ReinforceSight: Spatial-Reasoning VLM via SFT and GRPO

Jaideep Chawla
New York University, New York, USA
jc12751@nyu.edu

***Abstract*—Current multimodal language models struggle with *qualitative* spatial reasoning, routinely mis-classifying simple relations such as *left of* or *behind*. We introduce a compact 3-billion-parameter vision–language model that narrows this gap without sacrificing efficiency. Our three-stage pipeline: (i) Synthetic data generation: inspired by SPATIALVLM to lift 2-D detections from LOCALIZEDNARRATIVES into coarse 3-D scene graphs and sample 10k spatial QA pairs covering 2-D (left/right/above/below), depth ordering, and relative-distance relations; (ii) Supervised fine-tuning (SFT): initialised from Qwen2.5-VL-3B and trained with TRL on the synthetic corpus; (iii) Alignment: applied *Group Relative Policy Optimisation* (GRPO) in `r1_vlm` with a composite reward that balances answer correctness, chain-of-thought fidelity, and self-consistency.**

The resulting model attains 67.6 % accuracy on the qualitative spatial subset of CV-Bench (1 850 items), outperforming the frozen Qwen2.5-VL-3B baseline by +8.2%. Gains are most pronounced on 3-D tasks from the Omni3D split (+12.2%), demonstrating effective depth-aware reasoning.

***Index Terms*—Vision–Language Models, Spatial Reasoning, Reinforcement Learning, Chain of Thought, Synthetic Data Generation**

I. INTRODUCTION

II. INTRODUCTION

Robust spatial understanding involves the ability to parse *where* objects are with respect to one another and to the observer and is foundational for embodied agents, mixed-reality systems, and any vision–language model (VLM) that aspires to reason about real-world scenes. Humans acquire qualitative spatial concepts such as *left of*, *behind*, and *closer than* early in development, generalising them effortlessly across viewpoints and scales. By contrast, large multimodal language models (MLLMs) still struggle with these ostensibly "easy" relations. **CV-Bench** assembles human-verified questions spanning 2-D spatial relations, object counting, depth ordering, and relative distance; leading open-source MLLMs including LLaVA-1.6, Otter, and Qwen-VL all score <60 % on its spatial subset, far below their performance on generic VQA benchmarks. [1].

Geometry-aware VLMs.: A burst of work augments image–text pre-training with geometric signals. **SpatialVLM** injects metric depth tokens distilled from a monocular estimator and instruction-tunes on millions of synthetic spatial QA pairs, yielding strong improvements on depth-ordering tasks [2]. **GRIT** (released with Apple's *Ferret*) curates 1.1 M refer-and-ground samples that encode fine-grained spatial hierarchies, boosting localisation-demanding dialogue performance [3]. **LLaVA-3D** demonstrates that lightly adapting a 2-D backbone with 3-D positional embeddings can close much of the gap to specialised 3-D encoders while retaining 2-D skills [4]. Although effective in-distribution, these systems often memorise dataset biases, provide little transparent reasoning, and generalise poorly to out-of-domain scenes such as Omni3D.

Synthetic spatial supervision.: We build on the **VQASynth** pipeline and open-source implementation of SpatialVLM [2], which synthesises spatial questions from RGB-D reconstructions by: (i) estimating monocular depth, (ii) detecting objects, (iii) lifting them to a coarse 3-D scene graph, and (iv) sampling question templates covering 2-D relations, depth ordering, and relative distance. Unlike prior work that focuses on metric regression, the templates emphasise *qualitative* spatial language, producing 10k high-fidelity QA pairs across LocalizedNarratives [5].

Alignment via Guided Reinforcement Policy Optimisation.: Recent advances in reinforcement learning from human feedback (RLHF) show that relative-ranking objectives can stabilise long-sequence policy updates. **Group Relative Policy Optimisation** (GRPO) [6] replaces scalar value estimation with within-group advantages, enabling low-variance updates without a learned critic. We design a new "environment" [1] in the `r1_vlm` framework and design a composite reward balancing answer correctness, chain-of-thought (CoT) faithfulness, and self-consistency.:contentReference[oaicite:5]index=5

Comprehensive CV-Bench evaluation.: We benchmark on CV-Bench's spatial subset (1 850 questions) that probes both 2-D and 3-D reasoning across ADE20K, COCO, and Omni3D scenes. Our *synthetic-training* + *GRPO* regimen lifts a frozen Qwen2.5VL-3B baseline by **+8.2 percentage points** overall. Qualitative analysis shows that the model now articulates correct geometric rationales instead of guessing object frequencies.

A. Contributions

- Develop a synthetic data generation pipeline using VQASynth and LocalizedNarratives to create qualitative spatial reasoning questions with chain-of-thought explanations
- Implement a composite reward function for GRPO that balances answer correctness, reasoning quality, and consistency

[1]Concept was first introduced by William Brown as part of his early experiments over GRPO implemented in "verifiers" package https://x.com/willccbb/status/1884067125205356917

- Demonstrate the first application of GRPO to vision-language model alignment for spatial reasoning
- Attain **67.6%** accuracy on CV-Bench qualitative subset.
- Show particularly strong gains on 3D depth reasoning tasks

III. Dataset Construction

Source data. We use the LocalizedNarratives dataset from HuggingFace, which provides images with detailed captions and object annotations. The dataset contains diverse real-world scenes with complex spatial configurations, making it an ideal source for our task.

3D scene construction. For each image in the dataset, we use VQASynth's pipeline to:

- Estimate monocular depth using VGGT
- Detect and segment objects using SAM2
- Generate object embeddings and captions through CLIP and Molmo-2
- Fuse all information into a 3D scene representation

Question generation. Using the 3D scene representation, we generate three types of questions:

- **2D spatial relation**: Questions about left/right/above/below relationships based on objects' 2D positions in the image
- **3D depth ordering**: Questions about in-front-of/behind relationships based on estimated depth
- **3D relative distance**: Questions asking which of two objects is closer/farther to a reference object

Chain-of-Thought generation. For each question, we generate qualitative chain-of-thought explanations that avoid numerical values. For example:

> Looking at the image, I need to determine which object is closer to the camera: the chair or the table. I can see that the chair appears to be positioned more toward the foreground of the image. Objects closer to the camera appear larger relative to their actual size and often partially occlude objects that are further away. Based on these visual cues, I can determine that the chair is in front of the table.

Dataset balancing. We balance the dataset to ensure equal representation across all eight spatial relations (left, right, above, below, in front of, behind, closer, farther). We filter out ambiguous cases where depth differences are minimal to avoid confusing the model.

Dataset statistics. Our final dataset consists of 10,000 image-question-answer-explanation samples, split into train (80%), validation (10%), and test (10%) sets.

IV. Model Architecture

We adopt Qwen2.5-VL-3B-Instruct as our base model which couples a dynamic-resolution, window-attention ViT vision encoder to a 3-billion-parameter Qwen2.5 language decoder through a lightweight projector, all unified by upgraded multi-resolution RoPE positional encoding for end-to-end vision-language generation [7].

During SFT, we freeze the vision encoder and apply Low-Rank Adaptation (LoRA) with rank 16 to all attention projection matrices. We use 8-bit quantization to reduce memory requirements.

The model employs the standard VL architecture with a vision encoder connected to a text decoder via cross-attention mechanisms. We use the chat template provided by Qwen2.5-VL-Instruct to format our inputs as multimodal messages.

V. Training Procedure

Supervised fine-tuning (SFT).: We perform SFT using the TRL library with the following configuration:

- 3 epochs with batch size 64 and sequence length 2,048
- Learning rate 1.5e-4 with AdamW 8-bit optimizer
- Gradient accumulation steps: 4
- Gradient checkpointing and mixed precision (fp16)
- Early stopping based on validation exact match

Reward function.: We implement a composite reward that consists of:

$$R = \lambda_1 \cdot R_{\text{answer}} + \lambda_2 \cdot R_{\text{CoT}} + \lambda_3 \cdot R_{\text{consistency}} \tag{1}$$

where:

- R_{answer}: 1.0 for correct answers, -0.5 for incorrect answers
- R_{CoT}: Score (0-1) based on task-specific keywords, reasoning markers, visual references, and appropriate length
- $R_{\text{consistency}}$: Score (0-1) measuring alignment between the reasoning and the final answer

We use weights $\lambda = (1.0, 0.3, 0.2)$ based on ablation studies.

GRPO training.: We implement GRPO using the r1_vlm framework with:

- 8 rollouts × 9 candidates per prompt
- KL penalty coefficient of 0.12
- Learning rate 3e-5 with AdamW optimizer
- Batch size 32 with gradient accumulation steps of 2
- Checkpointing every 100 updates

VI. Results

We evaluate our model after 2 SFT epochs and 400 GRPO updates on the CV-Bench qualitative spatial subset.

A. Overall Spatial Performance

TABLE I
Overall qualitative spatial performance (CV-Bench).

Model	Q	Accuracy (%)	Δ
Qwen2.5-VL-3B	2 450	59.35	–
Ours (SFT + GRPO)	2 450	**67.57**	+8.22

TABLE II
ACCURACY BY DATASET SOURCE.

Source	Q	Base	Ours	Δ
ADE20K	650	67.70	67.70	0.00
COCO	600	70.19	71.87	+1.67
Omni3D	1 200	54.08	**66.25**	+12.17

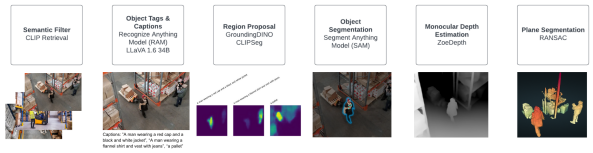

Fig. 1. VQASynth pipeline for spatial reasoning data generation.

B. Performance by Source

C. Key Findings

1) **Overall improvement:** Our approach achieves a significant +8.22 % improvement over the baseline, validating the effectiveness of our SFT+GRPO pipeline.
2) **Strong 3D reasoning:** The largest gains (+12.17%) are on the Omni3D subset, which focuses on 3D spatial relationships, suggesting our approach is particularly effective for depth reasoning.
3) **2D relation plateau:** We observe no improvement on ADE20K, which primarily contains 2D spatial queries. This suggests the baseline model already handles simple 2D relations well, or that our approach needs refinement for 2D tasks.
4) **CoT quality:** Qualitative analysis shows that our model produces more structured and relevant reasoning chains that align with the final answers.

VII. Conclusion

We present a compact vision-language model specialized for qualitative spatial reasoning through synthetic data generation, supervised fine-tuning, and GRPO alignment. Our approach demonstrates significant improvements over the baseline, particularly on challenging 3D spatial reasoning tasks requiring depth understanding. This shows the potential of combining synthetic data generation with reinforcement learning techniques for developing VLMs with stronger spatial reasoning capabilities.

Future work will explore scaling to larger models, incorporating explicit 3D representations, and improving performance on 2D spatial reasoning tasks where our current approach shows limited gains.

Acknowledgement

We thank the LLVM 2025 teaching staff and the NYU HPC team for compute support. We also acknowledge the creators of VQASynth, r1_vlm, and CV-Bench for their valuable open-source contributions.

References

[1] S. Tong, E. L. Brown II, P. Wu, S. Woo, A. J. Iyer, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang, X. Pan, R. Fergus, Y. LeCun, and S. Xie, "Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. [Online]. Available: https://openreview.net/forum?id=Vi8AepAXGy

[2] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. J. Guibas, and F. Xia, "SpatialVLM: Endowing vision–language models with spatial reasoning capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 14 455–14 466. [Online]. Available: https://arxiv.org/abs/2309.15569

[3] H. You, H. Zhang, Z. Gan, X. Du, B. Zhang, Z. Wang, L. Cao, S.-F. Chang, and Y. Yang, "Ferret: Refer and ground anything anywhere at any granularity," in *International Conference on Learning Representations (ICLR)*, 2024, introduces the GRIT dataset. [Online]. Available: https://arxiv.org/abs/2310.07704

[4] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu, "LLaVA-3D: A simple yet effective pathway to empowering lmms with 3d-awareness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. [Online]. Available: https://arxiv.org/abs/2409.18125

[5] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari, "Connecting vision and language with localized narratives," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020, pp. 647–665. [Online]. Available: https://arxiv.org/abs/2009.11539

[6] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, and *et al.*, "DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[7] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5–VL Technical Report," 2025. [Online]. Available: https://arxiv.org/abs/2502.13923