Unfold The World: Factorize 4D Properties in Reinforcing Spatial Reasoning

Yijun Yang[1,2,*], Shenghe Zheng[3,*], Wenbo Li[2,†], Jianhui Liu[4], Haoze Sun[2], Yanbing Zhang[2], Jiaxiu Jiang[2], Lin Song[2], Haoyang Huang[2], Nan Duan[2], and Lei Zhu[1,✉]

[1] The Hong Kong University of Science and Technology (Guangzhou)
leizhu@hkust-gz.edu.cn
[2] Joy Future Academy
[3] The Hong Kong University of Science and Technology
[4] The University of Hong Kong

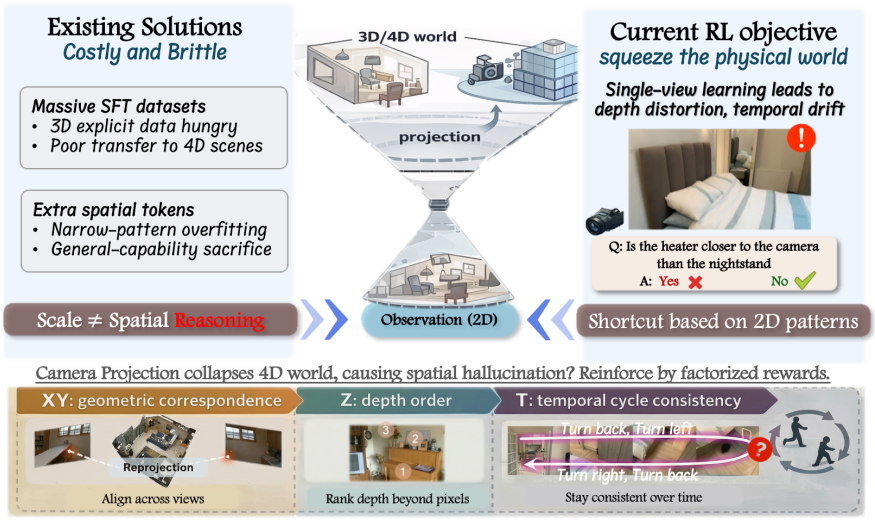

Fig. 1: Motivation. (1) Single-view learning contradicts how the physical world is structured, as real-world perception is inherently multi-view, depth-aware, and temporally continuous. (2) Scaling SFT or adding spatial tokens does not yield a coherent latent world model, while directly optimizing a 4D objective over space and time is computationally and algorithmically intractable.

Abstract. Despite the remarkable prowess of Vision-Language Models (VLMs) in general multimodal tasks, they remain fundamentally "flat" when reasoning about the physical world. We argue that this spatial bottleneck stems from a profound dimensional mismatch: while VLMs are trained to interpret 2D projections, true spatial reasoning demands the recovery of latent 3D geometry and temporal continuity. To conquer this high-dimensional complexity, we advocate a shift from monolithic learning to a "divide and conquer" paradigm. We present FactoSR, a factorized reinforcement learning framework that explicitly interpret the

* Equal Contribution, † Project Lead, ✉ Corresponding Author

dimensions collapsed by visual projection. At its core, FactoSR decomposes the monolithic problem of world-consistent reasoning into three orthogonal, geometric sub-objectives: planar correspondence (XY), depth consistency (Z), and temporal reversibility (T). By optimizing these verifiable constraints within a unified policy learning mechanism, we effectively transform an ill-posed projection recovery problem into a series of tangible reasoning steps. Extensive evaluations on multi-view and video benchmarks demonstrate that this elegant decomposition yields substantial gains in 3D and 4D reasoning, achieving a 5.9% boost on VSI-Bench and 4.5% on All-Angles-Bench. Our findings suggest that reinforcing explicit, factorized 4D consistency is a critical step toward evolving VLMs into robust, world-aware reasoners.

Code: https://github.com/scott-yjyang/FactoSR

Keywords: Spatial Intelligence · Reinforcement Learning · VLMs

1 Introduction

Vision-Language Models (VLMs) [1, 2, 14, 16, 39, 43] have achieved remarkable success in general visual tasks, yet a critical capability remains elusive, *i.e.*, spatial reasoning. This spatial reasoning ability is a fundamental component for VLMs in approaching real-world artificial general intelligence. While humans effortlessly identify spatial relationships in sequential visual environments, current VLMs struggle with even basic spatial queries [29, 40, 50], which requires interpreting the dynamic world beyond 2D projections. This limitation severely constrains their deployment in applications requiring dynamic spatial intelligence, from autonomous driving, robotics navigation to world models.

To mitigate this issue, many studies have synthesized massive spatial question-answering datasets for supervised fine-tuning (SFT) [9, 29, 51] or explored the incorporation of additional spatial tokens [7, 19, 46]. These approaches are often constrained by 3D explicit data hunger, and have poor transferable capability to 4D scenes. Even more critically, while they sample images from 4D scenes [4,10,17,55], they heavily rely on single-view-based question answer. Their static learning on 2D patterns would induce hallucination in dynamic spatial reasoning, due to the absence of true 4D physical-world perception. Consequently, they tend to infer spatial relationships heuristically, often making premature decisions without properly accounting for *geometric correspondence, depth estimation, or temporal consistency*, as illustrated in Fig. 1.

Recently, a few pioneers [30, 31, 42] alternatively investigate applying reinforcement learning with verifiable rewards (RLVR) to enhance the spatial reasoning abilities of VLMs. RLVR has demonstrated superior generalization over SFT by learning diverse reasoning strategies rather than static patterns [21, 26, 35]. However, existing RLVR suites for spatial reasoning [5, 23, 34, 42, 51] employ simple rewards inherited from general understanding, which focus only on final correctness. They also learn in the single-view spirit, which squeezes the dynamics of the latent world by depth distortion and temporal drift. Instead, *spatial*

intelligence requires explicit reasoning across dimensions of the physical world that are collapsed by the camera projection from reality to observations.

Unfortunately, formulating a unified 4D objective over both space and time is computationally and algorithmically intractable. To divide and conquer [6], we present FactoSR, a method beyond pixel-level understanding that factorizes spatial reasoning into plane, depth, and time through online policy reinforcement learning, with three parallel rewards that guide models toward explicit 4D reasoning. The training uses a multi-objective reward framework: format rewards ensure structured outputs; accuracy rewards prioritize correctness; XY rewards constrain re-projection consistency and correspondences to geometrically valid regions across views; Z rewards promote precise 3D localization and relative depth ordering; and T rewards enforce temporal cycle consistency and reversible camera-motion reasoning. Together with supervised spatial fine-tuning, our suite moves beyond image understanding toward models that reason across the dimensions collapsed by projection, supporting a process of observing, localizing, thinking, and answering.

Building on our constructed datasets, our contributions are threefold.

- **We present a learning suite that injects spatial knowledge into VLMs.** By cascading supervised fine-tuning (*FactoSR-SFT*) with reinforcement learning (*FactoSR-RL*), we establish a transition from initial spatial perception to explicit reasoning about the latent world.
- **We introduce a novel factorized reward framework for 4D Reasoning** that decomposes the complicated task into three complementary objectives: *XY-plane*, *Z-depth*, and *T-time*. This design recovers the dimensions typically collapsed by 2D projection, guiding the model to reason precisely across depth and temporal sequences.
- **We translate this design into significant spatial gains.** Our paradigm achieves state-of-the-art performance on multiple spatial benchmarks, particularly an improvement of *5.9% on VSI-Bench* and *4.5% on All-Angles-Bench*, while preserving decent general multi-modal capabilities.

2 Related Work

Spatial Intelligence in MLLMs. Recent large vision language models show strong perceptual abilities, yet their spatial intelligence remains limited. Spatial intelligence involves understanding geometric structure, relative positions, and viewpoint transformations, which requires consistent reasoning over spatial configurations rather than surface level recognition. Existing efforts improve spatial intelligence from three aspects. Architectural approaches such as Spatial-MLLM [45], VLM-3R [19], and 3DThinker [13] introduce geometric biases or 3D representations, while SpatialBot [8] and VILASR [47] leverage external perception tools. Data scaling works including SpatialVLM [11], SpatialRGPT [15], VST [51], and SenseNova-SI [9] expand spatial supervision. Reasoning oriented frameworks such as SpatialLadder [23], Cambrian-S [52], SpaceR [31], and MindCube [56] enhance structured spatial inference. However, these approaches do

not provide both depth and temporal analysis of feasible learning pathways for acquiring spatial intelligence. To address this gap, we propose a reinforcement learning framework with decoupled rewards to explicitly strengthen spatial reasoning and guide the model toward more effective behaviors.

Multimodal Reinforcement Learning. We focus on enhancing the general reasoning abilities of multimodal models via reinforcement learning. OpenVLThinker [18] leverages iterative SFT–RL cycles to facilitate reasoning emergence, while VL-Rethinker [41] and SRPO [59] incorporate self-reflection into RL to promote slow thinking. Open Vision Reasoner [44] studies cognitive behavior transfer from LLMs, and NoisyGRPO [32] and EvolvedGRPO [36] improve generalization and stability through noise modeling and progressive training. Generative RLHF-V [60] further advances multimodal reward modeling. Overall, prior work moves beyond simple outcome rewards toward more structured and robust RL paradigms for multimodal reasoning.

Reinforcement learning for spatial intelligence is a subfield of multimodal RL, mainly focusing on how to stably improve spatial reasoning through verifiable rewards. SVQA-R1 [42] and SpatialThinker [5] incorporate spatial relations into RL objectives via single-view-consistent or dense spatial rewards. Visionary-R1 [49] and SATORI-R1 [34] show that free-form reasoning suffers from shortcut learning, and introduce intermediate verifiable stages such as captioning or region localization. Perception-R1 [57] highlights the importance of perception-oriented rewards. Meanwhile, Visual Spatial Tuning [51] and SpatialLadder [23] adopt progressive training from perception to reasoning, while SpatialReasoner [30] explores explicit 3D representations for better generalization. Overall, prior work suggests that long CoT reasoning or sparse rewards alone is insufficient, motivating our structured solution with explicit representations and multiple verifiable rewards.

3 Preliminaries

Problem Formulation. In this work, we study 4D spatial-temporal intelligence in Vision-Language Models (VLMs), which requires joint reasoning over 3D geometric structures and temporal dynamics. Let $D = \{x_1, x_2, \ldots, x_N\}$ denote a spatial-temporal reasoning dataset, where each sample $x_i = (V_{1:t}, Q_{st}, a)$ consists of a sequence of visual observations $V_{1:t}$, a spatial-temporal query Q_{st}, and the corresponding ground-truth answer a. Here, $V_{1:t} = \{V_1, V_2, \ldots, V_t\}$ represents a time-ordered visual sequence that encodes dynamic 3D scene evolution.

Unlike conventional multimodal reasoning, the 4D intelligence task requires constructing an implicit spatio-temporal representation $S(V_{1:t})$, which captures geometric attributes (*e.g.*, distance, orientation, occlusion, topology) as well as temporal dynamics (*e.g.*, motion, interaction, and state transitions). Formally, given $x_i \in D$, the VLM aims to generate a textual token sequence y by reasoning over $S(V_{1:t})$ to resolve the spatial-temporal query Q_{st}, where the generated sequence $\hat{y}$ is expected to align with the ground-truth answer y.

Reinforcement Learning with Verifiable Rewards. RLVR departs from conventional RL by deriving rewards directly from ground-truth correctness

rather than relying on a learned reward model. This eliminates the need for auxiliary reward estimation, simplifies the training pipeline, reduces computational cost, and mitigates reward hacking by grounding supervision in objectively verifiable outcomes.

Current implementations of RLVR generally follow a two-part structure: a reward computation scheme that evaluates answer validity, and a policy optimization procedure built upon Group Relative Policy Optimization (GRPO) [21,33], which performs stable updates through intra-group comparative advantage estimation. GRPO is a reinforcement learning algorithm derived from PPO that improves policy stability via group-based relative advantage estimation. Its main advantage is that it does not require a value model, reducing memory and computational cost. The training objective of GRPO is to maximize:

$$J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim P(Q),\,\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(O|q)} \Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\Big(\min\big(r_{i,t}(\theta)A_i, \tag{1}$$

$$\text{clip}\big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\big)A_i\big) - \beta D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\Big)\Bigg], \tag{2}$$

where

$$r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid q,\, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid q,\, o_{i,<t})}, A_i = \frac{r_i - \text{mean}(\{r_1, r_2, \ldots, r_G\})}{\text{std}(\{r_1, r_2, \ldots, r_G\})}. \tag{3}$$

Here, π_θ and $\pi_{\theta_{\text{old}}}$ denote the updated and previous policies. q denotes the input query, $\{o_i\}_{i=1}^{G}$ are G sampled responses, $|o_i|$ is the length of the i-th response, and r_i is the reward assigned to response o_i.

4 Methodology

4.1 Overview

Our goal is to endow general vision-language models with explicit spatial reasoning capabilities beyond flat image understanding. Hence, we build our framework upon a strong general-purpose VLM backbone, Qwen3-VL [2], and train it through a two-stage pipeline, as shown in Fig. 2. The data statistics of two stages are introduced in Sec. 5.1.

Stage 1: Spatial Perception Fine-tuning. This stage aims to cultivate spatial grounding and foundational perception capabilities through supervised fine-tuning. Rather than introducing complex reasoning signals abruptly, we implement a short-to-long supervision curriculum to prepare the model for subsequent RL optimization.

We begin by jointly training the model on short-form paired data, where general multimodal understanding from LLaVA-OneVision [27] is synergized with specialized spatial perception tasks. These concise responses emphasize core grounding skills, such as localization, correspondence, and spatial relations, thereby enabling the model to align visual observations with spatial semantics.

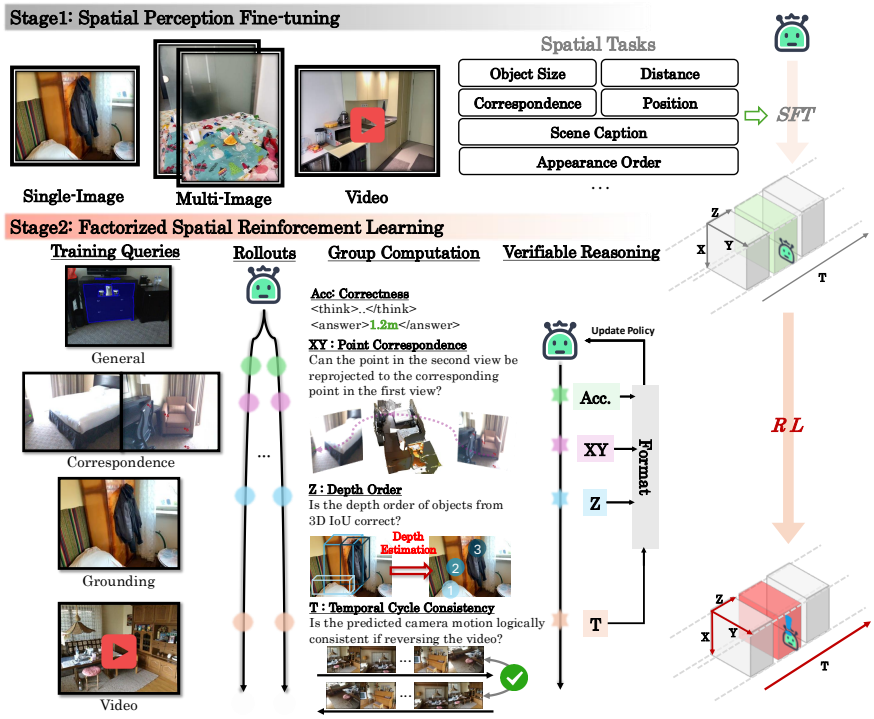

Fig. 2: The progressive training framework of FactoSR. Stage 1 performs supervised fine-tuning on diverse spatial tasks to build foundational spatial perception. Stage 2 introduces factorized reinforcement learning with verifiable rewards (Accuracy, XY, Z, T), explicitly enhancing correspondence, depth, and temporal reasoning.

During this process, visual tokens extracted from images serve as the conditioning context, and the model is optimized via a standard autoregressive objective:

$$\mathcal{L}_\theta(y \mid V_{1:t}, Q_{st}) = -\sum_{i=2}^{L} w_i \log p_\theta\left(y_i \mid V_{1:t}, Q_{st}, y_{1:i-1}\right) . \tag{4}$$

This mixed-task training is performed for a single epoch to inject spatial priors without compromising generalist performance.

Building upon this spatial foundation, we then introduce long-form data featuring structured "Anchor-Transfer-Verify" reasoning trajectories for cross-frame alignment, as illustrated in Fig. 4. The model undergoes further refinement over several hundred iterations using these extended sequences. By progressively scaling from basic grounding to reasoning, Stage I equips the model with incipient spatial logic while ensuring training stability. Crucially, this phase significantly facilitates convergence during the subsequent RL stage.

Stage 2: Factorized Spatial Reinforcement Learning. To bridge the gap between supervised awareness and physically consistent reasoning, we employ Group Relative Policy Optimization (GRPO) to further refine the model's reasoning trajectories. Building upon the "Anchor-Transfer-Verify" CoT framework

from Stage I, this reinforcement stage iteratively optimizes the model's thinking process against our curated verification dataset. Specifically, we transition from static supervision to a factorized rule-based reward that evaluates the consistency of the generated reasoning across three physical dimensions: XY (planar correspondence), Z (depth order), and T (temporal cycle-consistency). This structured feedback steers the model's latent thinking away from heuristic shortcuts toward a coherent, self-verifying 4D spatial logic. The detailed formulations of these rewards and the final optimization objective are introduced in the following subsections.

4.2 Factorized Spatial Reinforcement Learning

In this stage, we employ RL to further enhance the spatial reasoning capabilities of the stage-2 model. For this purpose, we utilize the revised GRPO algorithm [58], which bypasses the need for a value model by computing the relative advantage of each response within a group of responses to the same question. To facilitate this process, we curated a verification dataset comprising tasks related to spatial understanding, 3D object detection, and general multi-modal understanding. In the GRPO framework, we employ a mixed rule-based reward to evaluate the generated responses, including format, accuracy, XY, Z, and T rewards. For a given response $\hat{y}$ and its corresponding ground truth y, the basic accuracy reward function is used to ensure correctness and is defined as: $\mathcal{R}_{acc}(y, \hat{y}) = \mathbb{I}[\hat{y} = y]$.

XY Reward: Point Correspondence. For multi-view spatial reasoning, a model should not "guess" correspondence from 2D appearance alone. Instead, a correct correspondence must be *geometrically admissible*: the predicted point in the target view should agree with the reprojection of the reference point under camera intrinsics, poses, and depth. Therefore, our XY reward directly supervises *2D correspondence* by enforcing *reprojection consistency* and *overlap validity*, providing dense, physically grounded guidance during RL.

Given two views V_1, V_2 with depth maps $\mathcal{D}_1, \mathcal{D}_2$, intrinsics $\mathbf{K}_1, \mathbf{K}_2$, and camera-to-world poses $\mathbf{T}_1, \mathbf{T}_2$, we are provided a reference pixel $\mathbf{p}_1 = (u_1, v_1)$ in V_1 and a discrete candidate set $\{\mathbf{p}_2^{(A)}, \mathbf{p}_2^{(B)}, \mathbf{p}_2^{(C)}, \mathbf{p}_2^{(D)}\}$ in V_1. Let $\hat{\mathbf{p}}_2 = (\hat{u}_2, \hat{v}_2)$ be the model-selected candidate point in view 2. Specifically, we first compute the reprojection map from V_1 to V_2. For each pixel $\mathbf{p}_1 = (u_1, v_1)$ in view 1 with depth $d_1 = \mathcal{D}_1(\mathbf{p}_1)$, we unproject to camera coordinates:

$$\mathbf{x}_1 = d_1 \, \mathbf{K}_1^{-1} \, \tilde{\mathbf{p}}_1, \quad \tilde{\mathbf{p}}_1 = (u_1, v_1, 1)^\top . \tag{5}$$

We then transform it to world coordinates and reproject to view 2:

$$\mathbf{X} = \mathbf{T}_1 \, \tilde{\mathbf{x}}_1, \quad \mathbf{x}_2 = \mathbf{T}_2^{-1} \mathbf{X}, \quad \tilde{\mathbf{p}}_2 \sim \mathbf{K}_2 \mathbf{x}_2 \, , \tag{6}$$

where $\tilde{\mathbf{x}}_1 = (\mathbf{x}_1^\top, 1)^\top$ and $\sim$ denotes equality up to scale, and the resulting pixel coordinate is $\mathbf{p}_2^\star = (u_2^\star, v_2^\star) = \left(\frac{\tilde{p}_{2,x}}{\tilde{p}_{2,z}}, \frac{\tilde{p}_{2,y}}{\tilde{p}_{2,z}} \right)$. We also record the projected depth in view-2 camera coordinates $z_2^\star$ and a validity mask

$$\mathbb{I}_{\text{valid}}(\mathbf{p}_1) = \mathbb{I}[z_2^\star > 0] \cdot \mathbb{I}[\mathbf{p}_2^\star \in V_2] \, , \tag{7}$$

If $\mathbb{I}_{\text{valid}}(\mathbf{p}_1) = 0$, the correspondence is undefined, and we assign a zero reward. Otherwise, we compare the model prediction $\hat{\mathbf{p}}_2$ with the projected target $\mathbf{p}_2^\star$ in a normalized coordinate system:

$$d \;=\; \left\| \left(\frac{u_2^\star}{W_2}, \frac{v_2^\star}{H_2} \right) - \left(\frac{\hat{u}_2}{W_2}, \frac{\hat{v}_2}{H_2} \right) \right\|_2 , \tag{8}$$

where (H_2, W_2) is the size of V_2. We then assign a soft, distance-aware reward:

$$r_{\text{reproj}} = \exp\left(-\frac{d}{\sigma} \right) \cdot \mathbb{I}[d \le 3\sigma] , \tag{9}$$

where σ controls tolerance. The hard cutoff $\mathbb{I}[d \le 3\sigma]$ prevents rewarding far-away guesses and stabilizes RL by suppressing spurious gradients from grossly incorrect correspondences.

Reprojection alone may still reward points that are geometrically close but *not visible* in view 2 due to occlusion. To enforce physical plausibility, we construct an overlap mask on view 2 by checking depth consistency. For each valid reprojection, we mark $\mathbf{p}_2$ as visible if

$$|\mathcal{D}_2(\mathbf{p}_2) - z_2^\star| \le \delta , \tag{10}$$

where δ is a depth threshold. Thus, we construct the overlap mask $\mathbf{M}_2(\cdot) \in \{0, 1\}$, which identifies pixels in view 2 that are truly visible from view 1 under the given camera configuration. It is derived from the same reprojection process used to compute $\mathbf{p}_2^\star$. We gate the XY reward using the mask:

$$R_{XY} = r_{\text{reproj}} \cdot \mathbf{M}_2(\hat{\mathbf{p}}_2) \; . \tag{11}$$

This overlap gating turns the reward into a *visibility-aware* signal: even if $\hat{\mathbf{p}}_2$ is close to $\mathbf{p}_2^\star$, it receives zero reward if it falls outside the depth-consistent overlap region, discouraging correspondences on occluded or non-overlapping areas. Overall, the XY reward explicitly avoids appearance heuristics and promotes cross-view alignment that is both reprojection-consistent and visibility-valid.

Z Reward: Depth Order. After SFT establishes the model's basic grounding ability, we observe that further optimizing IoU localization during RL yields no benefit for spatial visual question answering. The core difficulty is not object detection itself, but reasoning about *relative depth* between objects in 3D space from 2D observations. Rather than supervising metric depth values, we directly optimize the correctness of depth order.

The policy is required to output a set of structured 3D bounding boxes. All predicted boxes are matched with ground-truth boxes using Hungarian assignment [22] with a 3D GIoU-based cost. For each matched object, we extract the depth of its center in the camera coordinate system, producing the predicted and ground-truth depth sequences. Given the predicted and ground-truth depth sequences $\hat{\mathbf{z}} = \{\hat{z}^{(1)}, \ldots, \hat{z}^{(n)}\}$ and $\mathbf{z} = \{z^{(1)}, \ldots, z^{(n)}\}$, we evaluate whether the predicted front–back relationships between objects are consistent with the ground truth using the Kendall-τ rank correlation.

Specifically, for every pair of objects (i, j) with $i < j$, we compare the ordering of their depths in the predicted and ground-truth sequences:

$$(\hat{z}^{(i)} - \hat{z}^{(j)})(z^{(i)} - z^{(j)}) > 0. \tag{12}$$

A pair is considered *concordant* if the predicted and ground-truth orders agree, and *discordant* if the two orders disagree. Let N_c and N_d denote the numbers of concordant and discordant pairs, respectively. The Kendall-τ coefficient is then defined as:

$$\tau = \frac{N_c - N_d}{\frac{n(n-1)}{2}}. \tag{13}$$

The coefficient $\tau \in [-1, 1]$ measures the consistency between predicted and ground-truth depth rankings. We further normalize it to $[0, 1]$ to obtain the depth ordering reward: $R_Z = \frac{\tau+1}{2}$.

Finally, this reward directly reinforce the model to recover the relative depth structure of the scene, which constitutes the core reasoning requirement for 3D spatial understanding.

T Reward: Temporal Cycle Consistency. While XY and Z rewards collaboratively reinforce 3D reasoning, they do not guarantee that a model truly understands motion over time. A model may answer a navigation question using static cues without reasoning about how the camera actually moves. To explicitly reinforce the temporal dimension, we introduce a T reward that enforces cycle consistency, *i.e.*, reasoning from the start view to the end view must be logically reversible when the process is queried in the opposite direction.

Each training sample, therefore, contains a forward question-answer pair together with a constructed inverse question and its expected answer. For instance, if the forward solution of camera motion corresponds to "Turn right, Turn back", the inverse problem, reasoning from the terminal state back to the start, should yield "Turn back, Turn left", ensuring a physically consistent cycle. During RL, the model rolls out both the forward prediction $\hat{y}$ and the inverse prediction $\hat{y}^{inv}$. The temporal cycle consistency reward evaluates whether both predictions match the expected answers:

$$\mathcal{R}(y, \hat{y}) = \mathcal{R}_{acc}(y, \hat{y}) \cdot \mathcal{R}_{acc}(y^{inv}, \hat{y}^{inv}). \tag{14}$$

T reward complements the accuracy reward by requiring the model to maintain logical reversibility between forward and inverse motion sequences, fostering a robust understanding of ego-motion and temporal causality.

Final Objective. To synergistically integrate the semantic accuracy with granular geometric constraints, we define a factorized total reward function $\mathcal{R}_{\text{total}}$. We stipulate that any reinforcement is strictly contingent upon the fulfillment of the format constraint $\mathbb{I}[\mathcal{R}_{\text{format}} = 1]$, thereby ensuring the structure of the model output. The final objective is formulated as a weighted composition:

$$\mathcal{R}_{\text{total}}(y, \hat{y}) = \mathbb{I}[\mathcal{R}_{format} = 1] \cdot (\lambda_1 \mathcal{R}_{acc} + \lambda_2 \mathcal{R}_{XY} + \lambda_3 \mathcal{R}_Z + \lambda_4 \mathcal{R}_T), \tag{15}$$

where $\lambda_{\{\cdot\}}$ denotes the task-specific importance of correspondence, depth reasoning, and temporal consistency.

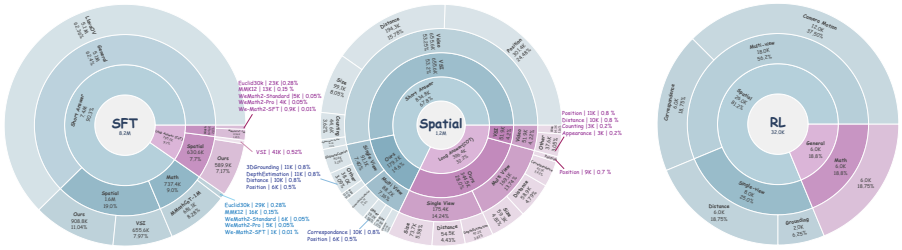

Fig. 3: Overview of (a) FactoSR-SFT and (b) FactoSR-RL datasets. Please zoom in.

By factorizing the reward space into explicit spatial, depth, and temporal dimensions, our RL framework effectively transitions from simple pattern matching to a deeper, physically-grounded understanding of 4D scenes.

5 Experiments

5.1 Data

Fig. 3 summarizes the training data used in our framework. The nested charts present hierarchical statistics, where each ring from the center outward corresponds to progressively finer-grained categories in the legend.

In the SFT stage, the dataset contains 8.2M samples, dominated by short-answer tasks (7.4M, 90.3%), while long-answer reasoning data (795K, 9.7%) provides additional chain-of-thought supervision. The short-answer portion mainly consists of general instruction data LLaVA-OneVision (5.1M) [27], followed by spatial reasoning (1.6M) and mathematical reasoning (737K). We build a new spatial reasoning dataset using our designed data pipeline. This dataset organizes 1.2M samples across both long-answer and short-answer formats, covering diverse spatial tasks, such as single-view depth estimation, multi-view correspondence, multi-view camera-relative motion, and video spatial understanding [52].

Finally, the RL stage employs a smaller but targeted dataset of 32K samples, where spatial reasoning dominates (81.2%), complemented by general mathematical reasoning data (18.8%).

5.2 Implementation Details

Stage 1. This training stage aims to establish a strong foundation of spatial understanding capabilities. For this stage, we use a global batch size of 128, a sequence length of 8,192, and a dynamic data packing strategy to accelerate the training process. We employ the AdamW [28] optimizer, setting the base learning rate to 5×10^{-5} and the vision encoder's learning rate to 5×10^{-6}. For the CoT cold-start, we continue training the model. The hyper-parameters are adjusted to a global batch size of 128, a base learning rate of 1×10^{-5}, a vision encoder learning rate of 1×10^{-6}, and a sequence length of 8,192.

Stage 2. In the RL stage, we refine the model from the first stage using the VeRL [37] framework. We adopt a revised version of the GRPO algorithm [58], using a rollout size of 8 samples per query and a sampling temperature of 1.0.

Table 1: Fine-grained comparison on 4D reasoning benchmarks. The best results among open-source VLMs are **bold**. "Base" denotes Qwen3-VL-8B-Instruct.

Methods	All-Angles-Bench [54]							VSI-Bench [50]									
	Avg.	Attr.	Pose	Cnt.	Manip.	Rel-Dir.	Rel-Dist.	Avg.	Cnt.	Obj-Size	Room-Size	Abs-Dist.	Dir-H	Dir-M	Dir-E	Rel-Dist.	Appr-Ord.	Route
Open-Source General/Spatial MLLMs																		
InternVL3-2B [61]	48.6	66.3	42.6	48.2	41.8	39.5	50.2	30.4	57.7	19.8	33.3	29.0	28.7	25.7	47.5	37.0	14.7	22.7
InternVL3-8B [61]	50.5	78.6	36.4	51.0	42.0	34.7	52.8	38.7	55.4	39.7	41.1	33.6	18.5	42.6	53.0	40.8	35.1	22.7
SpaceR-7B [31]	49.8	71.8	**51.1**	44.6	42.4	36.4	51.4	44.4	53.1	60.1	37.8	28.6	40.5	47.4	48.8	43.1	40.9	33.5
MiMo-VL-7B [39]	52.9	78.3	38.1	53.4	39.9	43.8	57.1	47.8	64.9	61.4	48.9	22.4	36.2	44.2	50.2	46.5	60.2	29.9
Qwen2.5-VL-7B-Instruct [3]	50.1	74.7	50.0	51.0	39.1	37.2	50.6	36.0	42.3	46.2	40.2	22.1	29.0	40.7	50.7	36.3	28.5	30.4
SpatialThinker-7B [5]	47.6	70.5	43.2	41.4	38.7	39.5	49.0	33.0	41.4	37.7	41.7	13.5	24.7	40.5	48.4	40.6	27.5	29.4
VST-7B-SFT [51]	49.5	76.5	35.8	45.8	42.0	41.8	48.2	55.3	68.0	73.2	57.6	39.8	44.8	54.0	51.6	51.5	52.8	43.3
VST-7B-Thinking [51]	49.0	75.2	39.2	47.4	42.4	36.9	48.0	52.6	61.9	73.0	58.3	33.2	40.5	46.8	50.2	50.3	53.9	41.2
Qwen3-VL-8B-Instruct [2]	49.5	76.5	21.0	51.0	37.6	43.8	53.6	55.6	63.8	73.0	56.3	44.3	42.3	52.1	53.0	53.3	57.3	32.5
Qwen3-VL-8B-Thinking [2]	50.9	78.9	24.4	47.4	43.7	42.9	53.0	49.8	47.0	65.6	43.9	38.8	42.3	44.4	51.6	52.5	54.7	35.1
Our Method																		
FactoSR-8B-SFT	51.6	77.5	40.9	50.2	45.6	38.9	50.8	58.5	**66.7**	73.7	58.2	45.8	46.4	53.7	50.2	59.3	63.1	38.2
FactoSR-8B-RL	**55.4**	**79.1**	44.3	**51.0**	**45.2**	**44.6**	**61.1**	**61.5**	66.4	**75.7**	**63.7**	**46.7**	**53.4**	**63.7**	**61.8**	**60.3**	**65.7**	**46.9**
Δ (Ours vs. Base)	**+5.9**	**+2.6**	**+23.3**	**+0.0**	**+7.6**	**+0.8**	**+7.5**	**+5.9**	**+2.6**	**+2.7**	**+7.4**	**+2.4**	**+11.1**	**+11.6**	**+8.8**	**+7.0**	**+8.4**	**+14.4**

Table 2: Quantitative Comparison with state-of-the-art VLMs on 13 3D/4D spatial benchmarks and general benchmarks.

Benchmarks	3/4D-Avg.	4D Reasoning		3D Reasoning							GeneralQA			
		VSI	AllAngles	BLINK	3DSR_C	CV-2D	CV-3D	ERQA	RealWorldQA	MMSI	MMB_CN	MMB_EN	MMStar	OCRB
Proprietary Models														
Gemini-2.5-Pro [16]	64.4	48.4	61.3	70.6	57.6	80.4	91.3	55.8	77.3	36.9	90.2	89.2	79.1	86.6
GPT-4o [1]	57.7	34.0	52.4	65.9	44.3	75.8	83.0	57.0	76.2	30.3	83.9	84.8	65.1	80.6
Open-Source General/Spatial VLMs														
InternVL3-2B [61]	50.6	30.4	48.6	52.8	46.4	71.9	77.3	36.2	65.5	25.9	77.1	78.3	61.5	83.7
InternVL3-8B [61]	56.2	38.7	50.5	55.7	52.7	80.6	86.0	40.5	70.6	30.9	81.9	82.0	68.2	88.0
SpaceR-7B [31]	53.3	44.4	49.8	54.3	47.5	73.9	76.2	40.5	64.2	29.4	80.3	83.0	61.6	85.9
MiMo-VL-7B [39]	58.2	47.8	52.9	59.7	56.1	76.9	86.9	41.0	**73.5**	29.3	80.9	80.8	**71.1**	84.5
Qwen2.5-VL-7B-Instruct [3]	52.8	36.0	50.1	55.3	49.0	75.6	73.8	41.0	68.1	26.5	82.3	82.3	70.9	87.9
SpatialThinker-7B [5]	53.2	33.0	47.6	55.9	47.6	78.7	83.3	39.5	67.6	25.7	79.8	81.0	63.4	87.7
VST-7B-SFT [51]	60.2	55.3	49.5	62.1	53.3	77.9	94.8	43.8	71.5	33.3	80.4	81.1	63.1	86.3
VST-7B-Thinking [51]	60.5	52.2	49.0	62.5	**57.6**	78.6	**95.5**	44.5	69.3	**34.9**	76.5	77.2	63.9	86.9
Qwen3-VL-8B-Instruct [2]	59.1	55.6	49.5	**66.1**	52.8	78.6	90.8	40.1	70.7	28.1	83.3	84.2	70.1	**90.3**
Qwen3-VL-8B-Thinking [2]	59.9	49.8	50.9	62.3	54.8	78.8	93.1	42.8	73.3	30.7	82.2	82.7	74.3	85.3
Our Method														
FactoSR-8B-SFT	60.3	58.5	51.6	62.6	51.9	81.1	90.7	45.3	71.6	29.7	83.2	85.1	67.4	87.7
FactoSR-8B-RL	**62.0**	**61.5**	**55.4**	66.0	52.9	**81.3**	91.7	**47.3**	71.6	30.7	**84.1**	**85.7**	69.1	87.6
Δ (Ours vs. Base)	**+2.9**	**+5.9**	**+5.9**	-0.1	**+0.1**	**+2.7**	**+0.9**	**+7.2**	**+0.9**	**+2.6**	**+0.8**	**+1.5**	-1.0	-2.7

This stage utilizes the AdamW optimizer with a constant learning rate of 1×10^{-6} and a global batch size of 128.

Evaluation. We assess the abilities of state-of-the-art and our models across three distinct capabilities: (1) 4D reasoning ability is benchmarked with All-Angles-Bench [54] and VSI-Bench [50]. We evaluate the fine-grained performance on sub-categories of the two core benchmarks. (2) 3D reasoning ability is benchmarked with BLINK [20], 3DSRBench (3DSR_C) [29], CVBench (CV-2D, CV-3D) [40], ERQA [38], RealWorldQA [48], and MMSI-Bench (MMSI) [53]. (3) General multi-modal understanding is evaluated across a suite of standard benchmarks: MMBench [24] (MMB_CN, MMB_EN), MMStar [12], and OCR-Bench [25]. All the models are evaluated using their native system prompt to ensure the fairness of comparisons. More details on prompts, SFT, and RL training setups, are provided in *Appendices*.

5.3 Quantitative Results

4D Reasoning Analysis. We compare FactoSR with a range of open-source spatial and general VLMs, including InternVL, Qwen-VL, MiMo-VL, SpaceR, SpatialThinker, and VST. As shown in Tab. 1, our method achieves the best performance on both All-Angles-Bench and VSI-Bench, two challenging 4D spatial reasoning benchmarks. Specifically, FactoSR-8B-RL achieves *61.5%* on VSI and

Fig. 4: Qualitative comparison between Qwen3-VL-8B Instruct and FactoSR-8B-RL on spatial reasoning tasks. While Qwen3-VL-8B Instruct often relies on appearance cues and makes inconsistent judgments, FactoSR-RL produces 4D-consistent reasoning across views, depth, and motion.

55.4% on AllAngles, outperforming the strongest open-source baselines by *+5.9* and *+2.5* points, respectively. The improvements are consistent across diverse sub-tasks such as attribute identification, camera pose estimation, and relative spatial reasoning on AllAngles, as well as video-based tasks including relative direction, spatial distance reasoning, and route planning on VSI.

Overall Benchmark Comparison. Tab. 2 further compares our model with state-of-the-art proprietary and open-source VLMs across *13* benchmarks.

FactoSR-8B-RL achieves the best spatial reasoning performance with the average of *62.0* across all 3D/4D spatial benchmarks, improving over the base model by *+2.9* points. Besides the strong gains on the 3/4D benchmarks, FactoSR also achieves competitive performance on general multimodal benchmarks, *e.g.*, MMBench-CN, MMBench-EN. *More results are provided in Appendices.*

5.4 Ablation Study

To quantitatively evaluate the contributions of the factorized rewards, we construct three specialized evaluation groups to analyze spatial reasoning improve-

ments: The correspondence group $\mathbf{\Delta}_{\text{Cor}}$ includes All-Angles-Bench (Attribute Identification), BLINK (Functional Correspondence, Semantic Correspondence, Visual Correspondence). The depth group$\mathbf{\Delta}_{\text{Depth}}$ includes CV-Bench-3D (Depth), BLINK (Relative Depth, Spatial Relation). The camera motion group $\mathbf{\Delta}_{\text{t}}$ includes RealWorldQA, MMSIBench (Motion-Cam), BLINK (Multi-view Reasoning), VSI-Bench (Route Planning). The 3D/4D Avg. Acc. is computed over 9 spatial benchmarks.

Tab. 3 shows that vanilla GRPO yields only marginal gains (+0.2), indicating that general RL signals alone are insufficient for spatial reasoning. Introducing factorized rewards leads to targeted improvements: XY mainly enhances correspondence reasoning (Δ_{Cor} +2.7), Z significantly improves depth understanding (Δ_{Depth} +1.3), and T notably strengthens temporal reasoning (Δ_t +7.9). Further analysis in Tab. 4 reveals that the vanilla grounding reward [51] is ineffective, causing overall performance drops (-0.2) and degraded spatial relation reasoning (-0.7). In contrast, Z reward consistently improves all depth-related metrics, especially Relative Depth (+3.2), validating that RL should emphasize relative depth of objects rather than the simple localization. Combining all rewards achieves the best overall improvement (1.7%), demonstrating that XY, Z, and T provide complementary supervision.

Table 3: Average accuracy across all 9 benchmarks and performance on three specialized tasks with relative improvements. *FactoSR* models consistently outperform SFT and vanilla GRPO.

Model	3D/4D Avg. Acc. (9)	$\mathbf{\Delta}_{\text{Cor}}$	$\mathbf{\Delta}_{\text{Depth}}$	$\mathbf{\Delta}_{\text{t}}$
Supervised Fine-Tuning				
FactoSR-SFT	60.3	69.9	86.3	45.7
Reinforcement Learning				
+ Vanilla GRPO	$60.5_{\pm 0.1}$ (+0.2)	$71.6_{\pm 0.8}$ (+1.7)	$85.8_{\pm 1.1}$ (-0.5)	$47.1_{\pm 2.2}$ (+1.4)
+ XY Reward	$60.7_{\pm 0.1}$ (+0.4)	$\mathbf{72.6}_{\pm 0.7}$ **(+2.7)**	$86.6_{\pm 1.0}$ (+0.3)	$47.2_{\pm 2.8}$ (+1.5)
+ Z Reward	$60.6_{\pm 0.2}$ (+0.3)	$69.8_{\pm 0.6}$ (-0.1)	$\mathbf{87.6}_{\pm 0.4}$ **(+1.3)**	$47.4_{\pm 2.0}$ (+1.7)
+ T Reward	$60.9_{\pm 0.2}$ (+0.6)	$69.9_{\pm 0.3}$ (+0.0)	$87.4_{\pm 0.8}$ (+1.1)	$\mathbf{53.6}_{\pm 1.1}$ **(+7.9)**
+ Full Rewards	$\mathbf{62.0}_{\pm 0.2}$ **(+1.7)**	$72.1_{\pm 0.7}$ (+2.2)	$87.1_{\pm 0.6}$ (+0.8)	$50.9_{\pm 1.3}$ (+5.2)

Table 4: Ablation on depth supervision showing the effectiveness of the Z reward and vanilla grounding reward over SFT models.

Model	3D/4D Avg. Acc. (9)	$\mathbf{\Delta}_{\text{Depth}}$	Depth	Relative_Depth	Spatial_Relation
+ Vanilla Grounding	-0.2	-0.1	-0.5	+0.8	-0.7
+ Z Reward	**+0.3**	**+1.3**	**+0.2**	**+3.2**	**+0.6**

5.5 Visualization

Analysis of Factorized Rewards. Fig. 4 illustrates how factorized rewards improve structured spatial reasoning. In multi-view correspondence, SFT models rely on size heuristics and fail to preserve identity. After RL, the model follows a cross-view CoT (anchor→transfer→verify), correctly tracking the target person across viewpoints, driven by the XY reward. In single-view spatial relation, the model adopts a depth-aware CoT, reasoning via occlusion and layering rather than size of the potted plant, enabled by the Z reward. In video camera motion, the model develops a temporal CoT by analyzing coherent foreground–background shifts around the mug, guided by the T reward. Overall, the factorized rewards promote consistent single-view and multi-view reasoning across space and time.

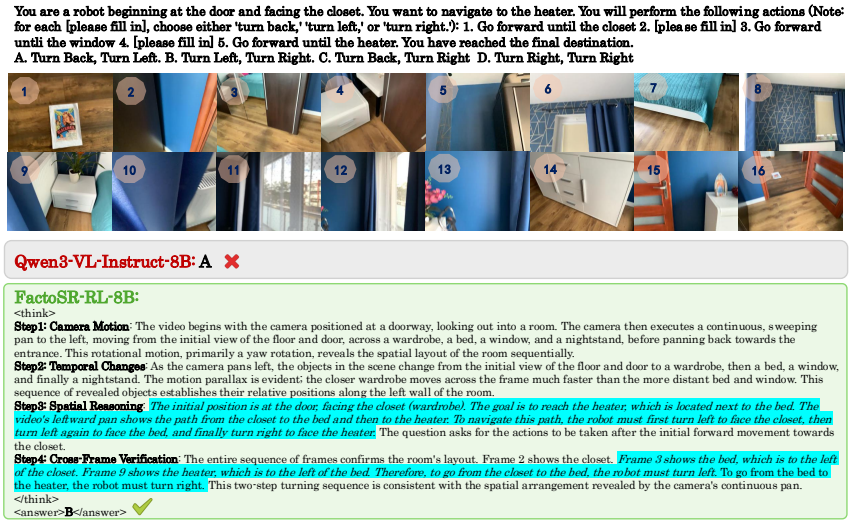

Fig. 5: The example of video route plan. Given a sequence of egocentric observations, our model infers a consistent navigation plan by reconstructing the underlying spatial layout and camera motion, effectively acting as a latent world model.

Generalization to Video Route Plan. The visual evidence from the FactoSR-RL-8B model in Fig. 5 demonstrates that the integration of camera motion data as a structural prior, coupled with a temporal cycle consistency reward in RL framework, significantly generalizes well to video route planning. While standard VLMs like Qwen3-VL often fail to maintain spatial orientation, as seen in its false choice of "Turn Back", the RL-tuned agent leverages motion parallax and yaw rotation (detailed in Steps 1 and 2) to construct a coherent topological map of the environment. Crucially, the model is encouraged to maintain a bijective mapping between temporal video observations and the robot's latent spatial state. This ensures that the transition from the closet to the bed, and ultimately to the heater, is grounded in cross-frame verification rather than isolated object recognition. As a result, the agent effectively decodes the underlying scene dynamics (Step 4), correctly identifying that "Turn Left, Turn Right" is the only path consistent with the reconstructed 4D navigation manifold.

6 Conclusion

Spatial intelligence requires the ability to reason freely within the vast 4D world, yet most vision-language models are shaped under a single-view paradigm that collapses this rich structure into flat image observations. To move beyond this limitation, we introduce FactoSR, a reinforcement learning framework that factorizes spatial reasoning into complementary dimensions of plane, depth, and time, allowing models to reason over the latent spatial structure underlying visual observations. Results on diverse 3D and 4D spatial benchmarks show that FactoSR pushes forward spatial reasoning, encourage multi-modal models to observe, reason, and interact with the latent structure of the physical world.

Acknowledgements

This work is supported by the Program of Guangdong Education Department. This work is also supported by Joy Future Academy and JD.com for research funding and computing resources.

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. Bai, S., Cai, Y., Chen, X., Huang, Q., Li, K., Lin, Z., Zhu, K., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025), https://arxiv.org/abs/2511.21631
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. Batra, H., Tu, H., Chen, H., Lin, Y., Xie, C., Clark, R.: Spatialthinker: Reinforcing 3d reasoning in multimodal llms via spatial rewards. arXiv preprint arXiv:2511.07403 (2025)
6. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Icml. vol. 2, p. 4 (2021)
7. 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)
8. Cai, W., Ponomarenko, I., Yuan, J., Li, X., Yang, W., Dong, H., Zhao, B.: Spatialbot: Precise spatial understanding with vision language models. arXiv:2406.13642 (2024)
9. Cai, Z., Wang, R., Gu, C., Pu, F., Xu, J., Wang, Y., Yin, W., Yang, Z., Wei, C., Sun, Q., et al.: Scaling spatial intelligence with multimodal foundation models. arXiv preprint arXiv:2511.13719 (2025)
10. 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)
11. 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)
12. 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)
13. Chen, Z., Zhang, M., Yu, X., Luo, X., Sun, M., Pan, Z., Feng, Y., Pei, P., Cai, X., Huang, R.: Think with 3d: Geometric imagination grounded spatial reasoning from limited views. arXiv preprint arXiv:2510.18632 (2025)
14. 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)

15. 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)
16. 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)
17. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
18. Deng, Y., Bansal, H., Yin, F., Peng, N., Wang, W., Chang, K.W.: Openvlthinker: An early exploration to complex vision-language reasoning via iterative self-improvement. arXiv e-prints pp. arXiv–2503 (2025)
19. 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)
20. 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)
21. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv:2501.12948 (2025)
22. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
23. Li, H., Li, D., Wang, Z., Yan, Y., Wu, H., Zhang, W., Shen, Y., Lu, W., Xiao, J., Zhuang, Y.: Spatialladder: Progressive training for spatial reasoning in vision-language models. arXiv preprint arXiv:2510.08531 (2025)
24. 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)
25. 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)
26. Liu, Z., Sun, Z., Zang, Y., Dong, X., Cao, Y., Duan, H., Lin, D., Wang, J.: Visual-rft: Visual reinforcement fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2034–2044 (2025)
27. LLaVA-Lab: Llava-onevision-data. `https://huggingface.co/datasets/lmms-lab/LLaVA-OneVision-Data` (2024)
28. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
29. 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)
30. Ma, W., Chou, Y.C., Liu, Q., Wang, X., de Melo, C., Xie, J., Yuille, A.: Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. arXiv:2504.20024 (2025)
31. 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)
32. Qiu, L., Ning, S., Sun, J., He, X.: Noisygrpo: Incentivizing multimodal cot reasoning via noise injection and bayesian estimation. arXiv preprint arXiv:2510.21122 (2025)
33. 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)

34. Shen, C., Wei, W., Qu, X., Cheng, Y.: Satori-r1: Incentivizing multimodal reasoning with spatial grounding and verifiable rewards. arXiv e-prints pp. arXiv–2505 (2025)
35. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
36. Shen, Z., Yu, Q., Li, J., Ji, W., Chen, Q., Tang, S., Zhuang, Y.: Evolvedgrpo: Unlocking reasoning in lvlms via progressive instruction evolution. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems
37. Sheng, G., Zhang, C., Ye, Z., Wu, X., Zhang, W., Zhang, R., Peng, Y., Lin, H., Wu, C.: Hybridflow: A flexible and efficient rlhf framework. In: Proceedings of the Twentieth European Conference on Computer Systems. pp. 1279–1297 (2025)
38. Team, G.R., Abeyruwan, S., Ainslie, J., Alayrac, J.B., Arenas, M.G., Armstrong, T., Balakrishna, A., Baruch, R., Bauza, M., Blokzijl, M., et al.: Gemini robotics: Bringing ai into the physical world. arXiv preprint arXiv:2503.20020 (2025)
39. Team, M.V.: Mimo-vl technical report (2025), https://arxiv.org/abs/2506.03569
40. 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)
41. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. arXiv preprint arXiv:2504.08837 (2025)
42. Wang, P., Ling, H.: Svqa-r1: Reinforcing spatial reasoning in mllms via view-consistent reward optimization. arXiv preprint arXiv:2506.01371 (2025)
43. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv:2409.12191 (2024)
44. Wei, Y., Zhao, L., Sun, J., Lin, K., Yin, J., Hu, J., Zhang, Y., Yu, E., Lv, H., Weng, Z., et al.: Open vision reasoner: Transferring linguistic cognitive behavior for visual reasoning. arXiv preprint arXiv:2507.05255 (2025)
45. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025)
46. Wu, H., Huang, X., Chen, Y., Zhang, Y., Wang, Y., Xie, W.: Spatialscore: Towards unified evaluation for multimodal spatial understanding. arXiv e-prints pp. arXiv–2505 (2025)
47. Wu, J., Guan, J., Feng, K., Liu, Q., Wu, S., Wang, L., Wu, W., Tan, T.: Reinforcing spatial reasoning in vision-language models with interwoven thinking and visual drawing. arXiv:2506.09965 (2025)
48. x.ai: Grok-1.5 vision preview (2024), https://x.ai/blog/grok-1.5v
49. Xia, J., Zang, Y., Gao, P., Li, S., Zhou, K.: Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. arXiv preprint arXiv:2505.14677 (2025)
50. 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)
51. Yang, R., Zhu, Z., Li, Y., Huang, J., Yan, S., Zhou, S., Liu, Z., Li, X., Li, S., Wang, W., et al.: Visual spatial tuning. arXiv preprint arXiv:2511.05491 (2025)
52. Yang, S., Yang, J., Huang, P., Brown, E., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., et al.: Cambrian-s: Towards spatial supersensing in video. arXiv preprint arXiv:2511.04670 (2025)

53. 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)
54. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025)
55. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
56. 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)
57. Yu, E., Lin, K., Zhao, L., Yin, J., Wei, Y., Peng, Y., Wei, H., Sun, J., Han, C., Ge, Z., et al.: Perception-r1: Pioneering perception policy with reinforcement learning. arXiv preprint arXiv:2504.07954 (2025)
58. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. arXiv:2503.14476 (2025)
59. Zhang, X., Wang, J., Cheng, Z., Zhuang, W., Lin, Z., Zhang, M., Wang, S., Cui, Y., Wang, C., Peng, J., et al.: Srpo: A cross-domain implementation of large-scale reinforcement learning on llm. arXiv preprint arXiv:2504.14286 (2025)
60. Zhou, J., Ji, J., Chen, B., Sun, J., Chen, W., Hong, D., Han, S., Guo, Y., Yang, Y.: Generative rlhf-v: Learning principles from multi-modal human preference. arXiv preprint arXiv:2505.18531 (2025)
61. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv:2504.10479 (2025)