Visual Spatial Tuning

Rui Yang[1,*], Ziyu Zhu[3,*], Yanwei Li[2,✉], Jingjia Huang[2], Shen Yan[2], Siyuan Zhou[2], Zhe Liu[1], Xiangtai Li[2], Shuangye Li[2], Wenqian Wang[2], Yi Lin[2], and Hengshuang Zhao[1,✉]

[1]The University of Hong Kong , [2]ByteDance Seed, [3]Tsinghua University
Code: https://github.com/Yangr116/VST

Abstract. Capturing spatial relationships from visual inputs is a cornerstone of human-like general intelligence, yet current Vision-Language Models (VLMs) still struggle to perform spatial understanding from everyday visual observations. To enhance the spatial ability within general architectures, we introduce Visual Spatial Tuning (VST), a comprehensive framework to cultivate VLMs with human-like visuospatial competence spanning spatial perception and reasoning. VST dataset comprises VST-P, a large-scale perception dataset with 4.1 million samples spanning 19 skills across single views, multiple images, and videos; and VST-R, a reasoning dataset with 135K samples that explicitly instruct models to reason in space. For spatial reasoning, we design Reconstruction with Text in Chain-of-Thought (RT-CoT), which represents scene layout as a textual intermediate and performs reasoning grounded on this reconstruction. We further propose a strategy termed prompting with BEV annotations to improve the faithfulness and coherence of spatial reasoning traces. Then, we adopt a progressive training pipeline: supervised fine-tuning to build foundational spatial perception abilities, followed by reinforcement learning to develop spatial reasoning, where a 3D-aware layout reward is proposed. Without the side-effect on general capabilities, the proposed VST consistently achieves state-of-the-art results on several spatial benchmarks, including 34.8% on MMSI-Bench and 61.2% on VSIBench. Moreover, VST substantially improves Vision-Language-Action (VLA) models, paving the way for more physically grounded AI.

Keywords: Spatial Intelligence · Vision-language Model

1 Introduction

Vision-Language Models (VLMs) [1,12,15,24,54] have achieved remarkable success across a wide range of domains, such as visual question answering [33,66], document understanding [34,37], and autonomous GUI agents [57]. However, these models exhibit limitations in capturing spatial relationships from sequential visual observations [61,63]. This spatial understanding ability is a foundational component of general intelligence and is present across a broad spectrum

* Equal Contribution, ✉ Corresponding Authors

Method	Data Type			Data Usage	
	SI	MI	Video	SFT	RL
SpatialVLM	✓	✗	✗	✓	✗
SAT	✓	✗	✗	✓	✗
MM-Spatial	✓	✗	✗	✓	✗
SPAR	✓	✓	✓	✓	✗
Space-R	✗	✗	✓	✓	✓
VLM-3R	✗	✗	✓	✓	✗
VST (ours)	✓	✓	✓	✓	✓

Table 1: Comparison with data-driven methods in data modality coverage (single-image, multi-image, video) and training usage.

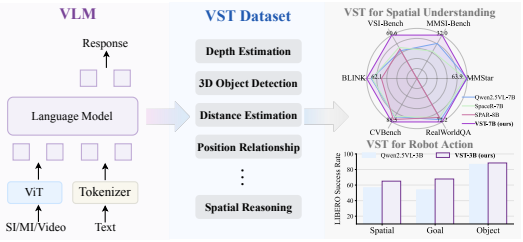


Fig. 1: Overview of the VST framework. It endows spatial awareness for VLMs by introducing the VST dataset, achieving improved spatial abilities, and boosting downstream VLA performance.

of animals, including humans [25,40]. The deficiency significantly constrains current VLMs to effectively interact with the physical world, thereby limiting their application in fields such as robotics [7,73], autonomous driving [50], and augmented/virtual reality (AR/VR) [23]. To mitigate this issue, several studies have explored the incorporation of additional expert encoders [5,20]. However, this approach often introduces extra complexity and can negatively impact the general capabilities of the models. Alternatively, other research efforts have focused on developing specialized datasets [9,17,39,58,65,70], aiming to enhance the spatial capabilities of VLMs from a data-driven perspective.

Nevertheless, these works have typically concentrated on limited or isolated aspects of spatial understanding, as summarized in Table 1. Existing studies, such as SpatialVLM [9] and the like [14,17], focus only on the supervised fine-tuning stage, while others are restricted to a single scenario [39,59], overlooking the diversity of visual input. To this end, we introduce a comprehensive and integrated framework, termed Visual Spatial Tuning (VST), which is designed to holistically cultivate human-like visuospatial abilities in VLMs. As illustrated in Figure 1, VST effectively augments the spatial capabilities of existing VLMs by leveraging a carefully curated dataset that encompasses diverse data modalities and task types, coupled with a progressive training approach. This enhancement proves advantageous for downstream Vision-Language-Action (VLA) tasks.

To develop the VST, we deconstruct spatial ability into two key components: *spatial perception* and *spatial reasoning*. We define spatial perception as the ability to discern the spatial relationships between objects, and spatial reasoning as the ability to build and mentally manipulate an internal model of an environment. These two components correspond to the concepts of perceptual and conceptual spatial ability, respectively, as proposed in cognitive science [40]. Effective spatial perception requires the model to possess foundational spatial knowledge—specifically, the ability to identify both "what is it?" and "where is it?" within its peripersonal space. While existing VLMs can accurately recognize objects and locate them within pixel space using 2D points or bounding boxes [18,54], their ability to determine object positions in 3D space, find multi-view consistency from multiple viewpoints, and capture spatiotemporal relationships from videos remains limited [51,61,63]. Therefore, we introduce

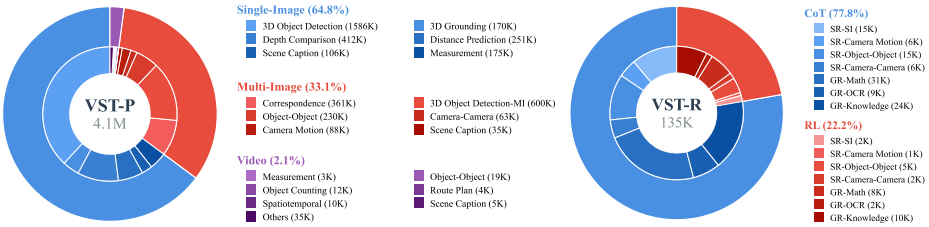
the VST-Perception (VST-P) dataset, comprising 4.1 million samples across 19 diverse tasks. The dataset incorporates single-image data to facilitate VLMs in discerning spatial relationships beyond the pixel level, which is an essential step towards bridging the gap between pixel space and 3D space. In addition, multi-image data is included to enhance the ability to comprehend spatial relationships from multiple viewpoints, and video data enables the capture of spatiotemporal relationships. More detail is in Section 3.1. Collectively, this dataset provides a foundation for advancing spatial perception in VLMs.

Beyond foundational spatial perception, we expect the model to mentally represent spatial relationships beyond its own body, thereby engaging in advanced spatial reasoning. To this end, we propose Reconstruction with Text in Chain-of-Thought (RT-CoT), where the model constructs a scene layout in text and reasons grounded on the textual reconstruction. Then, we introduce the VST-Reasoning (VST-R) dataset (see Section 3.2), which comprises samples featuring CoT processes to facilitate the spatial reasoning ability, as well as samples with rule-checkable answers to further enhance its reasoning capabilities. However, when generating spatial CoT, the limited spatial understanding of current large VLMs [61, 63] poses challenges for synthesizing accurate layout descriptions and coherent reasoning chains. Drawing inspiration from human cognition, we propose prompting with Bird's-Eye View (BEV) annotation. It leverages a top-down perspective to explicitly convey spatial relationships between objects, thereby improving the quality of both layout descriptions and CoT reasoning traces.

Building upon the introduced VST-P and VST-R datasets, we propose to inject visual spatial knowledge into VLMs through supervised fine-tuning (SFT) and further enhance spatial reasoning capabilities via reinforcement learning (RL). Note that we design a layout reward for 3D object detection tasks, which effectively improves the accuracy of predictions (see Section 4.2). This progressive approach mirrors the development of human spatial intelligence [40], i.e., establishing a foundation in spatial perception before developing higher-level spatial reasoning abilities. As a result, our proposed VST framework consistently achieves state-of-the-art performance on multiple spatial benchmarks, attaining 87.8% on CVBench [51], 34.8% on MMSI-Bench [63], and 61.2% on VSI-Bench [61], while preserving general multimodal capabilities. Furthermore, the spatial proficiency from VST demonstrably enhances broader robotic tasks. For instance, Qwen2.5VL-3B [3] fine-tuned on our VST yields an 8.6% improvement on the LIBERO benchmark [32].

Our contributions can be summarized as follows:

- We explicitly decompose spatial capability into spatial perception (single-image geometry, multi-view consistency, spatiotemporal understanding) and spatial reasoning, then create the large-scale VST dataset and validate a stage-wise curriculum via controlled ablations.
- For spatial reasoning, we propose RT-CoT, a learnable textual intermediate representation that explicitly reconstructs the scene layout within the reasoning trace. This formulation is more general than hint maps. Furthermore,



(a) Perception data distribution. **(b)** Reasoning data distribution.

Fig. 2: Overview of the VST dataset. (a) The distribution of VST-P, which is used for SFT. (b) The distribution of VST-R, which is used for CoT cold start and RL. 'SR' denotes spatial reasoning, and 'GR' denotes general reasoning.

we design a strategy termed prompting with BEV annotations, which helps the teacher model capture spatial relationships and generate more coherent reasoning traces.

– We design a 3D-aware layout reward to explicitly optimize layout learning during RL, leading to improvements beyond SFT (+5.7 AP_{15} on SUN-RGBD).

2 Related Work

VLMs have become a cornerstone of modern AI. Recent advancements on VLMs have focused on several areas: employing more powerful vision encoders [13], supporting dynamic resolutions [54], incorporating reasonable positional embeddings [3], and curating higher-quality data [30]. Despite these significant improvements, established benchmarks [61,63] reveal a persistent deficiency in the spatial understanding and reasoning capabilities of current VLMs. Early attempts introduced specialized datasets generated by expert models [9], while SAT [43] tackled data scarcity by simulation. Subsequent works [17, 70] have further expanded data to enhance these spatial abilities. There are also efforts [39, 55, 65] to improve spatial reasoning. In contrast to previous approaches that treat spatial understanding and reasoning as distinct tasks, the proposed VST integrates foundational spatial perception and reasoning capabilities, aiming to achieve a more holistic and robust model for spatial intelligence. Whereas previous methods like Shikra [10], Ferret [69], and LocVLM [41] operate in 2D pixel space, VST focuses on determining object relationships in 3D space.

3 Dataset

In this section, we introduce the VST dataset, specifically developed to enhance the spatial perception and reasoning capabilities of VLMs. First, we construct a large-scale dataset, **VST-Perception (VST-P)**, to equip VLMs with comprehensive spatial perception abilities: to determine object positions in 3D space, to find multi-view consistency from multiple viewpoints, and to capture spatiotemporal relationships from videos. Building upon this foundation, we further create the **VST-Reasoning (VST-R)** dataset to enable VLMs to reason in space.

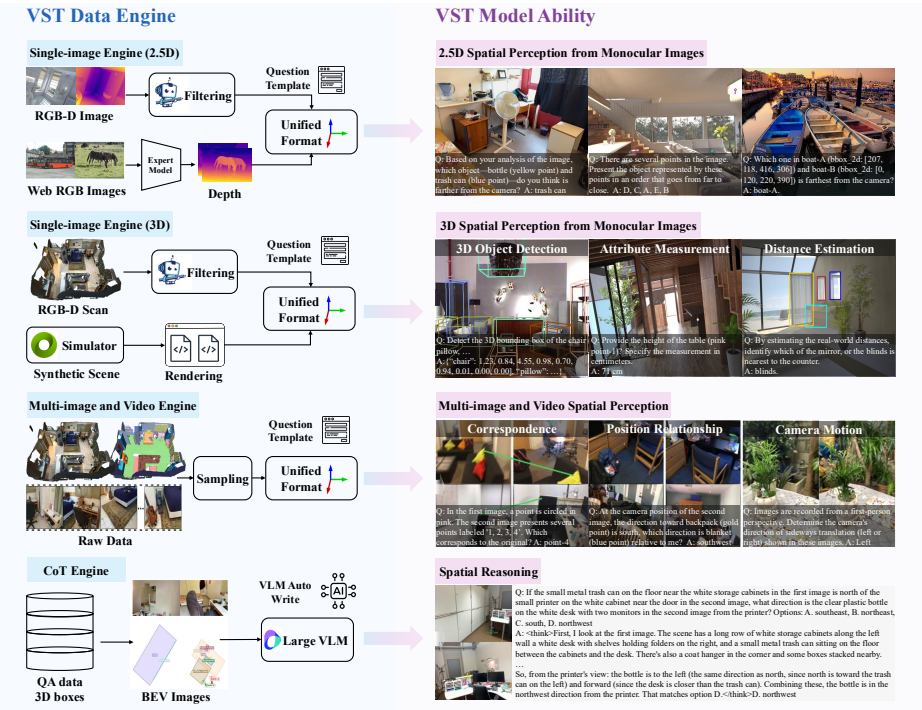


Fig. 3: Data engines of VST and the capabilities they enable in VST-Model.

3.1 VST-Perception

Recognizing that spatial awareness from single images [62], multi-view consistency [26], and spatiotemporal modeling [53] are essential for 3D spatial understanding, we constructed the VST-P dataset across these three aspects; as illustrated in Figure 2a, it comprises 4.1 M samples spanning 19 tasks within single-image, multi-image, and video scenarios.

Single-image. Since single-image data is easily obtainable, this part constitutes the largest category. This category primarily encompasses tasks such as relative depth estimation (2.5D), 3D object detection, and distance estimation. These tasks bridge the gap between 2D pixel coordinates and the 3D physical world, thereby facilitating the acquisition of spatial knowledge and the development of spatial awareness in VLMs. To curate this data, we create dedicated data engines to gather metadata with depth maps and 3D bounding box annotations, as shown in the top left of Figure 3. The depth data mainly comes from public real-world datasets [16,64] and synthetic datasets [45]. To increase the diversity of depth data, we use a depth expert model [62] to generate pseudo-labels for wild images [31], with these data accounting for 3.3% of the VST-P dataset. After obtaining the depth maps, we convert them to the same coordinate system and generate depth-related visual instruction samples. The reference formats for depth-related samples encompass text-based, point-based, box-based, and visual-prompt-based representations. These diverse formats enable VLMs to infer the

relative distance from objects to the camera plane. Some examples are shown in the Appendix.

For the 3D data engine, we also use two main approaches. The first approach relies on open-source datasets [2, 4, 8, 16, 45, 48]. Each dataset is designed for distinct applications, provides visual data in either video or image format, and annotates objects in varying coordinate systems. Therefore, we standardize all collected 3D bounding boxes to a unified camera coordinate system and process the raw visual data to reduce repetition and occlusion. The second approach is to generate data using the Isaac Sim[1] simulator. The virtual scenes are from the GUTopia [52]. With large-scale 3D bounding box data, we create visual instruction samples for 3D object detection, 3D grounding, attribute measurement, and distance estimation tasks. Notably, when projecting 3D annotations onto a 2D plane, spatial shifts and distortions may arise, leading to noisy 2D bounding boxes that do not accurately cover the target objects. As a result, object referring based on these projected loose 2D boxes becomes unreliable, potentially diminishing the model's original knowledge. To mitigate this issue, we leverage expert models [24,42] to evaluate object presence and generate masks for objects given the projected 2D boxes.

In 3D object detection, we predict the 9-DoF bounding box in the camera coordinate system. The 3D bounding box is defined by $(x, y, z, x_l, y_l, z_l, p, y, r)$, where (x, y, z) are the box centers, (x_l, y_l, z_l) are the lengths along the X, Y, and Z axes, and (p, y, r) are the rotation angles. When creating the instruction data, we **mix single-turn and multi-turn formats**. The multi-turn format allows each subsequent box to reference the previous one during training, helping the model learn layout information from context.

Furthermore, since current VLMs are language-centric, if we rely solely on template-based 3D object detection data for training, the VLM may learn a shortcut on specific tasks and fail to generalize spatial perception abilities. Therefore, to help the VLM better comprehend spatial information at the language level, we introduce the **Scene Caption**. Unlike general captions, which primarily describe image content, scene captions focus on the layout information and spatial relationships within the image. To obtain such scene captions, we prompt a large VLM [24] with ground-truth 3D bounding boxes and object relationships extracted from the scene graph [72]. The resulting scene captions not only describe the objects, but also provide detailed layout information and spatial arrangements. We demonstrate examples in the Appendix.

Multi-image. To enable the VLM to comprehend spatial relationships and find multi-view consistency across different viewpoints, we create multi-image data, which includes tasks such as multi-view 3D object detection, multi-view correspondence, object-object relationship understanding, and camera motion analysis. As illustrated in the third data engine of Figure 3, we sample multi-image data from RGB-D scans [4, 16, 64]. To unify object information across multiple images, we transform all objects into the camera coordinate system of the first image. For correspondence tasks, we utilize point clouds and depth maps

[1] https://developer.nvidia.com/isaac/sim

from various viewpoints to identify matched points. For camera motion data, we represent camera poses using Euler angles. After these unification steps, we generate template-based visual instruction samples. In the multi-image scenario, we also create the scene caption to reconstruct the scene layout with text and describe the spatial information represented by multiple RGB images.

Video. To enable the VLM to capture spatiotemporal relationships from videos, we create video data involving tasks such as identifying the order of appearances and counting objects. To construct the video dataset, we employ the same data engine used for multi-image data. The only difference is that we add the appearance time for each object. We also enhance the video dataset by sampling two-thirds of the data from VLM-3R [20], reorganizing it into a multi-turn format rather than a single-turn format.

With the introduction of the VST-P dataset, the VLM exhibits significantly enhanced fundamental capabilities for comprehending spatial relationships. Notably, there is a ∼**20**% improvement on CVBench-3D [51], a ∼**5**% increase on BLINK [21], and a ∼**16**% gain on VSIBench [61], as illustrated in Tables 5.

3.2 VST-Reasoning

To move beyond foundational spatial perception toward spatial reasoning, we develop the VST-Reasoning (VST-R) dataset. The dataset is designed to elicit and improve spatial reasoning in VLMs. As illustrated in Figure 2b, the VST-R consists of 135K samples, divided into two subsets: one part includes CoT traces to teach the model how to reason, and the other part provides rule-checkable data used in online RL to improve the reasoning ability. To obtain such spatial reasoning samples, we develop a data engine, as illustrated in the bottom left of Figure 3. Specifically, we sample question-answer pairs from VST-P and employ a large VLM [24] as the teacher to generate the detailed reasoning trace.

Prompting with BEV annotations. Since current large VLMs still lag behind their general multimodal capabilities in spatial understanding [61, 63], we introduce a novel strategy named *prompting with BEV annotations*. Given a multi-view sample with RGB images $I_{1:K}$ and ground-truth 3D bounding boxes $\mathcal{B}$, we prompt the teacher model $\mathcal{T}$ to produce reasoning traces:

$$y^{cot} = \mathcal{T}(\, I_{1:K},\ \Pi_{\text{bev}}(\mathcal{B}),\ \text{Meta}(\mathcal{B}),\ q,\ a\,), \tag{1}$$

where $\Pi_{\text{bev}}(\cdot)$ denotes a deterministic rendering function that renders 3D boxes onto the top-down plane with object categories and orientations. (q, a) is the question-answer pair, and $\text{Meta}(\mathcal{B})$ denotes compact textual metadata extracted from $\mathcal{B}$. The BEV image provides a global, viewpoint-invariant summary of scene layout, helping the teacher model establish consistent correspondences across views. As a result, the generated spatial reasoning traces are more coherent and accurate. The format of y^{cot} follows RT-CoT, as detailed in Section 4.2.

VLMs trained with the VST-R dataset exhibit significantly enhanced spatial reasoning performance. As illustrated in Table 8, there is an absolute gain of **8.9**% observed on MMSI-Bench [63].

4 Method

Our goal is to equip general VLMs with 3D knowledge for better spatial perception and reasoning from monocular images. Therefore, we choose Qwen2.5-VL [3] as the base model because it can accurately identify objects and locate them in the image plane. It follows the widely used "ViT-MLP-LLM" paradigm: a pre-trained Vision Transformer (ViT) is combined with a large language model (LLM) via an MLP merger.

4.1 Learning Spatial Perception by SFT

We continue training the base model to endow it with spatial perception abilities. Specifically, we fine-tune the base model on the proposed VST-P dataset. To preserve the original capabilities of the base model, we also incorporate a portion of general multimodal data from open-source datasets [18,30]. Let the base model be parameterized by θ, which can simultaneously process text, images, and video. For any given training sample $x = [x_1, \ldots, x_L]$ of length L, we employ visual tokens as the conditioning context for text prediction and adopt the standard auto-regressive objective:

$$\mathcal{L}_\theta(x) = -\sum_{i=2, x_i \in \text{text}}^{L} w_i \log p_\theta(x_i \mid x_1, \ldots, x_{i-1}), \tag{2}$$

where w_i refers to the loss weight of the i-th token.

FoV Unification. A major challenge in metric-related tasks is the large variation of camera intrinsics across mixed datasets, which leads to inconsistent pixel-to-ray geometry and harms metric generalization. To mitigate this issue, we propose to let each vision token share a consistent projection geometry. Specifically, we introduce a **FoV unification** module that normalizes images from different cameras into a canonical virtual camera with a predefined focal length: $I_{\text{uni}} = \mathcal{W}(I;\ f,\ f_{\text{uni}})$, where I is the original image, f denotes its focal length (in pixels), f_{uni} is a pre-defined unified focal length, and $\mathcal{W}(\cdot)$ is a geometry-consistent warping operator that projects I into a virtual camera. We apply this module only for metric-sensitive data, which reduces the intrinsic-related distribution shift and stabilizes metric learning across heterogeneous datasets.

The resulting model from the SFT stage exhibits foundational spatial perception abilities, as illustrated in Table 5.

4.2 Learning Spatial Reasoning by RL

With these spatial perception abilities, we further develop spatial reasoning by introducing the spatial chain-of-thought.

RT-CoT. Since text offers greater generality in the current VLM paradigm, we propose **RT-CoT** (**R**econstruction with **T**ext in **C**hain-**o**f-**T**hought), a text-based reasoning format that explicitly factorizes spatial reasoning into (i) reconstructing a scene layout in text and (ii) answering the question based on

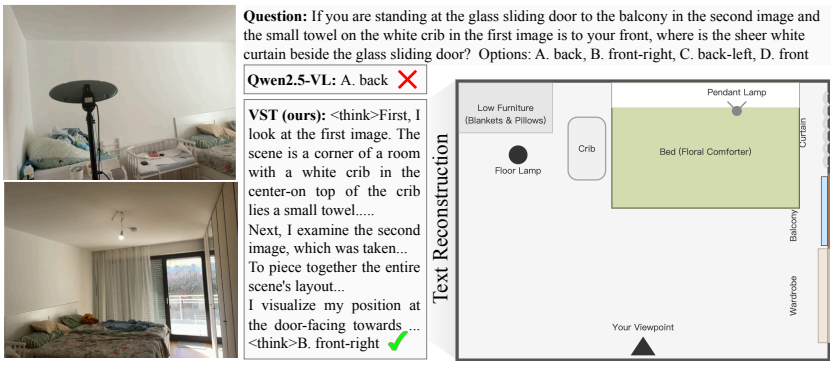


Fig. 4: Visualization of the textual reconstruction. The thinking content can be used to reconstruct the layout of a scene using SVG. Best viewed when zoomed in.

the reconstructed textual layout. The RT-CoT is organized into three segments: $y = [\texttt{[RECON]} \mid \texttt{[REASON]} \mid \texttt{[ANSWER]}]$, where `[RECON]` is a structured textual layout, `[REASON]` is the reasoning trace grounded on `[RECON]`, and `[ANSWER]` is the final answer. Figure 4 shows an example where the textual reconstruction can be rendered as an SVG. Compared to using a cognitive map [65] to reconstruct the scene layout, the proposed RT-CoT is more general (Table 8).

To enable the model to follow the RT-CoT, we conduct a cold-start stage. This stage leverages CoT data from VST-R to instruct the model to learn reasoning trajectories. The training objective remains consistent with Equation 2. The resulting model from this stage demonstrates foundational spatial reasoning capabilities (Table 8) and acts as the initial RL actor.

Then, we employ a reinforcement learning stage to further optimize the spatial reasoning ability. Specifically, we utilize the Group Relative Policy Optimization (GRPO) algorithm [46], which eliminates the need for an explicit value model by computing the relative advantage within a group of responses to the same question. The model after the cold-start stage serves as the initial RL actor. We categorize the training set into four task types: multiple-choice, open-ended, OCR, and 3D detection. In the GRPO framework, we design a mixed rule-based reward to score the generated responses. For a given task s with the predicted response $\hat{y}$ and its corresponding ground truth y, the overall reward function is defined as:

$$\mathcal{R}(s; y, \hat{y}) = \mathcal{G}(s; \mathcal{R}_{\text{acc}}(y, \hat{y}), \mathcal{R}_{\text{layout}}(y, \hat{y})) + \mathcal{R}_{\text{format}}(y, \hat{y}). \tag{3}$$

$\mathcal{G}(\cdot)$ performs hard routing based on the task type: it selects $\mathcal{R}_{\text{acc}}(\cdot,\cdot)$ for multiple-choice, open-ended, or OCR tasks and $\mathcal{R}_{\text{layout}}(\cdot,\cdot)$ for the 3D detection task. The accuracy reward $\mathcal{R}_{\text{acc}}(\cdot,\cdot)$ is calculated using standard evaluation protocols [22,36,37,47]. $\mathcal{R}_{\text{format}}(\cdot,\cdot)$ is the format reward, which incentivizes adherence to a specified output format.

Layout Reward. For the 3D object detection task, we propose a layout reward which is a linear combination of the 3D Intersection over Union (IoU) score and

the F1 score:

$$\mathcal{R}_{\text{layout}}(y, \hat{y}) = \alpha \mathcal{R}_{\text{iou}}(y, \hat{y}) + (1 - \alpha)\mathcal{R}_{\text{F1}}(y, \hat{y}), \tag{4}$$

where α is a hyperparameter that defaults to 0.5. In detail, given N predicted and M ground-truth 3D bounding boxes, we first establish a bipartite matching [29] between the predictions and the ground truth; $\mathcal{R}_{\text{iou}}(\cdot, \cdot)$ is then calculated as the average 3D IoU of the successfully matched pairs. To calculate $\mathcal{R}_{\text{F1}}(\cdot, \cdot)$, we define a true positive as a matched pair whose 3D IoU exceeds 0.25.

Following this stage, the model exhibits superior spatial reasoning abilities relative to the cold-start model, as shown in Table 8.

4.3 Probing Spatial-awareness using Vision-Language Action Model

With the spatially enhanced model, a natural question emerges: *does the integration of spatial priors improve the performance of Vision-Language-Action (VLA) models in robotic manipulation tasks?* To this end, we adapt the pre-trained VLM into a VLA model, following the methodology of OpenVLA [28]. Specifically, we formulate the action prediction problem as a vision-language task where, given an observation image and a natural language instruction, the model auto-regressively predicts the actions with a single-step policy. To accomplish this, we discretize the action space into 256 bins, where each bin corresponds to a special token in the language tokenizer. With the actions tokenized, the entire model is fine-tuned using the objective function defined in Eq. 2.

5 Experiments

5.1 Implementation Details

The training details can be found in the Appendix. For evaluation, we assess spatial understanding across three distinct modalities: single-image ability is benchmarked with CV-Bench [51] and 3DSRBench [35], multi-image ability with BLINK [21] and MMSI-Bench [63], and video-based ability with VSI-Bench [61]. The average score (S-AVG) across these benchmarks is used to quantify the overall spatial capabilities. To assess general capability, we also report the average score (M-AVG) across MMStar [11], MMBench (MMB) [33], RealworldQA (RWQA) [56], MMMU [66], OCRBench (OCRB) [34], and AI2D [27]. For 3D object detection, we evaluate the model on the SUN RGB-D [48] (Total3D version [38]) and ARKitScenes [4] (Omni3D version [6]).

5.2 Main Results

As shown in Table 2, our VST models achieve competitive results across both spatial and general benchmarks. Notably, VST-7B-SFT and VST-7B-RL deliver leading performance on mainstream spatial understanding tasks. On the

Methods	CV	3DSR	MMSI	BLINK	VSI	MMStar	MMB	RWQA	MMMU	OCRB	AI2D
GPT-4o [1]	76.0	45.3	30.3	65.9	34.0	65.1	84.3	76.2	70.7	80.6	84.9
Gemini-2.5-Pro [15]	-	-	36.9	70.6	-	77.5	90.1	78.0	81.7	86.6	88.4
Seed1.5-VL [24]	85.2	61.6	29.7	72.1	41.5	77.8	89.9	78.4	77.9	86.1	87.3
LLava-OneVision-7B [30]	61.9	54.4	26.6	48.2	32.4	61.7	80.8	66.3	48.8	62.2	81.4
InternVL3-8B [71]	81.0	55.7	25.7	55.5	42.1	68.2	83.4	70.8	62.7	88.0	85.2
MiMo-VL-7B-RL [67]	82.3	50.8	29.3	62.4	37.2	65.1	84.4	68.2	66.7	86.6	83.5
SpaceR-7B [39]	74.8	53.3	20.1	55.4	43.5	61.6	84.3	64.7	53.1	85.9	85.5
SPAR-8B [70]	80.7	57.5	-	43.9	41.1	-	79.9	64.7	-	-	-
Qwen2.5-VL-3B [3]	71.8	50.2	26.5	47.6	29.6	55.9	79.9	65.4	47.9	79.7	81.6
Qwen2.5-VL-7B [3]	80.6	53.2	25.9	56.4	38.9	63.9	83.5	68.5	51.2	86.4	83.9
VST-3B-SFT (ours)	84.4	54.1	30.2	59.1	57.9	58.0	80.9	68.4	45.2	83.7	82.5
VST-3B-RL (ours)	84.2	56.5	31.3	57.2	57.7	58.9	80.5	68.5	49.8	80.9	82.4
VST-7B-SFT (ours)	85.5	54.6	32.0	62.1	60.6	63.1	83.3	72.2	50.6	85.5	84.9
VST-7B-RL (ours)	**86.5**	**60.1**	**34.8**	**62.6**	**61.2**	63.5	83.0	68.5	49.4	86.1	83.5

Table 2: Comparison with state-of-the-art VLMs on spatial and general benchmarks. All results are evaluated using VLMEvalKit [19]. Qwen2.5-VL [3] serves as the base model for our proposed VST. Baseline models are in gray.

CV-Bench [49], VST-7B-SFT attains 85.5, surpassing the proprietary Seed1.5-VL [24] with 85.2. On MMSI-Bench [63], VST-7B-SFT achieves 32.0, outperforming GPT-4o [1] at 30.3, while RL further boosts VST-7B-RL to 34.8, approaching the proprietary state-of-the-art Gemini-2.5-Pro [15] at 36.9. Notably, the VSI-Bench [61] highlights the strength of our models in video spatial understanding. VST-7B-SFT reaches 60.6, and VST-3B-SFT achieves 57.9, substantially ahead of GPT-4o with 34.0. Detailed results are reported in Appendix. Without any specialized 3D encoder, VST-7B-RL delivers the best overall average among comparable VLMs, achieving 61.2. Although VLM-3R-7B [20] attains a similar score, it relies on an additional expert 3D encoder, whereas VST operates with a standard vision backbone. In addition, VST models provide a well-balanced performance on general benchmarks. These outcomes highlight the clear advantage of VST in spatial perception and reasoning while maintaining strong competitiveness in multimodal understanding. Table 3 summarizes results on the SUN RGB-D [48] 3D object detection benchmark. VST-7B-SFT reaches 41.6 AP_{15}, while VST-7B-RL improves to 44.2, ranking first among both general VLMs [15, 24] and expert methods [38, 68]. These findings show that VST, even without auxiliary 3D encoders, can achieve strong 3D object detection purely through visual spatial tuning.

5.3 Ablation Study

Larger model and data scale boost spatial abilities. We investigate the relationship between model and data scale and spatial abilities. The results presented in Table 4 demonstrate that both larger model size and increased data scale contribute to enhanced spatial abilities, as evidenced by improved performance on spatial tasks. In addition, tripling the data scale yields performance improvements of 1.1%, 1.5%, and 1.7% for the 3B, 7B, and 32B models, respectively, on spatial benchmarks. These findings suggest that larger models possess greater learning capacity.

Methods	AP_{15}
Seed1.5-VL	33.5
Gemini-2.0-Pro	32.5
Gemini Robotics-ER	**48.3**
Implicit3D	24.1
Total3DU	14.3
VST-3B-SFT (ours)	37.3
VST-3B-RL (ours)	40.1
VST-7B-SFT (ours)	41.6
VST-7B-RL (ours)	**44.2**

Table 3: Comparison AP_{15} on SUN RGB-D.

Model Size	VST Data Scale	S-AVG	Single-image			Multi-image		Video
			CV-2D	CV-3D	3DSR	MMSI	BLINK	VSI
3B	Qwen2.5-VL-3B [3]	49.6	71.2	72.5	50.2	26.5	47.6	29.6
	1×	60.6	75.1	93.1	54.0	31.3	55.6	54.7
	3×	61.7	75.0	93.8	54.1	30.2	59.1	57.9
7B	Qwen2.5VL-7B [3]	55.9	77.4	83.4	53.2	25.9	56.4	38.9
	1×	61.9	76.5	94.4	53.5	31.9	58.3	57.0
	3×	63.4	75.2	95.7	54.6	32.0	62.1	60.6
32B	Qwen2.5-VL-32B [3]	58.2	77.3	86.5	56.7	27.7	59.9	40.9
	1×	63.6	77.5	93.1	55.7	36.9	61.4	57.1
	3×	65.3	80.7	94.8	56.1	36.0	65.4	58.7

Table 4: Scaling model and data size. Baseline models are in gray.

Data	Data Type	S-AVG	Single-image			Multi-image		Video	MM AVG
			CV-2D	CV-3D	3DSR	MMSI	BLINK	VSI	
Baseline	-	49.9	71.2	72.6	50.5	26.1	49.2	29.6	68.3
+ 3D Object Detection	*si*	50.9	71.5	78.3	51.0	25.5	48.2	30.9	70.0
+ 3D Grounding	*si*	50.5	73.3	72.3	50.5	27.7	48.4	31.0	69.5
+ Scene Caption	*si*	52.8	72.4	83.8	52.2	25.9	49.4	33.1	69.8
+ Measurement	*si*	53.3	71.9	83.0	53.0	26.7	49.4	35.5	69.2
+ Depth and Distance	*si*	**56.4**$_{+6.5}$	73.5	93.4	53.2	28.8	50.6	38.7	69.7
+ Correspondence	*mi*	56.7	74.1	92.5	53.9	27.9	52.4	39.1	69.7
+ 3D Object Detection	*mi*	56.5	73.3	92.3	53.1	30.0	52.7	37.7	69.2
+ Object-object Relation	*mi*	57.1	74.3	91.9	53.2	31.9	53.2	38.3	69.4
+ Camera-camera Relation	*mi*	57.3	72.7	93.1	53.6	31.8	53.7	38.8	68.8
+ Scene Caption	*mi*	57.4	73.2	92.8	54.0	32.1	53.0	39.3	69.2
+ Camera Motion	*mi*	57.7	73.9	92.5	54.0	32.4	55.0	38.2	68.7
+ General Data	*mi*	**57.7**$_{+1.3}$	73.8	92.4	53.7	32.7	55.4	38.4	69.2
+ General Video	*vi*	57.9	74.3	92.3	53.5	31.9	57.6	38.1	69.4
+ VST-video	*vi*	**60.6**$_{+2.9}$	75.1	93.1	54.0	31.3	55.6	54.7	69.4

Table 5: Ablation on the effectiveness of different data. *si*, *mi*, and *vi* denote single-image, multi-image, and video data in VST-P.

Effectiveness of the curated data. Our baseline model is Qwen2.5-VL-3B fine-tuned on a general dataset of 800K samples. From this baseline, we progressively introduce different types of data and find that *the key bottleneck for single-image spatial perception is the lack of spatial awareness*. As presented in Table 5, the single-image data yields a 20.8% improvement on CV-Bench-3D [51]. In particular, the single-image data also improves performance on multi-image and video benchmarks, with improvements of +2.7% on MMSI-Bench [63] and +9.1% on VSI-Bench [61]. Based on the ability to infer spatial information from single images, we further integrate multi-image data to capture spatial relationships across diverse viewpoints and find that *cross-view spatial understanding requires multi-view consistency constraints*. The multi-image data results in a 1.3% increase in the average spatial understanding score. Notably, this approach yields a 3.9% improvement on MMSI-Bench [63] and a 4.8% improvement on BLINK [21]. Since single-image and multi-image data provide minimal improvement on video benchmarks, this suggests that *spatial ability does not emerge from general video VQA or multi-image data but needs explicit spatiotemporal supervision*. Thus, we construct video data to enhance spatial understanding for video inputs. The general video data from LLaVA-OneVision [30] exhibit limitations in visual spatial understanding tasks. By incorporating our VST-video data, the model achieves a 16.6% improvement on VSI-Bench [61].

Model	Used data	CV-2D	CV-3D	3DSR	MMSI	BLINK	VSI
InternVL2.5-8B [12]	–	74.2	80.0	55.0	29.1	55.0	40.5
SPAR-8B [70]	2.0M	72.3	89.1	57.5	-	43.9	41.1
VST-InternVL2.5-8B (**ours**)	1.3M	76.4	93.9	55.1	31.6	58.2	59.4

Table 6: Spatial tuning on InternVL2.5-8B, which is also the base model of SPAR-8B. Baseline models are in gray.

Method	Training Tokens	S-AVG	CV-2D	CV-3D	3DSR	MMSI	BLINK	VSI
Qwen2.5VL-3B (baseline)	-	49.6	71.2	72.6	50.2	26.5	47.6	29.6
Space-llava [44]	500M	50.1	69.7	75.3	49.9	26.6	47.1	32.0
SAT [43]	500M	41.9	61.2	39.3	47.9	28.7	44.6	29.9
SPAR [70]	500M	50.2	73.7	76.8	48.8	13.0	45.7	43.3
VST-P (**ours**)	500M	**56.2**	70.4	87.2	51.4	29.7	51.5	46.8

Table 7: Comparison across training data sources. We use the full datasets of Space-llava and SAT, and randomly sample 200K instances for SPAR and our VST-P. All methods are trained with 500M tokens. The baseline is highlighted in gray.

Stage	CoT	Data	Overall	Positional Relationship						Attribute		Motion		MSR
				CC	OO	RR	CO	OR	CR	M	A	C	O	-
baseline		-	26.4	22.6	28.7	17.3	39.5	38.8	28.9	20.3	18.2	21.6	28.9	25.3
CS	Num	OO	29.2	28.0	33.0	28.4	40.7	25.9	36.1	39.1	36.4	20.3	31.6	18.7
CS	RT	OO	30.0	35.5	31.9	25.9	44.2	36.5	32.5	32.8	30.3	21.6	27.6	21.2
CS	RT_{BEV}	OO	31.1	31.2	36.2	27.2	46.5	40.0	36.1	26.6	30.3	17.6	30.3	24.8
CS	RT_{BEV}	Mix	**31.7**	35.5	39.4	37.0	44.2	36.5	43.4	28.1	22.7	17.6	30.3	21.7
RL	-	Mix	**35.3**	35.5	38.3	30.9	54.6	35.3	47.0	26.6	48.5	33.8	31.6	22.7

Table 8: Exploring spatial reasoning on the MMSI-Bench [63]. 'CS' denotes the cold-start stage. CoT Type: 'Num.' denotes numerical type; 'RT' denotes reconstruction with text; 'RT_{BEV}' denotes the generated CoT is grounded on the BEV image. Data: OO refers to the object-object subset, and Mix includes all data.

Generalization beyond Qwen2.5-VL. VST is not tied to a specific base model. We use InternVL2.5-8B as the base model with the same suit as Table 5. As shown in Table 6, our data improves InternVL2.5-8B and outperforms SPAR-8B, even with fewer samples.

Clean, real-world data proves more effective for 3D spatial understanding. Existing datasets are constructed in different ways: Space-llava [44] generates samples using 3D foundation models, SAT [43] relies on a simulator, and SPAR [70] uses real-world scans but obtains 2D supervision by directly projecting 3D annotations, which can introduce noisy or misaligned labels. In contrast, VST-P is more diverse, with 97% of the data collected from real-world sources and accompanied by ground-truth 3D annotations. We further filter and correct the projected 2D annotations for the real-world portion, substantially improving label quality. As a result, VST-P achieves the best overall performance in Table 7.

The proposed RT-CoT is well-suited for Vision-Language Models. Our baseline is Qwen2.5-VL-7B [60], fine-tuned on one-third of VST-P single-image data, multi-view correspondence data, multi-image 3D object detection data, and 800K general samples. As summarized in Table 8, we first let the model represent the layout of scenes using 3D bounding boxes (Num-CoT), achieving 29.2% on MMSI-Bench. However, estimating camera poses across diverse viewpoints proves challenging, limiting the generalization. Thus, we propose reconstructing the scene using text within the CoT (RT-CoT), which outperforms Num-CoT.

Question: The images are taken continuously from a first-person perspective. At the moment of the last image, in which direction is the black chair relative to you? Options: A: Front right, B: Back left, C: Front left, D: Back right

VST: <think>To figure out the black chair's direction in the last image's perspective, I'll analyze changes between the two images. Looking at the first image, there's a black chair with curved armrests near a doorway, a large abstract painting on the wall, and dark wooden cabinets on the right. Now, in the second image, the camera has shifted. The black chair from the first image is no longer visible. Instead, the view centers on a bed with gray bedding, pillows, and a dark headboard. The dark cabinets and the large painting from the first image are on the right side of the second image, but the black chair is not in this new perspective. By comparing the two images, the camera has moved forward and to the right. Since the black chair was in the first image, when the camera shifts forward and right, the chair's position shifts out of the new camera's view. Thus, the black chair is in the direction that, from the second camera's position, is forward and to the left.</think> C. Front left

Fig. 5: Qualitative results on the MMSI-Bench [63]. The underlined text indicates reasoning hints. Purple text describes the first image, and blue text describes the second image.

A clearer reasoning trace from *prompting with BEV annotations* improves the learning of spatial reasoning, and the RL stage reinforces spatial reasoning abilities.
Furthermore, RT-CoT from prompting with BEV annotations (RT_{BEV}-CoT) yields an additional 1.1% improvement, illustrating that a high-quality CoT trace is conducive to the acquisition of spatial reasoning abilities. Finally, mixing all types of data raises the score to 31.7%. Using this resulting model as the actor, we proceed with a Reinforcement Learning (RL) phase. Upon completion of the RL stage, the model achieves a score of 35.3% on the MMSI-Bench [63], as detailed in the final row of Table 8. A qualitative example of the reasoning process is presented in Figure 5, which depicts the model first reconstructing the scene and subsequently deducing the outcome.
Ablation for the 3D object detection. We present ablation studies on 3D object detection settings in Table 9. First, FoV Unification (FoV-Uni) across different datasets yields a 2.5 AP improvement on ARKitScenes. Second, substituting Euler angles with quaternions results in reduced performance. Third, replacing all multi-turn data with an equivalent amount of single-turn data leads to a decrease of 1.7 AP on SUN RGB-D and 1.8 AP on ARKitScenes.
3D IoU combined with F1 score serves as an effective layout reward for the 3D object detection task. As recorded in Table 10, when we use 3D IoU and recall as the layout reward, performance drops markedly because each ground-truth box is matched with too many false-positive predictions. Thus, we switch to IoU and F1 score as the layout reward, yielding a 4.2 AP improvement.

5.4 Do spatial priors help with robotic learning?

As detailed in Section 4.3, we adapt our VST-tuned VLM into a VLA model. In contrast to the approach used by OpenVLA [28], *we do not utilize any pre-trained data on robotic learning*. Instead, we directly fine-tune the VLM and

Settings	SUN RGB-D					ARKitScenes				
	AP	AP_{15}	AP_{25}	AP_{50}	AP_{100}	AP	AP_{15}	AP_{25}	AP_{50}	AP_{100}
Baseline	18.5	28.5	19.1	**3.2**	**30.7**	**29.4**	**42.7**	**31.8**	**7.2**	**44.6**
w/o FoV-Uni	**18.6**	**29.1**	**19.2**	2.2	30.4	26.9	40.9	28.2	5.4	42.0
Euler to Quat	18.3	29.4	17.8	2.8	30.6	28.6	42.4	30.8	6.3	43.9
Single-turn Data	16.8	26.0	17.0	2.7	28.4	27.6	41.9	29.0	6.0	42.5

Table 9: Ablation for different settings for 3D object detection. Data: 600K 3D object detection and 800K general samples.

Layout Reward	SUN RGB-D				
	AP	**AP_{15}**	**AP_{25}**	**AP_{50}**	**AP_{100}**
Baseline	20.2	30.3	20.6	4.5	33.5
3D IoU + Recall	13.8	20.9	14.7	2.6	33.6
3D IoU + F1 Score	**24.4**	**36.0**	**26.3**	**5.8**	**37.2**

Table 10: Ablation for the layout reward $\mathcal{R}_{\text{layout}}$ on SUN RGB-D [48] 3D object detection.

VLA Backbone	Steps	Spatial	Object	Goal	10	AVG
Qwen2.5-VL-3B [3]	2.5K	56.6	86.6	53.8	15.2	53.1
VST–3B (**ours**)	2.5K	$65.0_{+8.4}$	$88.4_{+1.8}$	$67.8_{+4.0}$	$25.6_{+10.4}$	$61.7_{+\mathbf{8.6}}$
Qwen2.5-VL-3B [3]	10K	76.0	83.8	72.0	39.6	67.9
VST-3B (**ours**)	10K	$78.4_{+2.4}$	$87.8_{+4.0}$	$76.0_{+4.0}$	$41.0_{+1.4}$	$70.8_{+\mathbf{2.9}}$

Table 11: Success rate comparison on the LIBERO [32]. VST-3B is fine-tuned from Qwen2.5-VL-3B [3]

its action embeddings on the small-scale LIBERO benchmark [32] from scratch for action prediction. The results are presented in Table 11. Notably, the VLA model based on the VST-tuned model, which incorporates spatial knowledge, surpasses the one based on Qwen2.5-VL-3B [3] by an average of 8.6% in success rate. Even with a fourfold increase in training steps, the VST-tuned model still demonstrated an improvement of 2.9%. This improvement clearly demonstrates that **the integration of spatial knowledge helps with robotic learning**.

6 Conclusion

We present Visual Spatial Tuning (VST), a general and scalable framework that endows vision-language models with human-like spatial perception and reasoning abilities. With large-scale perception data (VST-P) and curated reasoning instructions (VST-R), VST acquires spatial awareness without degrading general capabilities. For spatial reasoning, we introduce RT-CoT, which reconstructs scene layouts in text and grounds reasoning on the reconstruction, and we further propose prompting with BEV annotations to improve the faithfulness of reasoning traces. The proposed approach achieves state-of-the-art performance on multiple spatial benchmarks, demonstrating that spatial abilities in foundation models can be systematically scaled. Moreover, Vision-Language-Action (VLA) models built on VST improve visuospatial skills, enabling more grounded interaction with the physical world. The generality, scalability, and effectiveness of VST highlight a promising direction for building physical AI systems that reason and act in space with human-like intelligence.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62422606) and Hong Kong Research Grant Council General Research Fund (No. 17213925).

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:2303.08774 (2023)
2. Ahmadyan, A., Zhang, L., Ablavatski, A., Wei, J., Grundmann, M.: Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In: CVPR (2021)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv:2502.13923 (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:2111.08897 (2021)
5. Bigverdi, M., Luo, Z., Hsieh, C.Y., Shen, E., Chen, D., Shapiro, L.G., Krishna, R.: Perception tokens enhance visual reasoning in multimodal language models. In: CVPR (2025)
6. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: CVPR (2023)
7. 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:2212.06817 (2022)
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: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: CVPR (2024)
10. Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R.: Shikra: Unleashing multimodal llm's referential dialogue magic. arXiv preprint arXiv:2306.15195 (2023)
11. 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? arXiv:2403.20330 (2024)
12. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv:2412.05271 (2024)
13. 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: CVPR (2024)
14. 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. NeurIPS (2024)
15. 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:2507.06261 (2025)
16. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
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. arXiv:2503.13111 (2025)

18. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR (2025)
19. Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al.: Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In: ACMM (2024)
20. 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:2505.20279 (2025)
21. 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: ECCV (2024)
22. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR (2017)
23. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., et al.: Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR (2022)
24. 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:2505.07062 (2025)
25. Hegarty, M., Montello, D.R., Richardson, A.E., Ishikawa, T., Lovelace, K.: Spatial abilities at different scales: Individual differences in aptitude-test performance and spatial-layout learning. Intelligence **34**(2), 151–176 (2006)
26. Jiang, H., Tan, H., Wang, P., Jin, H., Zhao, Y., Bi, S., Zhang, K., Luan, F., Sunkavalli, K., Huang, Q., et al.: Rayzer: A self-supervised large view synthesis model. In: ICCV (2025)
27. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: ECCV (2016)
28. Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., et al.: Openvla: An open-source vision-language-action model. arXiv:2406.09246 (2024)
29. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
30. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., Li, C.: Llava-onevision: Easy visual task transfer. arXiv:2408.03326 (2024)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
32. Liu, B., Zhu, Y., Gao, C., Feng, Y., Liu, Q., Zhu, Y., Stone, P.: Libero: Benchmarking knowledge transfer for lifelong robot learning. NeurIPS (2023)
33. 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: ECCV (2024)
34. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
35. Ma, W., Chen, H., Zhang, G., Chou, Y.C., de Melo, C.M., Yuille, A.: 3dsrbench: A comprehensive 3d spatial reasoning benchmark. arXiv:2412.07825 (2024)
36. Masry, A., Long, D.X., Tan, J.Q., Joty, S., Hoque, E.: Chartqa: A benchmark for question answering about charts with visual and logical reasoning. arXiv:2203.10244 (2022)

37. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: WACV (2021)
38. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: CVPR (2020)
39. Ouyang, K., Liu, Y., Wu, H., Liu, Y., Zhou, H., Zhou, J., Meng, F., Sun, X.: Spacer: Reinforcing mllms in video spatial reasoning. arXiv:2504.01805 (2025)
40. Piaget, J.: Child's Conception of Space: Selected Works vol 4. Routledge (2013)
41. Ranasinghe, K., Shukla, S.N., Poursaeed, O., Ryoo, M.S., Lin, T.Y.: Learning to localize objects improves spatial reasoning in visual-llms. In: CVPR. pp. 12977–12987 (2024)
42. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
43. Ray, A., Duan, J., Brown, E., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B.A., Krishna, R., et al.: Sat: Dynamic spatial aptitude training for multimodal language models. arXiv:2412.07755 (2024)
44. RemyxAI: vqasynth spacellava (2024), https://huggingface.co/datasets/remyxai/vqasynth_spacellava
45. 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: ICCV (2021)
46. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv:2402.03300 (2024)
47. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: CVPR (2019)
48. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: CVPR (2015)
49. Team, C.: Chameleon: Mixed-modal early-fusion foundation models. arXiv:2405.09818 (2024)
50. Tian, X., Gu, J., Li, B., Liu, Y., Wang, Y., Zhao, Z., Zhan, K., Jia, P., Lang, X., Zhao, H.: Drivevlm: The convergence of autonomous driving and large vision-language models. arXiv:2402.12289 (2024)
51. 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. NeurIPS (2024)
52. Wang, H., Chen, J., Huang, W., Ben, Q., Wang, T., Mi, B., Huang, T., Zhao, S., Chen, Y., Yang, S., et al.: Grutopia: Dream general robots in a city at scale. arXiv:2407.10943 (2024)
53. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
54. 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:2409.12191 (2024)
55. Wu, W., Mao, S., Zhang, Y., Xia, Y., Dong, L., Cui, L., Wei, F.: Mind's eye of llms: visualization-of-thought elicits spatial reasoning in large language models. NeurIPS (2024)
56. x.ai: Grok-1.5 vision preview (2024), https://x.ai/blog/grok-1.5v

57. 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. NeurIPS (2024)
58. 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:2505.17015 (2025)
59. 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:2505.17015 (2025)
60. Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., et al.: Qwen2 technical report. arXiv:2407.10671 (2024)
61. 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: CVPR (2025)
62. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. NeurIPS (2024)
63. 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:2505.23764 (2025)
64. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
65. 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. arXiv:2506.21458 (2025)
66. 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: CVPR (2024)
67. Yue, Z., Lin, Z., Song, Y., Wang, W., Ren, S., Gu, S., Li, S., Li, P., Zhao, L., Li, L., et al.: Mimo-vl technical report. arXiv preprint arXiv:2506.03569 **5** (2025)
68. Zhang, C., Cui, Z., Zhang, Y., Zeng, B., Pollefeys, M., Liu, S.: Holistic 3d scene understanding from a single image with implicit representation. In: CVPR (2021)
69. Zhang, H., You, H., Dufter, P., Zhang, B., Chen, C., Chen, H.Y., Fu, T.J., Wang, W.Y., Chang, S.F., Gan, Z., et al.: Ferret-v2: An improved baseline for referring and grounding with large language models. arXiv preprint arXiv:2404.07973 (2024)
70. Zhang, J., Chen, Y., Zhou, Y., Xu, Y., Huang, Z., Mei, J., Chen, J., Yuan, Y.J., Cai, X., Huang, G., et al.: From flatland to space: Teaching vision-language models to perceive and reason in 3d. arXiv:2503.22976 (2025)
71. 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:2504.10479 (2025)
72. Zhu, Z., Ma, X., Chen, Y., Deng, Z., Huang, S., Li, Q.: 3d-vista: Pre-trained transformer for 3d vision and text alignment. In: ICCV (2023)
73. Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., et al.: Rt-2: Vision-language-action models transfer web knowledge to robotic control. In: CoRL (2023)