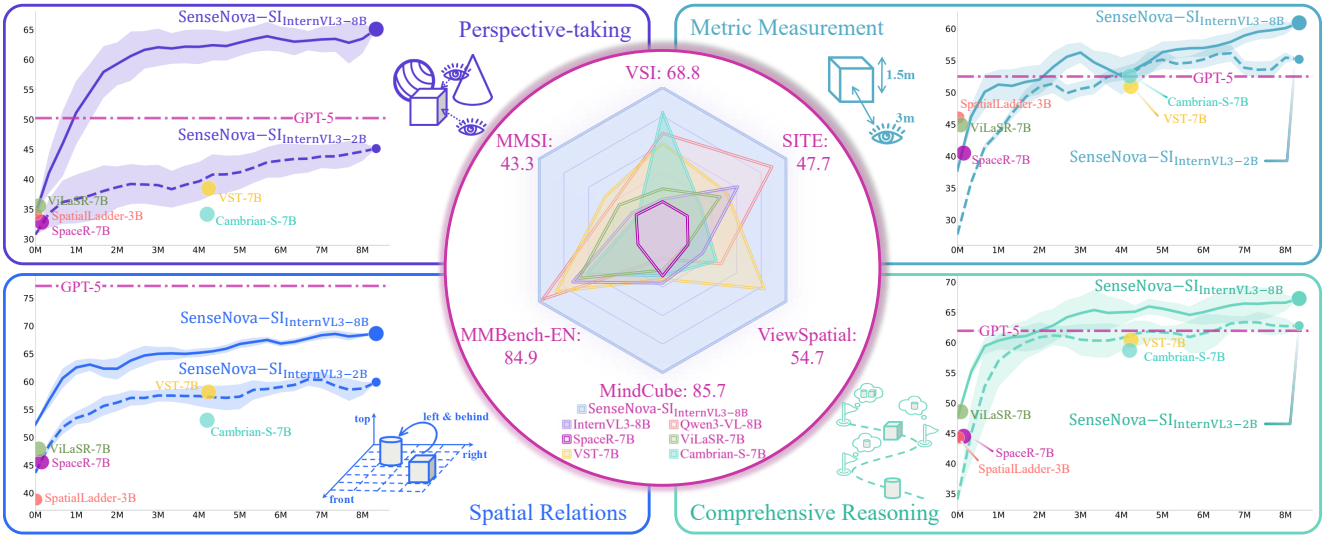

Figure 1. Guided by taxonomy of spatial intelligence [7], we scaled spatial data to construct **SenseNova-SI-8M**, which we leverage to investigate the impact of data scaling on cultivating spatial capabilities in various MLLMs. The four subfigures at the corners elaborate **SenseNova-SI**'s performance on four core spatial capabilities (*i.e.*, Perspective-taking, Spatial Relations, Metric Measurement, and Comprehensive Reasoning). Through data scaling, SenseNova-SI surpasses open-source models and even outperforms GPT-5 in specific spatial abilities, such as Perspective-taking. The lines denote the average performance across benchmark subtasks within each capability, while the shaded regions (confidence bands) represent ± 0.5 standard deviation. At center, we show **SenseNova-SI** achieves state-of-the-art (SoTA) results on five recent spatial intelligence benchmarks (VSI, MMSI, MindCube, ViewSpatial, and SITE) while maintaining strong performance on a general multimodal benchmark (MMBench-En).

Abstract

*Despite remarkable progress, multimodal foundation models still exhibit surprising deficiencies in spatial intelligence. In this work, we explore scaling up multimodal foundation models to cultivate spatial intelligence within the **SenseNova-SI** family[1], built upon established multimodal foundations including visual understanding models (*i.e.*, Qwen3-VL and InternVL3) and unified understanding and generation models (*i.e.*, Bagel). We take a principled approach to constructing high-performing and robust spatial intelligence by systematically curating SenseNova-SI-8M: eight million diverse data samples under a rigorous taxonomy of spatial capabilities. SenseNova-SI demonstrates unprecedented performance across a broad range of spatial intelligence benchmarks: 68.8% on VSI-Bench, 43.3% on MMSI, 85.7% on MindCube, 54.7% on ViewSpatial, 47.7% on SITE, 63.9% on BLINK, 55.5% on 3DSR, and 72.0% on EmbSpatial, while maintaining strong general*

[1]This paper is based on the v1.1 version of SenseNova-SI.

*multimodal understanding (*e.g.*, 84.9% on MMBench-En). More importantly, we analyze the impact of data scaling, discuss early signs of emergent generalization capabilities enabled by diverse data training, analyze the risk of overfitting and language shortcuts, present a preliminary study on spatial chain-of-thought reasoning, and validate the potential downstream application. All newly trained multimodal foundation models are publicly released.*

1. Introduction

In recent years, multimodal foundation models [3, 15, 74] have achieved groundbreaking progress across a wide spectrum of tasks. However, it has become evident that even the most advanced models still struggle with spatial intelligence: the ability to understand, reason about, and act within three-dimensional space, which is fundamental to embodied AGI that can perceive, adapt to, and interact with the physical world. Interestingly, such tasks are often considered trivial for humans [7]. One of the key limitations lies in the scarcity and imbalance of spatially grounded data. While recent efforts have introduced a surge of large-scale datasets targeting various facets of spatial reasoning, these resources remain fragmented and heterogeneous in scope and quality. Consequently, the community is still in the early stages of understanding how multimodal foundation models acquire and develop spatial intelligence, and what strategies are effective in fostering this capability.

In this work, we aim to provide timely insights into cultivating spatial intelligence within state-of-the-art multimodal foundation models by leveraging their powerful generalist backbones and scaling up diverse data collections. Our study investigates the data scaling laws of spatial intelligence through extensive experiments on the widely adopted InternVL3 multimodal foundation model family [74], and further extends the analysis to Qwen3-VL [3] as well as Bagel [15], a unified understanding and generation model. We envision the resulting models, denoted by the **SenseNova-SI** prefix, as open research platforms to advance studies in spatial intelligence. To preserve compatibility with existing research pipelines, we deliberately avoid altering the original architectures of the base models. Instead, we adopt a data-centric approach, emphasizing the role of data scaling and training strategies as the primary drivers of spatial understanding capability. Our systematic collection and synthesis of spatial data are guided by a principled taxonomy of fundamental spatial intelligence capabilities [7], resulting in *eight million* samples (named SenseNova-SI-8M) spanning five key domains: Metric Measurement (MM), Spatial Relations (SR), Mental Reconstruction (MR), Perspective-taking (PT), and Comprehensive Reasoning (CR). We analyze a diverse collection of public datasets for spatial intelligence, followed by strategic further scaling that places a special focus on perspective-taking, an underrepresented capability that is critical to spatial intelligence, while isolated from general multimodal capabilities [33].

We evaluate the SenseNova-SI foundation models across a broad suite of benchmarks, including VSI-Bench [64], MMSI [67], MindCube [70], ViewSpatial [31], SITE [57], BLINK [20], 3DSR [39], and EmbSpatial [16], following continued training on our comprehensive spatial intelligence data collection. The models achieve state-of-the-art performance among open-source models of comparable sizes, with the best performance achieving 68.8% on VSI-Bench, 43.3% on MMSI, 85.7% on MindCube, 54.7% on ViewSpatial, 47.7% on SITE, 63.9% on BLINK, 55.5% on 3DSR, and 72.0% on EmbSpatial, while retaining their original strengths on general multimodal understanding benchmarks such as MMBench-En (84.9%). Our analysis reveals several key findings: (1) Scaling law of spatial intelligence. We systematically investigate how spatial intelligence scales under mixed data regimes. Our analysis reveals distinct scaling behaviors across spatial capabilities and model sizes, and suggests that the observed saturation trends may signal that future advances require paradigm shifts built upon and beyond SenseNova-SI. (2) Emergent generalization through diverse data. We report surprising findings that point to early signs of emergent spatial intelligence: models trained on one set of spatial tasks exhibit nontrivial transfer to seemingly unrelated tasks, and demonstrate extrapolation to longer spatial contexts beyond the training distribution. (3) Robustness against overfitting and shortcuts. Through controlled experiments and circular test designs, we rigorously validate that SenseNova-SI genuinely acquires spatial capabilities rather than exploiting memorization, annotation biases, or unintended shortcuts in the training data. (4) Spatial chain-of-thought (CoT) may not be effective. We construct and evaluate three representative text CoT schemes and reinforcement learning, but find that they cannot reliably improve spatial reasoning beyond what is achieved through simple QA-style data scaling. These results suggest that extending text-based CoT paradigms to spatial intelligence is non-trivial and may require fundamentally different reasoning mechanisms. (5) Downstream task validation. To assess the practical utility of SenseNova-SI, we apply SenseNova-SI to robotic manipulation tasks without any finetuning, and achieve notable performance improvements on EmbodiedBench [65], demonstrating the potential of SenseNova-SI as a foundation for embodied AI (elaborated in Sec. H).

In summary, we introduce the SenseNova-SI series of multimodal foundation models, which achieve new state-of-the-art performance across major spatial intelligence benchmarks. Our study further validates that data scaling governs the progression of spatial intelligence. We envision

SenseNova-SI as a strong, robust baseline that future research can build upon to drive deeper advances.

2. Related Works

2.1. Multimodal Foundational Models

Recent studies [7, 33, 71] reveal that while models like GPT-5 demonstrate strong planar reasoning capabilities, they still lag significantly behind humans in Spatial Intelligence (SI). Furthermore, EASI [7] shows that the performance gap between open-source and closed-source models on SI tasks is relatively small. These findings motivate us to enhance the spatial intelligence of widely used open-source models (*e.g.*, QwenVL series [2–4, 54] and InternVL series [11, 56, 74]). This not only enables fairer comparisons among models of similar scale but also facilitates the community's direct use of our models for downstream tasks, (*e.g.*, VLA [28, 65, 75]), with minimal substitution costs.

2.2. Multimodal Models for Spatial Intelligence

Efforts to enhance spatial intelligence in multimodal models primarily follow two approaches: *leveraging 3D experts* or *curating spatial-specific datasets*. As spatial intelligence is inherently linked to 3D vision, an intuition is to employ 3D expert encoders that infer key 3D attributes from images [12, 51, 59]. Spatial-MLLM [59] incorporates VGGT [53] as an input-level encoder to capture 3D information, while VLM-3R [17] integrates 3D information using combined geometry and camera-view tokens. Recently, 3DThinker [12] aligns model-generated 3D features with VGGT-derived supervision at the output level. Conversely, some studies [9, 13, 60, 63] inject visual-spatial knowledge through dataset curation and training paradigm. SpatialVLM [9] pioneered this direction by synthesizing 2B VQA samples focused on two-object spatial relationships. SpaceR [43] uses RL for spatial reasoning, while MindCube [70] explores SFT and RL using QA and two types of cognitive maps. SpatialLadder [32] constructs a dataset with 26K samples and introduces a three-stage progressive training strategy. Concurrently, VST [66] adopts a two-phase training approach, using 4.1M samples for SFT on spatial perception and 135K samples for RL on spatial reasoning. Cambrian-S [68] develops VSI-590K dataset and employs a four-stage training framework to progressively enhance spatial video understanding. In this work, we systematically scale datasets targeting core spatial capabilities [7], addressing key gaps in existing datasets, particularly the previously overlooked perspective-taking tasks.

3. Data

The limitations in spatial intelligence mainly stem from high-quality, diverse data scarcity. In this work, we strategically scale data to expand coverage toward holistic spatial intelligence, rather than merely increasing data volume.

3.1. Task Taxonomy

We adopt a principled approach, following the EASI [7] protocol to decompose spatial intelligence into key fundamental capabilities. We focus on five capabilities that are closely aligned with real-world scenarios. For each, we analyze the core cognitive operation and derive tasks to ensure comprehensive coverage. Figure 2 illustrates the dataset constructed under this taxonomy.

Metric Measurement (MM). MM involves a basic understanding of the physical scale and typical object sizes. We include distances estimation between the camera and objects and pairs of objects, and size estimation across scales from individual objects to entire scenes.

Spatial Relations (SR). We define SR as the ability to impose and reason within a 3D coordinate system. In egocentric, local level of view, it unfolds into front–back, left–right, and up–down relations between subjects. In global, scene level, these relations extend to near–far and relative scale (large–small) comparisons.

Mental Reconstruction (MR). MR infers 3D object structure from limited 2D observations. We adopt a diagnostic task, which identifies which side of an object is visible. This requires the integration of sparse 2D cues to infer 3D geometry and align views in a canonical object-centric frame.

Perspective-taking (PT). PT addresses reasoning with changing camera viewpoints. We construct PT tasks in a progressively more challenging hierarchy:

- *View Correspondence* establishes correspondences of points or objects across views, recognizing entities under changes in viewpoint, scale, and occlusion.
- *Camera Motion Reasoning* infers relative camera motion between views, linking observational changes to camera transformations in the 3D space.
- *Allocentric Transformation* simulates viewpoint shifts and express spatial relations across coordinate systems, including camera, object-target, and self-oriented views.

This layered design ensures that PT goes beyond pattern matching across images, encouraging the model to build internal representations of how observations transform with viewpoint changes.

Comprehensive Reasoning (CR). CR tasks involve coordinating multiple spatial capabilities with extended memory and multi-step reasoning. Such data is scarce and often limited to simple scenarios. As these tasks lie beyond our main goal of scaling spatial QA and core spatial capabilities, we reuse existing datasets as a lightweight complement.

3.2. Data Sources

General QA. We collect a set of open-source general-purpose QA datasets for 2D image understanding. Specif-

Figure 2. **SenseNova-SI-8M** reorganizes 4M open-source data and scales 4.5M additional data, according to fundamental spatial capbilities [7]. It covers general visual understanding (Non-SI), 2D grounding, and five core spatial abilities: Metric Measurement (MM), Spatial Relationship (SR), Perspective-Taking (PT), Mental Reconstruction (MR), and Comprehensive Reasoning (CR). Notably, SenseNova-SI-8M addresses the previously overlooked PT tasks. How data from each source is mapped to the core spatial capabilities is illustrated at the top (with a scale in the upper-right corner indicating the number of QA pairs), while representative data samples are organized by core capability. The "Hugging Face" symbol indicates community datasets. The rest are curated for further scaling.

ically, we use VSR [34], SPEC [44], GQA [25], VQA [1], and IconQA [38], resulting in about 0.6M QA pairs.

Community Datasets on Spatial Intelligence. Among existing open-source resources, we identify several datasets that focus on spatial reasoning, including Open3D-VQA [73], CLEVR-series [27], REL3D [22], SAT [45], GRiD-3D [30], MultiSpa [63], MindCube [70], ViCA [18], VLM-3R [17], and VSI-590K [68]. We incorporate all of

these datasets, yielding in total about 3.3M QA pairs.

Further Scaling on Spatial Intelligence. Building on these open-source data, we find gaps in task coverage and data imbalance. MM and SR dominate the data, while PT and MR remain underrepresented. For point, object, scene level correspondence, only MultiSpa provides point level QAs. Camera motion is also mostly limited to MultiSpa. Allocentric viewpoint transformation, especially object-centric and hypothetical views, is largely unexplored, as real-world QA labels are scarce. Tasks such as object reconstruction remain unaddressed.

To address these gaps, we leverage richly annotated, scene-diverse 3D datasets, including MessyTable [6], ScanNet [14], ScanNet++ [69], SUN RGB-D [47], CA-1M [29], Ego-Exo4D [23], and Matterport3D [8], to generate large-scale, accurate and task-balanced QA pairs. This scaling process contributes 4.5M data, increasing the overall corpus size to 8.5M QA pairs.

4. Training

We adopt three multimodal foundation models in this study. **Qwen3-VL [3]** is the most capable multimodal model in the Qwen series to date. It adopts a strategy to scale from language foundation, that expand a strong LLM foundation to handle vision or audio modalities. **InternVL-3** [74] is natively multimodal, training vision and language jointly from scratch, thus enables stronger cross-modal alignment, more efficient scaling, and improved visual–language reasoning. **Bagel** [15] represents a new paradigm of unified understanding and generation. We include it in our study to examine whether such unified architectures can acquire strong spatial understanding capabilities.

Training Scheme. Each foundation model is trained for one epoch on the same dataset using 128 GPUs with batch size 2048, taking approximately three days. We employ AdamW [37] with a learning rate of 5×10^{-6} for all training runs. Maximum 16 frames are sampled for video data.

5. Experiments

5.1. Evaluation Benchmarks.

To assess SenseNova-SI under a broad range of scenarios, we select five newly released benchmarks for a complementary coverage of spatial intelligence. **VSI-Bench** [64] targets *video-based* visual-spatial reasoning, evaluating the ability to perceive and understand the 3D layout of indoor scenes over an extended context. We uniformly sample 32 frames from each video during testing. **MMSI-Bench** [67] extends spatial reasoning to *multi-image* settings, requiring models to integrate spatial cues across multiple views. MMSI is notably challenging: each question is manually crafted by researchers rather than mass-generated through templates. **MindCube** [70] targets *mental modeling* of scenes from limited observations, probing the ability to reconstruct occluded spaces and simulate viewpoints. Following the official setup, we train and evaluate on the non-overlapping MindCube-10K and MindCube-Tiny respectively. **ViewSpatial-Bench** [31] isolates *multi-perspective* localization, evaluating a model's perspective-taking ability to reason across egocentric (camera) and allocentric (human or object) viewpoints. **SITE** [57] provides a *broad cognitive coverage*, unifying over thirty datasets that span diverse facets of spatial intelligence. We adopt SITE to assess the generalization ability of SenseNova-SI, as it consists of highly abstract test cases. **BLINK** [20] covers Perspective-taking tasks such as visual correspondence and multi-view reasoning. **3DSR** [39] evaluates spatial reasoning on natural images. Its paired-view setup further tests model robustness. **EmbSpatial** [16] targets on egocentric views, assessing object relation reasoning in embodied scenes.

5.2. Main Results

We compare SenseNova-SI against leading open-source and proprietary multimodal models. As shown in Tab. 1, we observe three key findings: (1) SenseNova-SI outperforms all general open-source models by clear margins, and even surpasses strong proprietary ones such as GPT-5 [42], revealing persistent knowledge gaps in existing foundation models. (2) SenseNova-SI also achieves superior performance over all dedicated spatial-intelligence models, suggesting that algorithmic innovation alone may be premature when the benefits of large-scale spatial data have not yet been fully realized. Notably, SenseNova-SI surpasses two recent strong baselines (VST [66] and Cambrian-S [68]) even when using comparable amounts of training data (Fig. 1) and a smaller model (2B parameters). We attribute these gains to the inclusion of extensive perspective-taking data, which is central to spatial intelligence. (3) While InternVL3 [74], Qwen3-VL [3], and Bagel [15] exhibit slightly different behaviors, SenseNova-SI consistently improves upon all three families. This further validates the effectiveness of our scaling strategy across diverse architecture designs and pretraining paradigms.

Moreover, we include model performance on general undestanding benchmarks (MMBench-EN [35], MMStar [10], AI2D [24], OCRB [36], DocVQA [41], MMVP [50], V* [61], MMMU [72] and Vid-MME [19]) in Tab. 8, and find that data diversity is crucial: incorporating a wide coverage of multimodal data and varied general knowledge sources effectively mitigates catastrophic forgetting and preserves overall multimodal competence.

5.3. Scaling

Effectiveness. As shown in Fig. 1, scaling spatial intelligence data leads to steady improvements across all key capability dimensions. We highlight three observa-

Models	Avg.	VSI-Bench [64]	MMSI-Bench [67]	MindCube* [70]	ViewSpatial [31]	SITE [57]	BLINK [20]	3DSR [39]	EmbSpatial [16]
Metric		MRA, Acc	Acc	Acc	Acc	CAA	Acc	Acc	Acc
Human	-	**79.2**	**97.2**	**94.5**	-	**67.5**	**95.67**	**95.7**	**90.33**
Random Choice	-	34.0	25.0	33.0	26.3	0.0	38.09	45.8	25.0
Proprietary Models									
Seed-1.6-2025-06-15 [48]	54.2	49.9	38.3	48.8	43.9	54.6	65.9	56.9	75.4
Gemini-2.5-Pro-2025-06 [49]	58.0	53.6	38.0	57.6	46.1	57.1	73.5	59.3	78.8
Grok-4-2025-07-09 [62]	53.3	47.9	37.8	63.6	43.2	47.0	56.4	54.9	75.5
GPT-5-2025-08-07 [42]	58.8	**55.0**	41.8	56.3	45.6	61.9	68.0	60.3	81.6
Gemini-3-Pro-Preview [21]	**63.8**	52.5	**45.2**	**70.9**	**50.4**	**62.3**	**76.0**	**68.9**	**84.3**
Open-source General Models									
Bagel-7B-MoT [15]	45.3	31.4	31.0	34.7	41.3	37.0	63.6	50.2	73.1
Qwen2.5-VL-3B-Instruct [4]	39.1	27.0	28.6	37.6	32.0	33.1	48.7	43.5	62.3
Qwen2.5-VL-7B-Instruct [4]	43.1	32.3	26.8	36.0	36.9	37.6	55.9	47.5	71.8
Qwen3-VL-2B-Instruct [3]	44.6	50.4	28.9	34.5	37.0	35.7	53.2	47.5	70.1
Qwen3-VL-8B-Instruct [3]	50.6	57.9	31.1	29.4	42.2	45.8	**66.7**	53.9	**77.7**
InternVL3-2B [74]	39.8	33.0	26.5	37.5	32.6	30.0	50.8	47.7	60.1
InternVL3-8B [74]	45.7	42.1	28.0	41.5	38.7	41.1	53.5	44.3	76.3
Open-source Spatial Intelligence Models									
MindCube-3B-RawQA-SFT [70]	22.0	17.2	1.7	51.7	24.1	6.3	35.1	2.8	37.0
SpatialLadder-3B [32]	40.9	44.9	27.4	43.5	39.9	28.0	43.0	42.8	58.2
Spatial-MLLM-4B [59]	35.6	46.3	26.1	33.5	34.7	18.0	40.5	36.2	50.0
SpaceR-7B [43]	41.8	41.6	27.4	38.0	35.9	34.3	49.6	40.5	66.9
ViLaSR-7B [60]	43.7	44.6	30.2	35.1	35.7	38.7	51.4	46.6	67.3
VST-3B-SFT [66]	47.7	51.4	28.8	36.0	52.9	35.9	58.8	48.7	69.0
VST-7B-SFT [66]	50.8	55.5	32.5	39.7	50.5	39.7	61.9	53.1	73.7
Cambrian-S-3B [68]	42.0	56.1	27.0	38.4	41.0	31.0	37.7	41.4	63.5
Cambrian-S-7B [68]	45.1	62.9	27.1	37.9	41.3	36.1	37.9	45.0	72.8
Ours									
SenseNova-SI $_{\text{Bagel-7B-MoT}}$	48.6(+3.3)	41.5(+10.1)	34.5(+3.5)	46.8(+12.1)	46.9(+5.6)	42.0(+5.0)	65.4(+1.8)	42.4(-7.8)	69.0(-4.1)
SenseNova-SI $_{\text{Qwen3-VL-8B}}$	58.1(+7.5)	64.8(+6.9)	38.1(+7.0)	73.8(+44.4)	51.2(+9.0)	**49.6**(+3.8)	61.9(-4.8)	53.2(-0.7)	72.5(-5.2)
SenseNova-SI $_{\text{InternVL3-2B}}$	49.4(+9.6)	63.7(+30.7)	34.2(+7.7)	41.8(+4.3)	52.7(+20.1)	36.8(+6.8)	52.4(+1.6)	50.5(+2.8)	62.8(+2.7)
SenseNova-SI $_{\text{InternVL3-8B}}$	**61.5**(+15.8)	**68.8**(+26.7)	**43.3**(+15.3)	**85.7**(+44.2)	**54.7**(+16.0)	47.7(+6.6)	63.9(+10.4)	**55.5**(+11.2)	72.0(-4.3)

Table 1. **Evaluation on key spatial intelligence and general benchmarks.** All results are evaluated on EASI [7], using the official EASI-8 protocol. MindCube* denotes MindCube-Tiny. **Dark purple** highlights the best result and light purple indicates the second-best result within Proprietary and Open-source models, respectively.

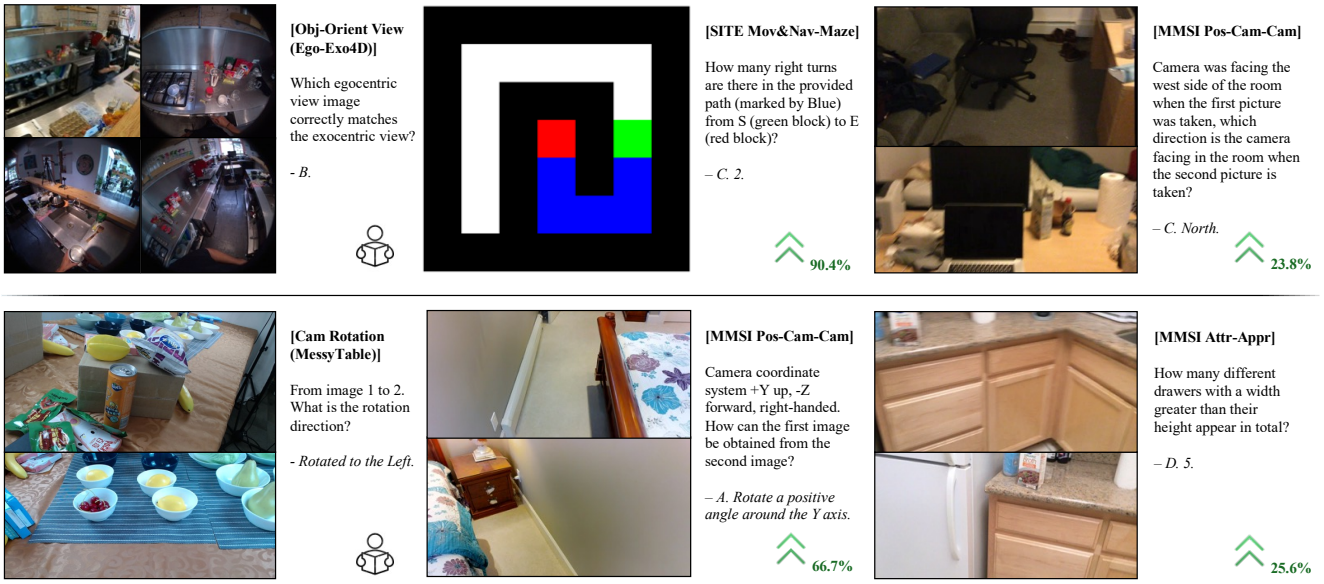

Figure 3. Observations on **generalization ability** from a single data source and single task. The upper example demonstrates how training on ego-exo association task enhance performance on task required imagined first-person perspectives. The lower example demonstrates how a camera rotation task, based on cross-view visual correspondence, generalizes to tasks with distinct questions and visual appearances. These findings suggest the potential existence of *meta-tasks* in PT, which may enable related spatial capabilities.

tions: (1) Data mixing is highly effective. By aggregating a wide collection of public datasets and further enlarging the spatial intelligence corpus, SenseNova-SI surpasses existing 7B spatial-intelligence baselines with models one size smaller (2B) under comparable data budgets. (2) Model size impacts capability trends. While InternVL3 2B and 8B variants exhibit similar performance trajectories on MM, SR, and CR tasks, their behaviors diverge sharply on PT tasks. We hypothesize that the 2B model lacks sufficient capacity to robustly learn viewpoint transformations: a challenging but essential component of spatial intelligence. (3) Capability-wise differences reveal data-driven gains. Proprietary models such as GPT-5 [42] are notably strong on SR, yet show clear deficiencies in PT. In contrast, SenseNova-SI-InternVL3-8B convincingly outperforms GPT-5 on PT, benefiting from the large-scale, comprehensive perspective-taking data introduced during continued scaling. Interestingly, even though we include very limited CR data during training, SenseNova-SI still gradually surpasses GPT-5 in CR performance. This suggests the presence of capability synergy, where advances in fundamental spatial tasks (*e.g.*, PT and SR) transfer to more complex reasoning skills (more discussions in Sec. 5.4).

Saturation. As shown in Fig. 1, the performance gains gradually diminish as the amount of training data increases. While it remains unclear whether continued scaling will eventually reach a tipping point that triggers stronger emergent capabilities (though we note some early signs discussed in Sec. 5.4), we concur with the broader community that data scaling alone is unlikely to achieve human-level spatial intelligence [68]. Motivated by this, we commit to fully open-sourcing the weights of SenseNova-SI, allowing the community to bypass the costly scaling stage and instead focus on advancing algorithmic innovation on top of a strong, spatially capable foundation.

5.4. Capability Emergence

We present interesting observations from scaling that may suggest early signs of emerging spatial intelligence.

Spill-Over. Large-scale mixed-domain training inevitably exposes models to a broad distribution of scenarios, making it increasingly difficult to determine whether downstream improvements stem from genuine, generalizable spatial reasoning or from incidental overlap with training data. To more rigorously examine spatial capability spill-over, we therefore conduct controlled experiments in which models are trained on a single dataset and evaluated on tasks drawn from entirely different domains. As shown in Fig. 3, we observe clear emergence and transfer of spatial understanding across tasks. The view-transformation dataset, constructed from Ego-Exo4D [23], requires models to translate between egocentric and exocentric viewpoints—forcing them to infer cross-view geometric relationships. This ability transfers strongly to downstream tasks such as Maze Pathfinding [57] and Pos-Cam-Cam [67], both of which depend on sequential viewpoint simulation and aggregating information across views. Similarly, the dataset built from MessyTable [6] images requires models to identify shared objects and infer spatial relationships between two viewpoints. This yields notable gains on benchmark subcategories such as MMSI [67] Pos-Cam-Cam and Attr-Appr, both of which rely on robust spatial correspondence identification between paired images.

Model	Benchmark	# Frames			
		16	32	64	128
Cambrian-S-7B [68]	VSI [64]	58.6	63.61	66.4	**67.5**
	VSI-Debiased [5]	49.7	55.6	59.1	**59.9**
SenseNova-SI $_{\text{InternVL3-8B}}$	VSI [64]	64.6	68.7	**68.8**	66.3
	VSI-Debiased [5]	58.9	**62.8**	62.4	59.7

Table 2. Ablation on inference frames. Our model was trained on maximum 16 frames per sample, while Cambrian-S-7B [68] was trained on 64/128 frames. SenseNova-SI demonstrates strong extrapolation capabilities beyond the training number of frames. Interestingly, SenseNova-SI shows a clear lead over Cambrian-S-7B [68] on two benchmarks, even with fewer frames at inference.

Models	Standard	Soft cir.	Hard cir.	w/o. Vis.
Gemini-3-Pro-Preview [21]	70.9	75.4	59.6	39.7
MindCube-SFT-RawQA [68]	51.7	45.8	23.1	50.7
SenseNova-SI $_{\text{InternVL3-8B}}$	85.6	84.0	75.6	52.5

Table 3. Analysis on MindCube [70]. *Soft cir.* and *Hard cir.* stands for Soft circular and Hard circular following [7]. *w/o. Vis.* indicates testing without visual as input, following [52].

Extrapolation. A surprising observation is that although SenseNova-SI is trained with at most 16 frames per sample, it generalizes effectively to sequences of 32 frames or more at inference time, as shown in Tab. 2. This suggests that SenseNova-SI learns to construct coherent spatial structure rather than merely repeating patterns confined to the supervised training window. Interestingly, while SenseNova-SI does not continue to extrapolate beyond 64 frames, unlike Cambrian-S [68], which is explicitly trained with much longer context windows of 64 or 128 frames, SenseNova-SI nevertheless achieves performance comparable to Cambrian-S while using substantially fewer frames at inference. This indicates that SenseNova-SI possesses a stronger spatial understanding capability that enables it to form meaningful connections across larger temporal gaps, without relying on densely sampled frame sequences.

5.5. Overfit and Shortcut Analysis

Recent studies [5] have shown that multimodal models can exploit language shortcuts to answer questions without gen-

Model	#Token	VSI-Bench
InternVL3-8B	1.0	42.1
Train set: Rel. Dir. Subset		
No CoT	3.4	40.6
CoT-GPT-5	1175.5	26.5
CoT-MindCube-Aug-CGMap	3940.7	17.0
CoT-SenseNova-SI-CGMap	2534.5	31.8
Train set: Full set (QA + CoT)		
CoT-SenseNova-SI-CGMap	190.8	49.2
+ RL (GRPO)	1299.2	43.1

Table 4. Impact of training schemes involving text-based Chain-of-Thought (CoT) and reinforcement learning on VSI-Bench.

uine visual reasoning. To ensure that the improvements of SenseNova-SI are not due to overfitting to QA text, we conduct targeted analyses on VSI [64] and MindCube [70].

The recently proposed VSI-Debiased [5] is a specifically designed variant of VSI to eliminate text-only shortcuts by removing questions that can be answered correctly without visual understanding. As shown in Tab. 2, when evaluated on VSI-Debiased, SenseNova-SI exhibits a substantially smaller performance drop compared to Cambrian-S-7B [68], indicating that SenseNova-SI relies less on textual heuristics and more on spatially grounded understanding.

For MindCube, we follow the protocol in [52] and evaluate models *without visual inputs*. Surprisingly, as shown in Tab. 3, the previous open-source SoTA on MindCube, *MindCube-RawQA-SFT* [70] achieves a score of 50.7 without any images, which is almost identical to its performance with full visual inputs, revealing a heavy dependence on language priors rather than visual reasoning.

In contrast, SenseNova-SI drops from 85.6 to 52.5 in the no-vision setting, validating that it genuinely uses visual information rather than relying on language shortcuts. Notably, both models converge to a score around 50 in the absence of images, underscoring the importance of debiasing benchmarks, as argued in VSI-Debiased [5].

To further verify that SenseNova-SI does not overfit to text option ordering, we conduct circular tests [7, 33, 35], which reorders the choices in the questions to eliminate dependency on certain text patterns. As reported in Tab. 3, SenseNova-SI exhibits minimal degradation under the Soft circular test [33]. Even in the Hard circular test [35], which requires robust handling of all rotations of answer choices, SenseNova-SI drops 10 points, whereas MindCube-RawQA-SFT drops nearly 30 points. This demonstrates that SenseNova-SI is far less sensitive to superficial text patterns.

5.6. Spatial Chain-of-Thought

Chain-of-Thought (CoT) [58] has become the *de facto* approach for complex reasoning tasks. However, despite numerous recent attempts [26, 55, 67, 70], incorporating CoT variants typically yields only marginal gains (often $\sim$2%), which are consistently overshadowed by improvements derived from large-scale curated spatial datasets.

In Tab. 4, we present a preliminary evaluation of different CoT styles. We examine three paradigms: (1) CoT-GPT-5, which directly uses a large language model (GPT-5 [42]) to annotate CoT given the question and ground-truth answer; (2) CoT-MindCube-Aug-CGMap, which follows MindCube [70] and constructs a JSON-style cognition map (CogMap) within the CoT; (3) CoT-SenseNova-SI-CGMap, our extended CogMap that provides step-by-step tracking of objects across frames, maps them to a world coordinate system with precise (rather than coarse-grid) coordinates, and reasons about relative spatial relationships more explicitly. (4) CoT-SenseNova-SI-CGMap, followed by followed by reinforcement learning based on GRPO [46].

We train each variant on roughly 100K examples, reasonably large compared to typical CoT studies, and evaluate on VSI's Object Relative Direction task, a challenging subset known to impede strong baselines such as InternVL3. We find that (1) our elaborated CoT achieves the highest improvement among the three, but (2) all CoT variants yield limited absolute gains, insufficient to justify their computational overhead, especially given the extra tokens required during both training and inference. (3) RL does not yield clear performance gains over a strong baseline. We hypothesize RL in LLM may not be readily helpful for spatial reasoning as long text may not be ideal for spatial reasoning: as discussed in Sec. J, we discover that long spatial CoT is prone to inconsistency and internal mistakes, that undermines the performance.

Our findings suggest that while carefully engineered CoT can offer modest benefits, text-based reasoning alone may be neither the most efficient nor the most effective paradigm for spatial intelligence. This may signal the need for a broader paradigm shift beyond conventional CoT. Moreover, multimodal RL for spatial reasoning remains largely underexplored, in line with SpatialReasoner [40].

6. Conclusion

In this work, we validate the effectiveness of scaling spatial intelligence across multiple multimodal foundation models, and achieve significant performance gains across the board. We further validate that the enhanced foundation models retain their general capabilities, and start to develop generalization capabilities that were not possible without training on large-scale, diverse data. We hope our study lays a solid foundation for future research on developing spatial intelligence in multimodal foundation models.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2015. 4, 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2(1):1, 2023. 3

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2, 5, 6, 11, 12, 13, 14, 15, 16, 17, 18

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 11, 12, 13, 14, 15, 16, 17, 18

[5] Ellis Brown, Jihan Yang, Shusheng Yang, Rob Fergus, and Saining Xie. Benchmark designers should" train on the test set" to expose exploitable non-visual shortcuts. *arXiv preprint arXiv:2511.04655*, 2025. 7, 8

[6] Zhongang Cai, Junzhe Zhang, Daxuan Ren, Cunjun Yu, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, and Chen Change Loy. Messytable: Instance association in multiple camera views. In *Proceedings of the European Conference on Computer Vision*, pages 1–16. Springer, 2020. 5, 7, 1, 2

[7] Zhongang Cai, Yubo Wang, Qingping Sun, Ruisi Wang, Chenyang Gu, Wanqi Yin, Zhiqian Lin, Zhitao Yang, Chen Wei, Xuanke Shi, et al. Has gpt-5 achieved spatial intelligence? an empirical study. *arXiv preprint arXiv:2508.13142*, 2025. 1, 2, 3, 4, 6, 7, 8, 11, 12

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 5

[9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 3

[10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 5, 4

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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,*, pages 24185–24198, 2024. 3

[12] Zhangquan Chen, Manyuan Zhang, Xinlei Yu, Xufang Luo, Mingze Sun, Zihao Pan, Yan Feng, Peng Pei, Xunliang Cai, and Ruqi Huang. Think with 3d: Geometric imagination grounded spatial reasoning from limited views. *arXiv preprint arXiv:2510.18632*, 2025. 3

[13] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 3

[14] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5828–5839, 2017. 5, 1

[15] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 2, 5, 6, 11, 12, 13, 14, 15, 16, 17, 18

[16] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. *arXiv preprint arXiv:2406.05756*, 2024. 2, 5, 6, 18

[17] Zhiwen Fan, Jian Zhang, Renjie Li, Junge Zhang, Runjin Chen, Hezhen Hu, Kevin Wang, Huaizhi Qu, Dilin Wang, Zhicheng Yan, et al. Vlm-3r: Vision-language models augmented with instruction-aligned 3d reconstruction. *arXiv preprint arXiv:2505.20279*, 2025. 3, 4

[18] Qi Feng. Towards visuospatial cognition via hierarchical fusion of visual experts. *arXiv preprint arXiv:2505.12363*, 2025. 4

[19] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis, 2025. 5, 4

[20] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *Proceedings of the European Conference on Computer Vision*, pages 148–166. Springer, 2024. 2, 5, 6, 16

[21] Gemini. Gemini 3 Pro Model Card. Technical report, Gemini, 2025. Accessed: 2025-11-18. 6, 7, 16, 17, 18

[22] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. *Advances in Neural Information Processing Systems*, 33:10514–10525, 2020. 4, 3

[23] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar

Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 5, 7

[24] Tuomo Hiippala, Malihe Alikhani, Jonas Haverinen, Timo Kalliokoski, Evanfiya Logacheva, Serafina Orekhova, Aino Tuomainen, Matthew Stone, and John A. Bateman. Ai2d-rst: a multimodal corpus of 1000 primary school science diagrams. *Proceedings of the Language Resources and Evaluation*, 55(3):661–688, 2020. 5, 4

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019. 4, 3

[26] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025. 8

[27] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,*, pages 2901–2910, 2017. 4, 3

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. Openvla: An open-source vision-language-action model. In *Proceedings of the Conference on Robot Learning*, pages 2679–2713. PMLR, 2025. 3

[29] Justin Lazarow, David Griffiths, Gefen Kohavi, Francisco Crespo, and Afshin Dehghan. Cubify anything: Scaling indoor 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22225–22233, 2025. 5, 1

[30] Jae Hee Lee, Matthias Kerzel, Kyra Ahrens, Cornelius Weber, and Stefan Wermter. What is right for me is not yet right for you: A dataset for grounding relative directions via multi-task learning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 1039–1045. International Joint Conferences on Artificial Intelligence Organization, 2022. 4

[31] Dingming Li, Hongxing Li, Zixuan Wang, Yuchen Yan, Hang Zhang, Siqi Chen, Guiyang Hou, Shengpei Jiang, Wenqi Zhang, Yongliang Shen, et al. Viewspatial-bench: Evaluating multi-perspective spatial localization in vision-language models. *arXiv preprint arXiv:2505.21500*, 2025. 2, 5, 6, 14

[32] Hongxing Li, Dingming Li, Zixuan Wang, Yuchen Yan, Hang Wu, Wenqi Zhang, Yongliang Shen, Weiming Lu, Jun Xiao, and Yueting Zhuang. Spatialladder: Progressive training for spatial reasoning in vision-language models. *arXiv preprint arXiv:2510.08531*, 2025. 3, 6, 11, 12, 13, 14, 15, 16, 17, 18

[33] Yijiang Li, Qingying Gao, Tianwei Zhao, Bingyang Wang, Haoran Sun, Haiyun Lyu, Robert D Hawkins, Nuno Vasconcelos, Tal Golan, Dezhi Luo, et al. Core knowledge deficits in multi-modal language models. *arXiv preprint arXiv:2410.10855*, 2024. 2, 3, 8

[34] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 4, 3

[35] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*, pages 216–233. Springer, 2024. 5, 8, 4

[36] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12), 2024. 5, 4

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[38] Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. *arXiv preprint arXiv:2110.13214*, 2021. 4

[39] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Celso M de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024. 2, 5, 6, 17

[40] Wufei Ma, Yu-Cheng Chou, Qihao Liu, Xingrui Wang, Celso M de Melo, Jianwen Xie, and Alan Yuille. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. In *Advances in Neural Information Processing Systems*, 2025. 8

[41] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. 5, 4

[42] OpenAI. GPT-5 System Card. Technical report, OpenAI, 2025. Accessed: 2025-08-10. 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18

[43] Kun Ouyang, Yuanxin Liu, Haoning Wu, Yi Liu, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Spacer: Reinforcing mllms in video spatial reasoning. *arXiv preprint arXiv:2504.01805*, 2025. 3, 6, 11, 12, 13, 14, 15, 16, 17, 18

[44] Wujian Peng, Sicheng Xie, Zuyao You, Shiyi Lan, and Zuxuan Wu. Synthesize diagnose and optimize: Towards fine-grained vision-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13279–13288, 2024. 4, 3

[45] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, et al. Sat: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 4, 3

[46] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 8

[47] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 567–576, 2015. 5, 1

[48] ByteDance Seed Team. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025. 6, 11, 12, 13, 14, 15, 16, 17, 18

[49] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 6, 11, 12, 13, 14, 15, 16, 17, 18

[50] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024. 5, 4

[51] Haochen Wang, Yucheng Zhao, Tiancai Wang, Haoqiang Fan, Xiangyu Zhang, and Zhaoxiang Zhang. Ross3d: Reconstructive visual instruction tuning with 3d-awareness. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 3

[52] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. *Advances in Neural Information Processing Systems*, 37:75392–75421, 2024. 7, 8

[53] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,*, pages 5294–5306, 2025. 3

[54] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[55] Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, Minnan Pei, Zheng Tian, Haifeng Zhang, and Jun Wang. Spatialviz-bench: Automatically generated spatial visualization reasoning tasks for mllms. *arXiv preprint arXiv:2507.07610*, 2025. 8

[56] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3

[57] Wenqi Wang, Reuben Tan, Pengyue Zhu, Jianwei Yang, Zhengyuan Yang, Lijuan Wang, Andrey Kolobov, Jianfeng Gao, and Boqing Gong. Site: towards spatial intelligence thorough evaluation. *arXiv preprint arXiv:2505.05456*, 2025. 2, 5, 6, 7, 15

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 8

[59] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 3, 6, 11, 12, 13, 14, 15, 16, 17, 18

[60] Junfei Wu, Jian Guan, Kaituo Feng, Qiang Liu, Shu Wu, Liang Wang, Wei Wu, and Tieniu Tan. Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. *arXiv preprint arXiv:2506.09965*, 2025. 3, 6, 11, 12, 13, 14, 15, 16, 17, 18

[61] Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094, 2024. 5, 4

[62] xAI. Grok 4, 2025. Model announcement. 6, 11, 12, 13, 14, 15, 16, 17, 18

[63] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025. 3, 4

[64] Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,*, pages 10632–10643, 2025. 2, 5, 6, 7, 8, 11, 20, 22

[65] Rui Yang, Hanyang Chen, Junyu Zhang, Mark Zhao, Cheng Qian, Kangrui Wang, Qineng Wang, Teja Venkat Koripella, Marziyeh Movahedi, Manling Li, et al. Embodiedbench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents. *arXiv preprint arXiv:2502.09560*, 2025. 2, 3, 4, 5

[66] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, Yi Lin, and Hengshuang Zhao. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025. 3, 5, 6, 4, 11, 12, 13, 14, 15, 16, 17, 18

[67] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 2, 5, 6, 7, 8, 12

[68] Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, et al. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*, 2025. 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18

[69] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 5, 1

[70] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chan-

drasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop*, 2025. 2, 3, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 18

[71] Songsong Yu, Yuxin Chen, Hao Ju, Lianjie Jia, Fuxi Zhang, Shaofei Huang, Yuhan Wu, Rundi Cui, Binghao Ran, Zaibin Zhang, et al. How far are vlms from visual spatial intelligence? a benchmark-driven perspective. *arXiv preprint arXiv:2509.18905*, 2025. 3

[72] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 5, 4

[73] Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang. Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025. 4

[74] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 2, 3, 5, 6, 11, 12, 13, 14, 15, 16, 17, 18

[75] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Proceedings of the Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 3