SpatialThinker: Reinforcing 3D Reasoning in Multimodal LLMs via Spatial Rewards

1 Extended Abstract

Multimodal large language models (MLLMs) have advanced rapidly in vision–language tasks, yet they remain weak in 3D spatial reasoning, which is essential for embodied AI applications such as robotics, navigation, and augmented reality. Existing spatial MLLMs often depend on massive datasets, explicit 3D inputs, or architectural modifications, and when reinforcement learning (RL) is applied, it typically uses sparse rewards that provide little guidance for grounded reasoning. We propose SpatialThinker, a 3D-aware MLLM that learns to observe, localize, and reason about object relations through structured grounding and dense reward optimization.

SpatialThinker introduces two key contributions: (1) **STVQA-7K**, a high-quality dataset of 7.5K spatial VQA pairs generated from Visual Genome scene graphs, enriched with 34 additional spatial relations and aligned to localized subgraphs; and (2) a **multi-objective dense spatial reward** with lexicographic gating. The reward integrates four components aligned with human-like spatial reasoning stages: format (structured outputs with tags and valid JSON scene graphs), count (guides focus to task-relevant regions and mitigates reward hacking from excessive object predictions), accuracy (final answer correctness), and spatial (CIoU-based bounding box supervision gated on correctness). Rewards are optimized via Group Relative Policy Optimization (GRPO) (Shao et al., 2024; DeepSeek-AI et al., 2025), providing stable learning from dense signals without requiring critic networks. This design enforces a pipeline of observe $\rightarrow$ localize $\rightarrow$ think $\rightarrow$ answer.

We train SpatialThinker on Qwen2.5-VL backbones (3B and 7B) using only RGB images and the 7K STVQA samples, without any supervised fine-tuning. Despite this minimal training, the models achieve substantial improvements over both supervised and sparse-RL baselines across six benchmarks: CV-Bench, 3DSRBench, MMVP, SpatialBench, and RealWorldQA. Table 1 summarizes performance on the major benchmarks. Compared to sparse RL training, dense spatial rewards nearly doubles the base-model gain (+6.5% vs. +3.6%), and matches or surpasses GPT-4o (+12.1% on 3DSRBench). These results showcase the effectiveness of combining spatial supervision with reward-aligned reasoning in enabling robust 3D spatial understanding with limited data and advancing MLLMs towards human-level visual reasoning.

Model	**3DSRBench**	**CV-Bench**		**Avg.**	**BLINK**		**Avg.**
		2D	3D		Spatial Relation	Relative Depth	
Proprietary Models							
GPT-4o	44.3	75.8	**83.0**	**79.4**	82.5	**78.2**	**80.4**
Open-Source General MLLMs							
Qwen2.5-VL-3B	44.0	59.9	60.2	60.1	66.4	54.0	60.2
Qwen2.5-VL-7B	48.4	69.1	68.0	68.6	84.0	52.4	68.2
VLAA-Thinker-7B	52.2	60.8	60.3	60.6	81.2	71.0	76.1
LLaVA-NeXT-8B	48.4	62.2	65.3	63.8	-	-	-
Cambrian-1-8B	42.2	72.3	72	72.2	-	-	-
Open-Source Spatial MLLMs							
RoboPoint-13B	-	-	61.2	-	60.8	61.3	61.1
SpaceThinker-Qwen2.5-VL-3B	51.1	65.1	65.9	65.5	73.4	59.9	66.7
SpaceLLaVA-13B	42.0	-	68.5	-	72.7	62.9	67.8
SpatialBot-3B	41.1	-	69.1	-	67.8	67.7	67.8
Spatial-RGPT-7B w/ depth	48.4	-	60.7	-	65.7	**82.3**	74.0
Method Comparison (Trained on STVQA-7K)							
Qwen2.5-VL-3B + SFT	50.8	53.9	68.4	61.2	65.0	66.9	66.0
Qwen2.5-VL-3B + Vanilla GRPO	50.1	70.6	66.6	68.6	73.4	55.6	64.5
SpatialThinker-3B (Ours)	52.9	71.0	76.3	73.7	81.8	66.9	74.4
Qwen2.5-VL-7B + SFT	53.6	56.1	71.3	63.7	75.5	64.5	70.0
Qwen2.5-VL-7B + Vanilla GRPO	54.7	68.9	76.5	72.7	80.4	75.0	77.7
SpatialThinker-7B (Ours)	**56.4**	**77.7**	**78.7**	**78.2**	**86.0**	72.6	**79.3**

Table 1: Performance over 2D & 3D Spatial Understanding Benchmarks across different model types.

Submitted to WiML Workshop @ 39th Conference on Neural Information Processing Systems (NeurIPS 2025).