SpatialReasoner: Towards Explicit and Generalizable 3D Spatial Reasoning

Wufei Ma* Yu-Cheng Chou* Qihao Liu*
Xingrui Wang Celso M de Melo† Jianwen Xie° Alan Yuille

Johns Hopkins University, †DEVCOM Army Research Laboratory, ° Lambda Inc

Abstract

Despite recent advances on multi-modal models, 3D spatial reasoning remains a challenging task for state-of-the-art open-source and proprietary models. Recent studies explore data-driven approaches and achieve enhanced spatial reasoning performance by fine-tuning models on 3D-related visual question-answering data. However, these methods typically perform spatial reasoning in an implicit manner and often fail on questions that are trivial to humans, even with long chain-of-thought reasoning. In this work, we introduce SpatialReasoner, a novel large vision-language model (LVLM) that addresses 3D spatial reasoning with explicit 3D representations shared between multiple stages–3D perception, computation, and reasoning. Explicit 3D representations provide a coherent interface that supports advanced 3D spatial reasoning and improves the generalization ability to novel question types. Furthermore, by analyzing the explicit 3D representations in multi-step reasoning traces of SpatialReasoner, we study the factual errors and identify key shortcomings of current LVLMs. Results show that our SpatialReasoner achieves improved performance on a variety of spatial reasoning benchmarks, outperforming Gemini 2.0 by 9.2% on 3DSRBench, and generalizes better when evaluating on novel 3D spatial reasoning questions. Our study bridges the 3D parsing capabilities of prior visual foundation models with the powerful reasoning abilities of large language models, opening new directions for 3D spatial reasoning. Our project page is available here.

1 Introduction

3D spatial reasoning studies how models perceive, understand, and reason about 3D object properties and spatial relationships. It is not only a fundamental task for vision-language models to achieve human-level intelligence, but also crucial to a range of downstream applications in robotics [14, 18] and embodied AI [10]. Despite the recent advancements of large multi-modal models, such as GPT-4o and Qwen2.5-VL, their 3D spatial reasoning capabilities remain limited and fall far behind human-level performance [9, 31, 51]. Recent 3D-aware large vision-language models (LVLMs) injected 3D knowledge by fine-tuning the model on synthetic 3D-related question-answer pairs [9, 33] and achieved improved performance on spatial reasoning benchmarks [31, 45, 51]. On the other hand, large proprietary models such as Gemini 2.0 [15] have advanced 3D parsing by directly predicting 3D object bounding boxes, enabling the development of powerful generalist robotics models [43].

We identify two key challenges in 3D spatial reasoning: (1) *3D thinking*–the ability to decompose a complex 3D spatial reasoning question into small, manageable steps, and (2) *3D computation*–the ability to solve these thinking steps in a consistent and accurate manner. As shown in Figure 1, prior reasoning methods adopt long chain-of-thought reasoning to tackle the problem but do not

*Equal first author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

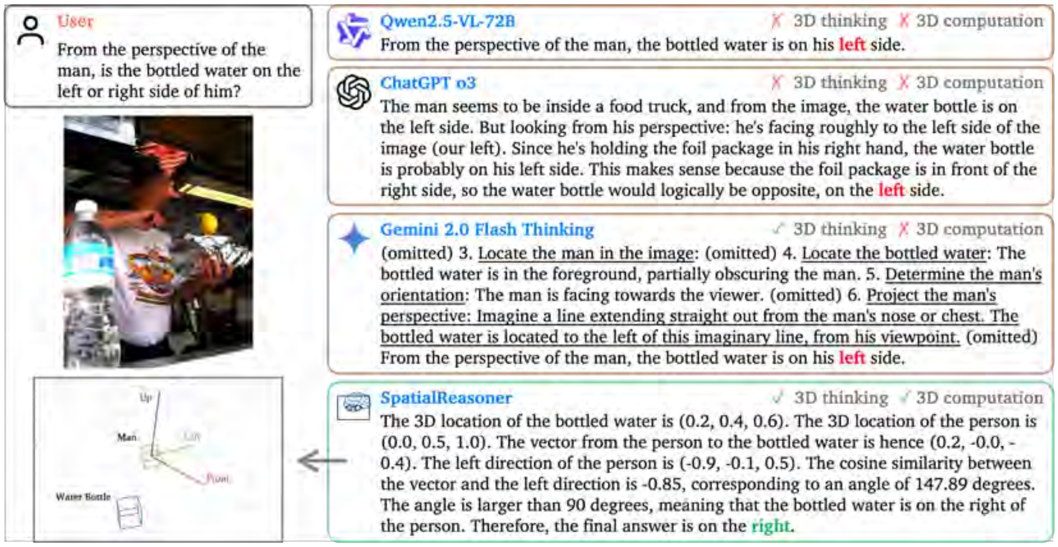


Figure 1: **Comparing 3D spatial reasoning of our SpatialReasoner with previous state-of-the-art models.** Our SpatialReasoner builds on explicit 3D representations, performs 3D computation, and reasons about the final answer. Although Gemini 2.0 can also break down complex 3D spatial reasoning questions into small and tractable steps, it lacks reliable 3D computation and the long chain-of-thought reasoning ultimately leads to a wrong answer.

have explicit 3D computation. ChatGPT o3 [36] fails to adopt a systematic approach to solving the problem and instead relies on other visual cues to assist the reasoning (*e.g.*, location of the foil package). In contrast, the Gemini 2.0 thinking model [15] employs an organized strategy to tackle the problem but ultimately fails to arrive at the correct answer due to a lack of reliable 3D computation.

In this work, we present SpatialReasoner, a novel large vision-language model (LVLM) built with (1) explicit 3D representations and (2) enhanced and generalizable 3D thinking. Specifically, our SpatialReasoner adopts explicit 3D representations, such as 3D locations and orientations, as an interface that enables coherent and reliable reasoning across multiple stages, *i.e.*, 3D perception, computation, and reasoning. On the other hand, we would like to learn enhanced 3D thinking capabilities that generalize to novel question types not seen during training. Hence we explore a two-stage training strategy following prior works [17]. In Stage I, we apply supervised fine-tuning (SFT) to equip the LVLM with explicit 3D representations, enhancing 3D perception and computation capabilities of the model. Then in Stage II, we leverage reinforcement learning (RL) to develop robust and generalizable 3D thinking built on explicit 3D representations.

Inspired by previous 3D pseudo-annotation pipelines in [9, 11, 33], we synthesize basic 3D perception, *i.e.*, detection and pose estimation, and 3D computation question-answering data interleaved with explicit 3D representations. We further generate standard spatial reasoning question-answer pairs with chain-of-thought reasoning that breaks down complex 3D spatial reasoning questions into multiple steps–3D perception, computation, and reasoning. Experimental results demonstrate that our SpatialReasoner with explicit 3D representations can significantly enhance 3D spatial reasoning abilities of LVLMs and generalize to novel question types.

Besides enhancing 3D spatial reasoning capabilities of LVLMs, reasoning with explicit 3D representations allows us to interpret the reasoning process and to study the failure modes of LVLMs. We find that the accuracy of 3D perception lags significantly behind that of 3D computation, suggesting that most errors in downstream VQA tasks still stem from failures in 3D perception. Moreover, by predicting key 3D information, such as 3D object locations and orientations, as intermediate results, our SpatialReasoner enables compositional reasoning for 3D spatial tasks [46]. This not only allows our method to generalize better to novel spatial reasoning questions, but also makes it easily extensible to other tasks that build on our explicit 3D representations.

Besides improving spatial reasoning performance on a variety of benchmarks, we experiment on various LVLMs fine-tuned with different combinations of data and training methods to study the key factors toward improved 3D spatial reasoning. Our empirical results lead to the following insights: (1) For 3D-aware VLMs, SFT offers a more scalable approach than RL that requires high-quality 3D-aware data, which is often difficult to obtain. Our recipe of RL followed by SFT achieves the best overall performance; (2) 3D-aware LVLMs fine-tuned with RL generalize better than SFT when tested on novel 3D spatial reasoning questions, echoing prior findings [12]; (3) Standard LVLMs often exploit 2D reasoning as a shortcut to tackle 3D spatial reasoning problems, whereas our SpatialReasoner avoids the spurious correlations and always reasons with explicit 3D representations, achieving improved and robust performance on challenging real-world 3D spatial reasoning datasets.

In summary, our contributions are as follows: (1) We introduce SpatialReasoner, a novel LVLM that solves 3D spatial reasoning problems with a compositional approach based on explicit 3D representations. (2) Reasoning with 3D representations allows us to interpret the reasoning process and to study the failure modes of LVLMs. (3) Our SpatialReasoner effectively improves the 3D spatial reasoning performance on a range of benchmarks and extensive experimental results provide valuable findings on data and training strategy designs for future development of 3D-aware LVLMs.

2 Related Works

3D Spatial reasoning. 3D spatial reasoning explores how models perceive and reason about 3D object properties and relationships. Early works [46, 47, 48] studied this in simulated environments or datasets with object-level annotations [20, 25, 7, 11, 45, 33]. More recent efforts construct benchmarks based on real-world imagery [31, 51] and improve model performance via synthetic QA data [9, 11, 33, 39]. However, these models often rely on implicit reasoning, offering little interpretability or intermediate 3D computation. Moreover, it remains unclear whether the 3D spatial reasoning capabilities acquired from data-driven fine-tuning can generalize to novel question types that require complex 3D computation over different combinations of 3D perception outputs.

Explicit 3D representations. Explicit 3D representations simplify reasoning and expose model failures. Simulation-based works [35] used neural-symbolic methods to parse object-level structures, while PO3D [46] and its extensions [47, 48] showed that structured visual modules enable interpretable reasoning [32]. These results in simulation systems reveal the limitations of current LVLMs and highlight the importance of the visual module for a successful reasoning model.

Post-training. Post-training aligns pre-trained models with downstream objectives via SFT [49, 54] and RL [5, 6, 37, 41], improving formatting and reward-guided alignment [21, 44]. Recent systems like GPT-4 [1], Claude-3.5 [2], and DeepSeek-R1 [17] showcase these techniques. Despite recent advances, post-trained models still struggle with generalization under distribution shifts and novel tasks. While SFT stabilizes outputs but often overfits, RL improves adaptability [12], and combining both shows promise [17], making post-training vital for robust, aligned LLMs.

Test-time scaling. Scaling inference-time compute improves performance without retraining. Approaches include beam search [30, 16], best-of-N sampling [42], and MCTS [13], alongside prompting methods like CoT [50] and ToT [53]. Strategically allocated test-time compute has proven effective for enhancing reasoning in both unimodal and multimodal models. Building on this, we apply CoT reasoning to 3D spatial reasoning by fine-tuning models to generate step-by-step rationales, providing both accurate and interpretable results.

3 SpatialReasoner

3.1 Overview

In this section, we introduce our SpatialReasoner for explicit and generalizable 3D spatial reasoning. Our SpatialReasoner features two key designs: (1) explicit 3D representations that serves as interface to support multi-stage spatial reasoning, *i.e.*, 3D parsing, computation, and reasoning (see Figure 2), and (2) generalizable spatial reasoning from multi-stage training (see Figure 4).

In Section 3.2, we present the explicit 3D representations and describe how our model is trained to predict and to interpret the 3D representations for spatial reasoning. Then in Section 3.3 we discuss our training strategies, exploring standard supervised fine-tuning, reinforcement learning, as well as

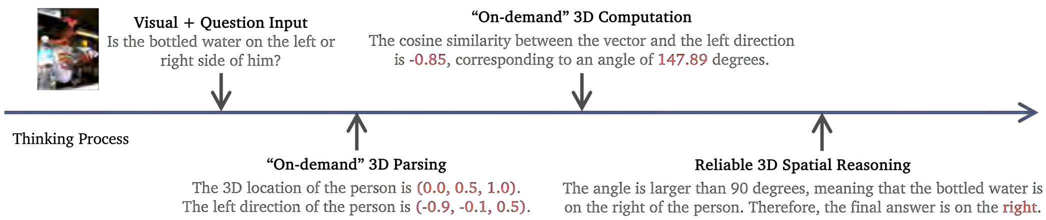


Figure 2: **Overview of our SpatialReasoner design.** Our SpatialReasoner adopts explicit 3D representations as an interface to enable coherent and reliable multi-stage spatial reasoning, *i.e.*, 3D parsing, computation, and reasoning.

3D-aware process rewards. Lastly we introduce our 3D-aware data generation pipeline and different variants of training data used at different stages in Section 3.4.

3.2 Learning Explicit 3D Representations

Despite improved spatial reasoning abilities achieved by 3D-aware VLMs, such as SpatialRGPT [11] and SpatialLLM [33], and advanced proprietary models like Gemini 2.0 [15], these methods lack explicit 3D representations and rely on natural language to perform 3D spatial reasoning. For example, Gemini 2.0 describes object poses as "facing towards the viewer and slightly to its right" and estimates 3D distances with phrases such as "far behind some other object". Such natural language descriptions are inefficient and often not accurate enough for complex 3D spatial reasoning.

Therefore, we propose integrating LVLMs with explicit 3D representations, such as 3D locations and poses, to serve as an accurate and reliable interface shared across stages of 3D spatial reasoning (see Figure 2). Our SpatialReasoner can predict explicit 3D representations as intermediate results or take them as inputs to perform basic 3D computations or complex spatial reasoning tasks.

Explicit 3D representations. We define explicit 3D representations in a calibrated camera 3D space, which is a standard camera 3D space calibrated with the extrinsics of the camera. As illustrated in Figure 3, the calibrated camera 3D space has its z-axis aligned with the z-axis of the 3D world space, and the origin on the z-axis is positioned close to the ground plane. Although estimating 3D object locations is easier in original camera 3D space, adopting explicit 3D representations in calibrated camera 3D space offers many advantages for subsequent spatial reasoning: (1) the z-coordinates directly correspond to object heights, (2) estimating 3D spatial relationships such as "above" and "below" is largely simplified, and (3) objects often are on a plane parallel to the ground plane, which reduces many 3D spatial relationships to simpler 2D problems.

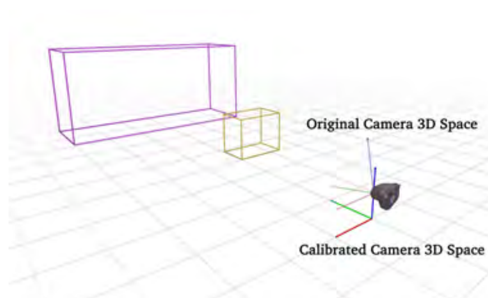


Figure 3: **Comparison between original and calibrated camera 3D space.** Our explicit 3D representations are defined within calibrated camera 3D space that simplifies subsequent 3D computation and reasoning.

A unified interface for explicit 3D spatial reasoning. Explicit 3D representations serve as an interface enabling coherent and accurate 3D spatial reasoning across stages (see Figure 2). For 3D perception, the model predicts object locations and orientations as 3D vectors, then estimates explicit distances or angles based on these predictions. Finally, the model aggregates explicit 3D information from earlier stages to reason about and answer 3D spatial questions.

3.3 Training Strategies

Supervised fine-tuning (SFT). We adopt a two-stage post-training strategy to equip the model with explicit, generalizable 3D spatial reasoning capabilities. In the first stage, SFT serves as a critical initialization step, aligning the pre-trained LVLM with curated 3D-annotated datasets. By

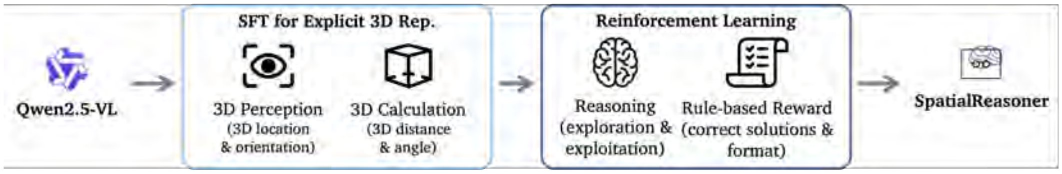


Figure 4: **Overview of our SpatialReasoner training.** We adopt a multi-stage training strategy: In Stage I, we apply supervised fine-tuning (SFT) to equip the LVLM with explicit 3D representations; then in Stage II, we leverage reinforcement learning (RL) to develop robust and generalizable 3D spatial reasoning built on explicit 3D representations.

optimizing a maximum likelihood objective over paired input-output sequences, the model learns to identify objects in 3D scenes and predict explicit 3D representations such as object locations and orientations. This structured supervision enables the model to expose interpretable intermediate reasoning traces during spatial computations—*e.g.*, calculating relative distances and angles. However, due to its reliance on static demonstrations, SFT shows limited capacity to generalize beyond observed reasoning patterns. As shown in Section 4.3, SFT-trained models tend to memorize spatial templates in the training set and struggle with novel compositions or combinatorial variations in 3D queries.

Reinforcement learning (RL). To overcome this limitation, we further post-train the SFT-initialized model using reinforcement learning (RL) with rule-based rewards. Treating the spatial reasoning task as a sequential decision process, each reasoning step is framed as an action within a Markov Decision Process (MDP), and policy gradients are optimized using GRPO [41]. Importantly, we integrate a reward scheme that provides structured reward signals reflecting both the correctness of final answers and the coherence of intermediate 3D computation steps. This enables multi-turn optimization where the model learns to revise inaccurate inferences and explore alternative reasoning paths. Empirically, this RL stage substantially improves generalization to out-of-distribution 3D spatial questions (Table 3), particularly in settings involving multi-object arrangements. The combined SFT+RL strategy therefore balances response formatting and perceptual grounding with adaptive reasoning and robustness, making it especially suited for complex 3D spatial understanding.

Reward and policy optimization design. To train SpatialReasoner via reinforcement learning, we design a composite reward scheme capturing both answer correctness and reasoning quality. For outcome rewards, we use an accuracy reward—aligned with the multiple-choice evaluation metric from MMBench [28]—which gives positive feedback only when the model selects the correct answer, and a format reward [17] that encourages structured, readable outputs. We also explore process rewards to assess whether models can learn structured reasoning without SFT. These include a reasoning steps reward, which promotes use of structured indicators (*e.g.*, “First”, “Next”), and a 3D-aware process reward that checks for necessary spatial terms like distance or orientation. The reward signal is then calculated through the accuracy of the presence of each term. Though not used in the final model, these rewards serve as diagnostics for emergent reasoning behavior. Finally, we ablate the KL divergence term in GRPO and find that it harms training accuracy despite stabilizing output lengths (Section 4.3). We thus remove it in our final setting.

3.4 Training Data

To enable LVLMs to predict and reason with explicit 3D representations and to post-train LVLMs to solve various challenging 3D spatial reasoning questions, we generate a series of 3D-aware training data. We extend the data generation pipeline in [9, 11, 33]. Our process begins with generating 3D pseudo-annotations, followed by optional human verification, and ends with constructing various VQAs and chain-of-thought reasoning based on the 3D pseudo-annotations.

Pseudo 3D ground-truths. We extend the 3D pseudo-annotation pipeline proposed in [9, 33] and generate various 3D annotations, such as object category, 3D location, and 3D pose, on unlabeled images from the OpenImages dataset [22]. Based on the object-level 3D annotations, we then apply rule-based methods to derive ground-truth labels for a range of 3D spatial relationships. Despite significant progress in visual foundation models for segmentation [38], metric depth estimation [52], and object pose estimation [34], we notice many factual errors in the generated 3D ground-truth,

Method	Mean	Height	Location	Orientation	Multi-Object
Open-Sourced Generalist					
LLaVA-v1.5-7B [26]	38.1	39.1	46.9	28.7	34.7
LLaVA-Next-8B [24]	48.4	50.6	59.9	36.1	43.4
Cambrian-1-8B [45]	42.2	23.2	53.9	35.9	41.9
Qwen2.5-VL-3B-Instruct [4]	43.9	45.2	56.8	35.7	35.7
Qwen2.5-VL-7B-Instruct [4]	48.4	44.1	62.7	40.6	40.5
Qwen2.5-VL-72B-Instruct [4]	54.9	53.3	71.0	43.1	46.6
Open-Sourced Specialist					
SpaceLLaVA [40]	42.0	49.3	54.4	27.6	35.4
SpatialBot [8]	41.0	40.4	54.4	31.9	33.5
SpatialLLM [33]	44.8	45.8	61.6	30.0	36.7
SpatialRGPT [11]	32.7	**55.9**	39.0	27.8	20.0
SpatialRGPT *w/* depth [11]	48.4	55.9	60.0	34.2	42.3
Proprietary					
GPT-4o-mini	39.7	44.3	52.4	21.0	36.5
GPT-4o	44.2	53.2	59.6	21.6	39.0
Claude 3.5 V Sonnet	48.2	<u>53.5</u>	63.1	31.4	41.3
Gemini 2.0 Flash	49.8	49.7	68.9	32.2	41.5
Gemini 2.0 Flash (thinking)	51.1	53.0	67.1	35.8	43.6
QwenVLMax	52.0	45.1	70.7	37.7	44.8
Ours					
SpatialReasoner-Zero	54.0	46.4	67.3	48.4	47.2
SpatialReasoner-SFT	<u>58.3</u>	51.9	<u>73.5</u>	<u>50.7</u>	<u>50.3</u>
SpatialReasoner	**60.3**	52.5	**75.2**	**55.2**	**51.8**

Table 1: **Comparison with previous state-of-the-art methods on 3DSRBench [31].** Our SpatialReasoner outperforms previous open-source and proprietary methods on challenging 3D spatial reasoning problems in 3DSRBench.

particularly when mistakes (*e.g.*, missing objects or inaccurate object poses) propagate through the data pipeline. Therefore, we adopt a series of aggressive filtering steps to ensure the quality of our training data, including: (1) removing images with densely cluttered scenes, (2) excluding object categories that are difficult to segment or estimate pose for, and (3) discarding boundary cases that could lead to ambiguity.

Human verification. Despite leveraging state-of-the-art visual foundation models in our data generation pipeline and applying multiple filtering strategies, the 3D pseudo-annotations remain susceptible to factual errors. Specifically, small issues such as missing objects or inaccurate 3D pose predictions propagate through later stages of the pipeline, leading to factual errors in the final spatial relationship pseudo-annotations. To assess the impact of data quality, we create a smaller but higher-quality dataset by manually verifying the correctness of the 3D pseudo-annotations.

Training data variants. Based on the obtained 3D ground-truth, we can generate different variants of data for fine-tuning. Specifically, we consider the following: (1) *Basic3D-QA* consists of basic 3D perception and 3D computation question-answering data. This can be used to learn explicit 3D representations without training on various 3D spatial relationships considered in downstream tasks. (2) *SR-QA* contains visual question-answering pairs about various 3D spatial relationships, following previous 3D-aware datasets [11, 33]. (3) *SR-CoT* extends SR-QA and comprises chain-of-thought reasoning with explicit 3D representations. Questions are answered in a step-by-step manner.

4 Results

4.1 Experimental Setup

Baselines. We compare our SpatialReasoner with the following three types of baseline models. (1) *Open-sourced generalists*: such as LLaVA [26], Cambrian-1 [45], and Qwen2.5 [4] that are trained on

Method	CV-Bench-3D Depth	Distance	GQA Mean	Choose	Compare	Logical	Query	Verify
Qwen2.5-VL-7B-Instruct	82.5	**83.2**	58.8	82.7	71.5	78.9	40.6	82.5
SpatialReasoner-Zero	77.5	<u>81.8</u>	60.2	81.6	67.9	<u>79.2</u>	43.8	82.1
SpatialReasoner-SFT	<u>85.2</u>	71.5	**62.0**	<u>82.8</u>	<u>72.2</u>	**81.4**	**45.5**	**83.2**
SpatialReasoner	**87.3**	73.3	<u>61.8</u>	**83.2**	**81.1**	71.8	<u>45.2</u>	<u>82.9</u>

Table 2: **Performance on CV-Bench-3D and GQA.** Our SpatialReasoner also improves the spatial reasoning performance on GQA [19] and depth-related questions in CVBench-3D [45]. For distance-related questions, unlike Qwen2.5 that exhibits excessive dependence on 2D shortcuts, our SpatialReasoner employs rigid 3D spatial reasoning and achieves compelling performance. Regarding the performance on distance-related questions, see Section 4.2 and Section C for detailed discussions.

Method	In-Distribution Mean	Height	Location	Orientation	Novel Multi-Object
Qwen2.5-VL-7B-Instruct	48.4	44.1	62.7	40.6	40.5
SpatialReasoner-Zero	<u>53.7</u>	40.6	68.4	<u>50.2</u>	**46.6**
SpatialReasoner-SFT	52.2	<u>44.9</u>	<u>69.5</u>	48.9	40.0
SpatialReasoner	**56.4**	**52.5**	**72.6**	**54.1**	<u>43.4</u>

Table 3: **Evaluation of generalization ability by finetuning on simpler (in-distribution) 3D spatial reasoning questions and evaluating on complex (novel) questions types unseen during training.** Our SpatialReasoner-Zero and SpatialReasoner demonstrates superior zero-shot generalization, indicating RL fosters more robust and flexible reasoning than SFT.

general vision-language data. (2) *Open-sourced specialists*: We evaluate SpaceLLaVA [40] (a public re-implementation of SpatialVLM [9]), SpatialBot [8] that enhances fine-grained spatial reasoning and robot control with RGB and depth images, and SpatialLLM [33] that fine-tunes a LLaVA model with multi-stage 3D-informed training. Note for fair comparison, we evaluate SpatialBot with RGB inputs only. (3) *Proprietary models* such as GPT-4o [1] and Claude 3.5 [3] that were trained on abundant web-scale data and for Gemini 2.0 [15], additional 3D-aware post-training.

Evaluation benchmarks. We evaluate spatial reasoning abilities of various models on three spatial reasoning benchmarks. *3DSRBench* [31] is a comprehensive 3D spatial reasoning benchmark with 2,100 questions and studies various 3D awareness and reasoning abilities with a robust evaluation setup. *CVBench* [45] is a vision-centric benchmark that assesses models at classic vision tasks with a range of 2D and 3D understanding VQAs. In this work, we focus exclusively on 3D-related questions, *i.e.*, CVBench-3D, as 2D left-right relationships can lead to ambiguity with 3D left-right questions considered in 3DSRBench. *GQA* [19] is a widely adopted benchmark that studies visual reasoning and compositional question answering on a range of spatial relationships between objects.

4.2 Advancing 3D Spatial Reasoning

We evaluate SpatialReasoner on 3DSRBench [31], CVBench-3D [45], and GQA [19] to assess 3D perception, computation, and reasoning. As shown in Table 1, SpatialReasoner achieves a new state-of-the-art 60.3% mean accuracy on 3DSRBench, outperforming Gemini 2.0 Flash (49.8%) and Claude 3.5 Sonnet (48.2%). It shows notable improvements on Location (+6.3%) and Orientation (+14.6%) questions, and achieves 51.8% (+8.2%) on Multi-Object reasoning relative to the second-best model, indicating stronger 3D perception and complex spatial understanding.

On CVBench-3D and GQA (Table 2), SpatialReasoner achieves 87.3% on depth-related questions in CVBench-3D, and improved performance on GQA's Compare category, further supporting its multi-object reasoning capability. These gains validate our multi-stage training: SFT equips the model with explicit 3D representations, while RL enhances adaptive reasoning and generalization.

Despite this, we observe a performance drop on CVBench-3D distance questions, in contrast to substantial gains on similar 3DSRBench cases ("multi-object-closer-to", 34.3% $\rightarrow$ 70.9%). We attribute

Method	Mean
Qwen2.5-VL-7B-Instruct	48.4
SR-SFT	58.3
SR-SFT (*w/o* exp. 3D rep.)	51.9

Table 4: **Comparisons between SpatialReasoner with and without explicit 3D representations and CoT reasoning.** Results highlight the benefits of explicit 3D representations as an interface to support enhanced 3D spatial reasoning.

Method	Mean
Qwen2.5-VL-7B-Instruct	48.4
SpatialReasoner-SFT	58.3
SpatialReasoner-SFT (+HQ SFT)	54.7
SpatialReasoner-Zero	54.0
SpatialReasoner-Zero (*w/* KL)	52.4
SpatialReasoner-Zero (*w/* 3D Rwd)	54.6

Table 5: **Ablation study on various design choices in RL and SFT.** Notably with 3D-aware rewards, SpatialReasoner-Zero produces coherent chain-of-thought reasoning on explicit 3D representations and improves benchmark performance.

this gap to shortcut-driven 2D biases in CVBench-3D. Unlike existing LVLMs, SpatialReasoner relies on explicit 3D reasoning, offering greater robustness to such spurious cues (Section C).

4.3 Analyses and Findings

Generalization abilities. While SFT stabilizes outputs, it often overfits training distributions, limiting generalization to out-of-distribution (OOD) variations [12]. In contrast, outcome-based RL promotes transferable reasoning strategies and perceptual adaptability. To examine this, we compare SpatialReasoner-SFT (SFT-only), SpatialReasoner-Zero (RL-only), and SpatialReasoner (SFT+RL) on multi-object reasoning, with all multi-object data removed during training.

As shown in Table 3, SpatialReasoner-Zero achieves 46.6%, outperforming SpatialReasoner-SFT (40.0%) and SpatialReasoner (43.4%), confirming RL's superior zero-shot generalization ability. Although SFT+RL improves over SFT, it does not fully match RL-only performance, indicating that RL fosters more flexible reasoning than memorization-prone SFT.

When multi-object training data are available (Table 1), SFT attains 50.3% via pattern matching, but drops to 40.0% when data are withheld, whereas RL-based models retain stable performance (46.6% vs. 47.2%). Mean accuracy further supports this: SpatialReasoner-Zero maintains performance (53.7% vs. 54.0%), while SpatialReasoner-SFT degrades significantly (58.3% to 52.2%), reinforcing that RL enables more robust generalization.

Compositional reasoning. We assess whether structured 3D reasoning can emerge from process rewards alone, without curated CoT. As shown in Figure 8, SpatialReasoner-Zero trained with 3D-aware rewards generates coherent multi-step reasoning grounded in 3D cues, improving interpretability over the baseline relying on shallow shortcuts. This demonstrates the model's ability to synthesize novel compositional reasoning strategies from basic 3D perception and computation skills. Quantitatively, it achieves a performance gain from 54.0% to 54.6% (Table 5), indicating that outcome-driven reward shaping can induce coherent reasoning without human-crafted CoT. Nonetheless, for the final SpatialReasoner, we use SFT to guide reasoning traces, as it outperforms process rewards alone.

Scaling of training computation. We ablate on the scaling of training computation for SpatialReasoner-SFT and SpatialReasoner in Figure 5. Results show that with more training steps, SpatialReasoner-SFT starts to overfit and exhibit decreased performance, while SpatialReasoner trained with RL retains a stable and competitive performance.

Scaling of training data. 3D-related tasks often lack high-quality data with 3D (pseudo-)annotations due to the cost and expertise required. While synthetic or pseudo-labeled data is more scalable, they introduce noise and domain gaps. To study data scaling, we mix 1.2K verified and 24K unverified samples and compare SpatialReasoner with and without RL. As shown in Figure 6, RL performs best with only the 1.2K verified data, whereas SpatialReasoner-SFT benefits from more unverified data. This suggests that SFT is more tolerant to noisy pseudo-annotations, while RL favors fewer but higher-quality samples.

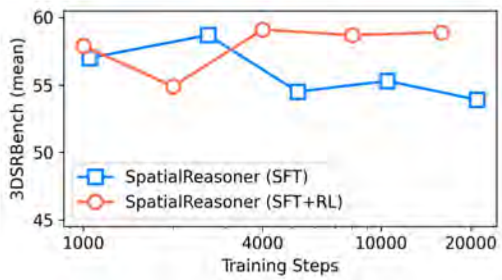


Figure 5: **Scaling of training computation.** SpatialReasoner-SFT may overfit to training data while SpatialReasoner trained with RL retains a stable and competitive performance.

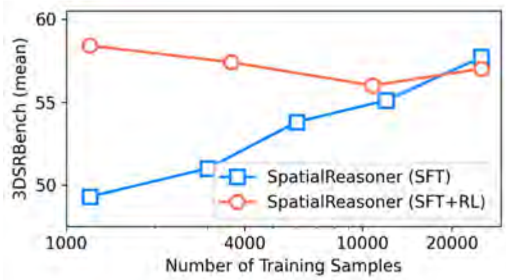


Figure 6: **Scaling of training data by mixing 1.2K human-verified with 24K unverified data.** Results show SFT remains a compelling choice when scaling with potentially noisy training data.

Ablation study on explicit 3D representations. To study the importance of explicit 3D representations and step-by-step 3D computations, we train a variant of SpatialReasoner on the same data but without explicit 3D representations. Results in Table 4 show that explicit 3D representations can enhance 3D spatial reasoning by a wide margin.

Ablation study on KL divergence. We evaluate KL divergence regularization in GRPO and find it harms learning in our setting. As shown in Figure 9, adding KL causes training to collapse, while removing it yields stable improvements. The shorter completions without KL may reflect mitigated length bias [29], improving token efficiency. As shown in Table 5, SpatialReasoner-Zero reaches 54.0% accuracy without KL vs. 52.4% with it. We thus omit KL for better spatial reasoning.

Ablation Study on SFT+RL vs. SFT+SFT To verify the performance gains are not merely from additional data exposure, we compare sequential SFT+RL against SFT+SFT. From Table 5, while SpatialReasoner-SFT achieves 58.3% accuracy, applying a second round of SFT (SFT+SFT) degrades performance to 54.7%. In contrast, RL after SFT (SFT+RL) improves accuracy to 60.3%. These results indicate repeated SFT exacerbates overfitting to training patterns, whereas RL effectively builds on the structured outputs from SFT to promote adaptive and generalizable 3D spatial reasoning.

4.4 Interpreting Failure Modes

Explicit 3D representations enhance both spatial reasoning and interpretability [46]. We categorize 3D spatial reasoning into two stages: 3D perception (parsing key 3D information from the image input) and 3D reasoning (computing 3D metrics and deriving final answer). To quantitatively analyze model behavior, we manually verify 300 questions with human-verified 3D answers. For 3D perception, we evaluate questions on object location and orientation; for 3D reasoning, models estimate distances, depths, and angles from 3D inputs, with accuracy (within $\pi/6$) for angles and mean error for positions. As shown in Table 7, models reason more accurately than they perceive–indicating most VQA errors stem from flawed 3D perception, consistent with our qualitative visualizations. Additionally, RL slightly degrades perception and reasoning accuracy, likely due to rewards focusing only on final answers and reasoning format, not intermediate 3D outputs.

5 Conclusions

In this paper, we presented SpatialReasoner that performs explicit and generalizable 3D spatial reasoning by predicting intermediate 3D representations across perception, computation, and reasoning stages. Our two-stage post-training pipeline with SFT followed by RL achieves state-of-the-art results on multiple benchmarks while generalizing well to novel tasks. By analyzing model behavior, we show that explicit 3D reasoning improves both accuracy and interpretability, and reveals key limitations in 3D perception. This work highlights the value of structured reasoning and adaptive training for robust 3D understanding, paving the way for future research in multimodal spatial reasoning.

Acknowledgements. W.M. and A.Y. acknowledges support from the Office of Naval Research with N00014-23-1-2641 and ARL award W911NF2320008.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anthropic. The claude 3 model family: Opus, sonnet, haiku.

[3] Anthropic. Claude 3.5 sonnet.

[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.

[5] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[6] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

[7] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3D: A large benchmark and model for 3D object detection in the wild. In *CVPR*, Vancouver, Canada, June 2023. IEEE.

[8] Wenxiao Cai, Yaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*, 2024.

[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.

[10] An-Chieh Cheng, Yandong Ji, Zhaojing Yang, Zaitian Gongye, Xueyan Zou, Jan Kautz, Erdem Bıyık, Hongxu Yin, Sifei Liu, and Xiaolong Wang. Navila: Legged robot vision-language-action model for navigation. *arXiv preprint arXiv:2412.04453*, 2024.

[11] 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. *arXiv preprint arXiv:2406.01584*, 2024.

[12] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. Sft memorizes, rl generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

[13] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In *International conference on computers and games*, pages 72–83. Springer, 2006.

[14] Jensen Gao, Bidipta Sarkar, Fei Xia, Ted Xiao, Jiajun Wu, Brian Ichter, Anirudha Majumdar, and Dorsa Sadigh. Physically grounded vision-language models for robotic manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024.

[15] Google. Introducing gemini 2.0: our new ai model for the agentic era, 2024. Accessed: Dec 2024.

[16] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012.

[17] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[18] Wenlong Huang, Chen Wang, Yunzhu Li, Ruohan Zhang, and Li Fei-Fei. Rekep: Spatio-temporal reasoning of relational keypoint constraints for robotic manipulation. *arXiv preprint arXiv:2409.01652*, 2024.

[19] 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.

[20] 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 conference on computer vision and pattern recognition*, pages 2901–2910, 2017.

[21] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[23] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[24] Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024.

[25] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14963–14973, 2023.

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[28] 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 *European conference on computer vision*, pages 216–233. Springer, 2024.

[29] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025.

[30] Bruce P Lowerre and B Raj Reddy. Harpy, a connected speech recognition system. *The Journal of the Acoustical Society of America*, 59(S1):S97–S97, 1976.

[31] Wufei Ma, Haoyu Chen, Guofeng Zhang, Celso M de Melo, Alan Yuille, and Jieneng Chen. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. *arXiv preprint arXiv:2412.07825*, 2024.

[32] Wufei Ma, Angtian Wang, Alan Yuille, and Adam Kortylewski. Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. In *European Conference on Computer Vision*, pages 492–508. Springer, 2022.

[33] Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2025.

[34] Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 37:96127–96149, 2024.

[35] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *International Conference on Learning Representations*, 2019.

[36] OpenAI. Introducing openai o3 and o4-mini, 2025. Accessed: Apr 2025.

[37] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[38] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[39] Arijit Ray, Jiafei Duan, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A Plummer, Ranjay Krishna, Kuo-Hao Zeng, et al. Sat: Spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

[40] remyxai. Remyxai/spacellava, 2024.

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

[42] Hanshi Sun, Momin Haider, Ruiqi Zhang, Huitao Yang, Jiahao Qiu, Ming Yin, Mengdi Wang, Peter Bartlett, and Andrea Zanette. Fast best-of-n decoding via speculative rejection. *arXiv preprint arXiv:2410.20290*, 2024.

[43] Gemini Robotics Team, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Travis Armstrong, Ashwin Balakrishna, Robert Baruch, Maria Bauza, Michiel Blokzijl, et al. Gemini robotics: Bringing ai into the physical world. *arXiv preprint arXiv:2503.20020*, 2025.

[44] Guiyao Tie, Zeli Zhao, Dingjie Song, Fuyang Wei, Rong Zhou, Yurou Dai, Wen Yin, Zhejian Yang, Jiangyue Yan, Yao Su, et al. A survey on post-training of large language models. *arXiv preprint arXiv:2503.06072*, 2025.

[45] Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

[46] Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36:58717–58735, 2023.

[47] Xingrui Wang, Wufei Ma, Angtian Wang, Shuo Chen, Adam Kortylewski, and Alan Yuille. Compositional 4d dynamic scenes understanding with physics priors for video question answering. *arXiv preprint arXiv:2406.00622*, 2024.

[48] Xingrui Wang, Wufei Ma, Tiezheng Zhang, Celso M de Melo, Jieneng Chen, and Alan Yuille. Pulsecheck457: A diagnostic benchmark for comprehensive spatial reasoning of large multi-modal models. *arXiv preprint arXiv:2502.08636*, 2025.

[49] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[50] 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.

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

[52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

[53] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[54] Chunting Zhou, Pengfei Liu, Puxin Xu, Srinivasan Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36:55006–55021, 2023.

A Limitations

Our system has a few limitations. First, despite the use of reinforcement learning (RL), supervised finetuning data with chain-of-thought reasoning and explicit 3D representations remains crucial as a warm-up. Training with RL alone–*i.e.*, SpatialReasoner-Zero–fails to produce high-quality reasoning processes. Second, SpatialReasoner consistently performs explicit multi-step 3D reasoning, regardless of the difficulty of the problems. Ideally the model should adopt a hybrid strategy–leveraging commen sense and visual cues for simpler questions while reserving explicit 3D computation and reasoning for more challenging problems. This would retain many strong and generalizable knowledge from the base Qwen2.5-VL model while improving inference efficiency.

B Implementation Details

We train SpatialReasoner using different combinations of curated datasets and training objectives. Starting from the Qwen2.5-VL-7B [4] base model, we first apply SFT with 24k curated SR-CoT data alongside 24k randomly sampled LLaVA [27] data, resulting in SpatialReasoner-SFT. Next, we further train SpatialReasoner-SFT with RL using 1.2k SR-QA examples, leading to our final SpatialReasoner. For comparison, if we instead fine-tune SpatialReasoner-SFT with SFT on the same 1.2k SR-QA set, we obtain SpatialReasoner-SFT (+HQ SFT), serving as an ablation study baseline. In parallel, we directly fine-tune the base model with RL using the 1.2k SR-QA data without prior SFT, resulting in SpatialReasoner-Zero. Additionally, to investigate if explicit 3D perception ability can enable the model to self-organize reasoning trajectories under process rewards, we train the base model with SFT using 12k Basic3D-QA data alongside 12k randomly sampled LLaVA data before applying RL training on the 1.2k SR-QA data, yielding SpatialReasoner-Zero (*w/* 3D Rwd).

We conduct all training experiments using 4×NVIDIA H100 80GB HBM3 GPUs. For SFT, we train the model for 10 epochs (approximately 20K steps with a batch size of 6) on the combined 24k SR-CoT and 24k LLaVA datasets. For RL training, we train for 100 epochs (approximately 13K steps with a batch size of 12) on the 1.2k SR-QA dataset, using 1 GPU with vLLM [23] for efficient inference acceleration. For the experiments withholding multi-object training examples, we double the number of training epochs to compensate for the reduced training set size. We set the learning rate to 5e-6 for SFT and 5e-7 for RL, both following a cosine learning rate scheduler with a warm-up ratio of 0.1. In the KL divergence ablation study, we set the KL penalty weight to 0.04. We monitored the model every 1K training steps and reported results based on the best-performing checkpoint.

C 2D Reasoning as a Shortcut

As shown in Table 1 and Table 2, our SpatialReasoner outperforms previous open-source and proprietary models on 3DSRBench [31], and achieves notable improvements on GQA [19] (from 58.8% to 61.8%) and depth-related questions in CVBench-3D [45] (from 82.5% to 87.3%). However, if we focus on multi-object 3D distance-related questions in CVBench-3D [45] and 3DSRBench [31], we observe contradictory results: SpatialReasoner achieves a substantial improvement of 21.5% on 3DSRBench, but exhibits a notable performance drop of 9.9% on CVBench-3D (see Table 6).

We attribute this discrepancy to the abundant shortcuts in distance-related questions in CVBench-3D. From the qualitative examples in Figure 7, the provided 2D bounding boxes in CVBench-3D can be exploited as shortcuts to answer the 3D spatial reasoning question. Rather than reasoning about 3D distances between objects, we can easily derive the correct answer by comparing the 2D distances between the red and blue boxes and between the red and green boxes. Meanwhile, 3DSRBench is a human-collected VQA dataset and manually avoid such spurious correlation, *e.g.*, "objects closer in 3D space are also closer in 2D image plane". For the 3DSRBench example in Figure 7, the bounding box of the dog is actually closer to the bounding box of the man in black that farther away in 3D space.

Given the 2D bounding box annotations in CVBench-3D and 3DSRBench, we derive a simple heuristic to answer distance-related spatial reasoning questions by simply comparing L2 distances between 2D centers of object bounding boxes. We achieve an 80.2% accuracy on distance questions in CVBench-3D and 34.3% in 3DSRBench. **This demonstrates that baseline models such as Qwen2.5-VL are largely exploiting 2D spatial reasoning as a shortcut to answer complex 3D spatial**

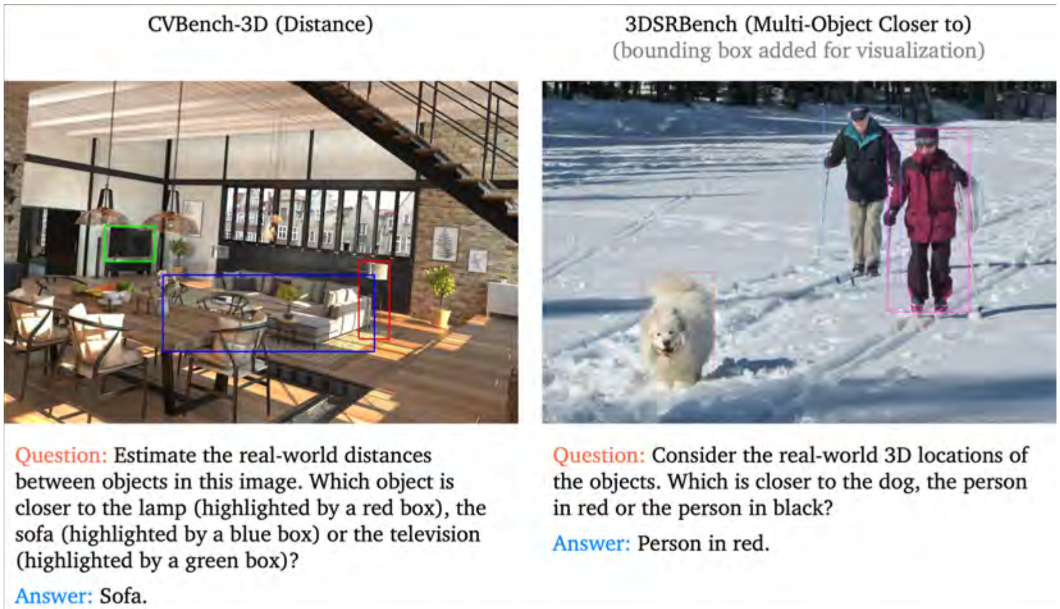


Figure 7: Comparison between (multi-object) distance-related questions in CVBench-3D [45] and 3DSRBench [31].

Method	CVBench3D Distance	3DSRBench multi-object-closer-to
2D Heuristic	80.2	34.3
Qwen2.5-VL-7B-Instruct [4]	83.2	49.4
SpatialReasoner	73.3	70.9

Table 6: Comparison between Qwen2.5-VL [4] and SpatialReasoner on (multi-object) 3D distance-related questions in CVBench-3D [45] and 3DSRBench [31].

reasoning questions. Meanwhile, with our 3D-aware post-training, our SpatialReasoner adopts explicit 3D representation for 3D spatial reasoning questions. **The trade-off between exploiting 2D shortcuts and adopting explicit 3D representations results in slightly lower performance of SpatialReasoner on test data with abundant spurious correlations, but more importantly, leads to a robust and largely improved performance on a challenging real-world dataset.**

Visual cues or explicit 3D reasoning? LVLMs that exploit 2D reasoning as shortcuts may achieve improved performance on test data with abundant spurious correlations (*e.g.*, distance questions in CVBench-3D). However, they cannot genuinely solve 3D spatial reasoning problems and fall far behind SpatialReasoner that builds on explicit 3D spatial reasoning when tested on challenging real-world problems in 3DSRBench.

D Additional Results

We present the following additional results:

1. **Table 7.** We study the failures modes of SpatialReasoner. Results show that 3D computation is more accurate than 3D perception, indicating that most VQA errors stem from flawed 3D perception.
2. **Figure 8.** Qualitative comparisons between different variants of SpatialReasoner.
3. **Figure 9.** Ablation study results on KL divergence.

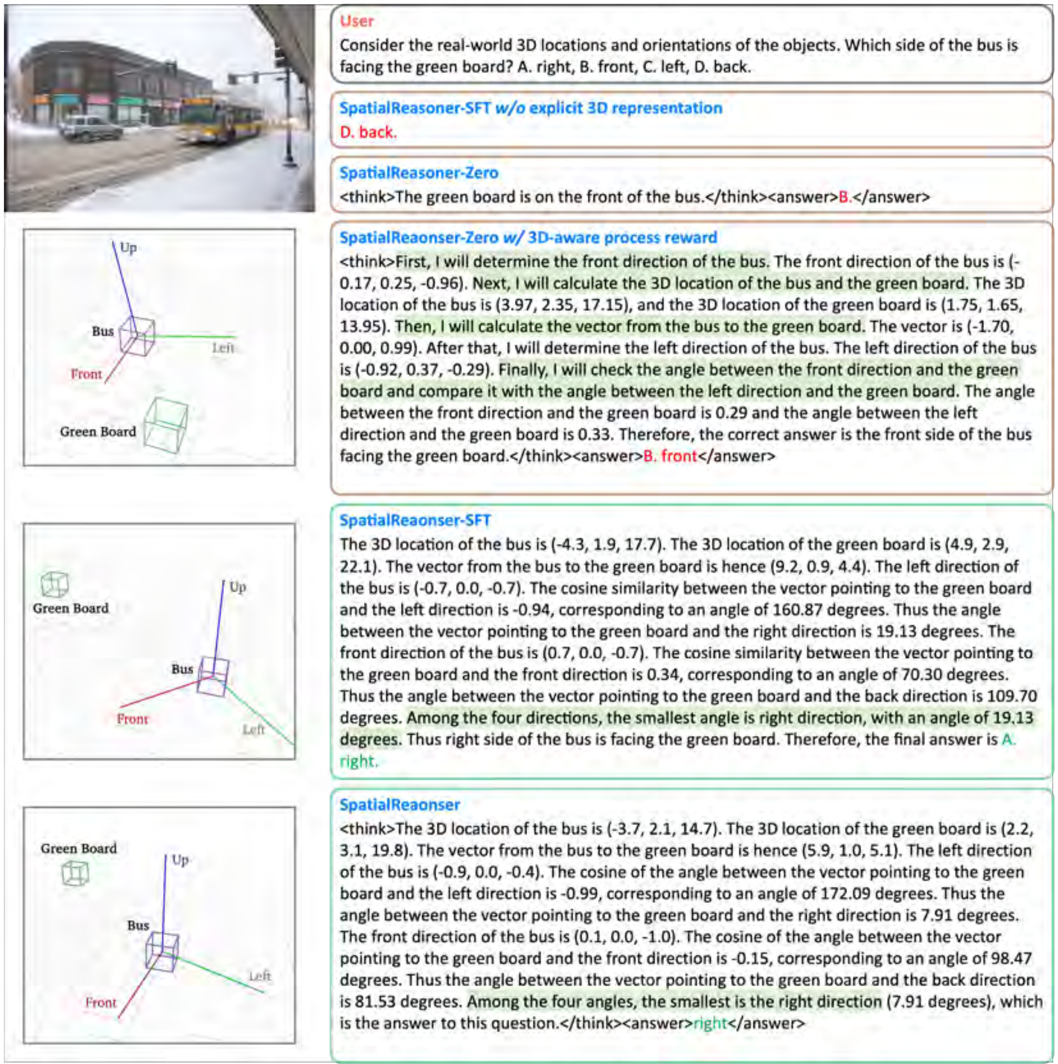


Figure 8: **Qualitative comparisons.** Our explicit 3D spatial reasoning improves interpretability over the baseline that relies on shallow shortcuts.

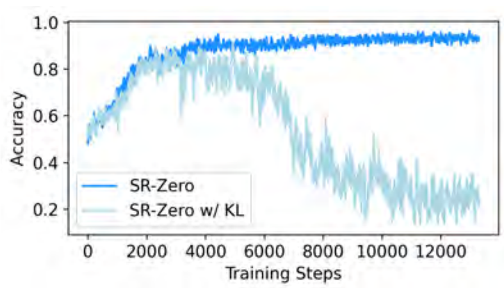
E Open Access

All code, data, and models will be available on our project page to support reproducibility and benefit the research community.

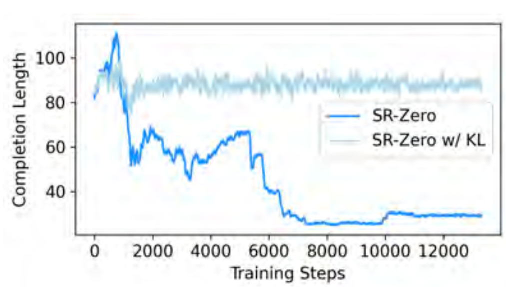
1. Codebase for our full 3D-aware data generation pipeline.
2. Codebase for our SFT and RL finetuning.
3. Synthesized 3D-aware training data.
4. Weights of our SpatialReasonerand SpatialReasoner-SFT.

Method	3D Perception		3D Reasoning		
	Orientation (↑)	Location (↓)	Angle (↑)	Distance (↓)	Depth (↓)
SpatialReasoner-SFT	35.5	0.91	55.0	0.17	0.13
SpatialReasoner	31.0	1.05	52.5	0.19	0.25

Table 7: **Studying failure modes of SpatialReasoner.** We observe that 3D reasoning can estimate angles, distances, and depths a lot more accurately than 3D perception that predicts orientations and locations from visual features.



(a) Accuracy of SpatialReasoner-Zero *w/* and *w/o* KL divergence.



(b) Completion length of SpatialReasoner-Zero *w/* and *w/o* KL divergence.

Figure 9: **Training curve between SpatialReasoner-Zero and SpatialReasoner-Zero *w/* KL.**

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims made in the abstract and introduction are supported by our quantitative results in Section 4 and qualitative results in Figure 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Section A.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: Our paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclose all information needed to reproduce the main results in Section B, including dataset, hyperparameters, and training setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All our code and data will be made public as detailed in Section E.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We present the training and test details in Section 4.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Due to limited computational resources, we were unable to perform multiple runs for each experimental settings that are necessary for statistically robust error estimation. However, we re-run our main method with different random seeds and obtain a standard deviation of 0.20, which shows that our experimental results are sound and significant. Lastly we will open-source our code and data to support public reproduction.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The compute resources are detailed in Section B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work builds on public academic datasets and codebases and we carefully follow the NeurIPS Code of Ethics in each stage of our research.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work is a foundational research on 3D spatial reasoning, enabling large vision language models to reason in 3D space. Our work does not have direct societal impacts.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper studies the 3D spatial reasoning of large vision-language models. Our data and models do not have risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: For all the codebase and data used in this work, we properly cite the creators in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper introduced new code, data, and models that will be released to public upon paper acceptance. All assets are well documented and will be actively maintained in the future.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.