

Spatial-SSRL: Enhancing Spatial Understanding via Self-Supervised Reinforcement Learning

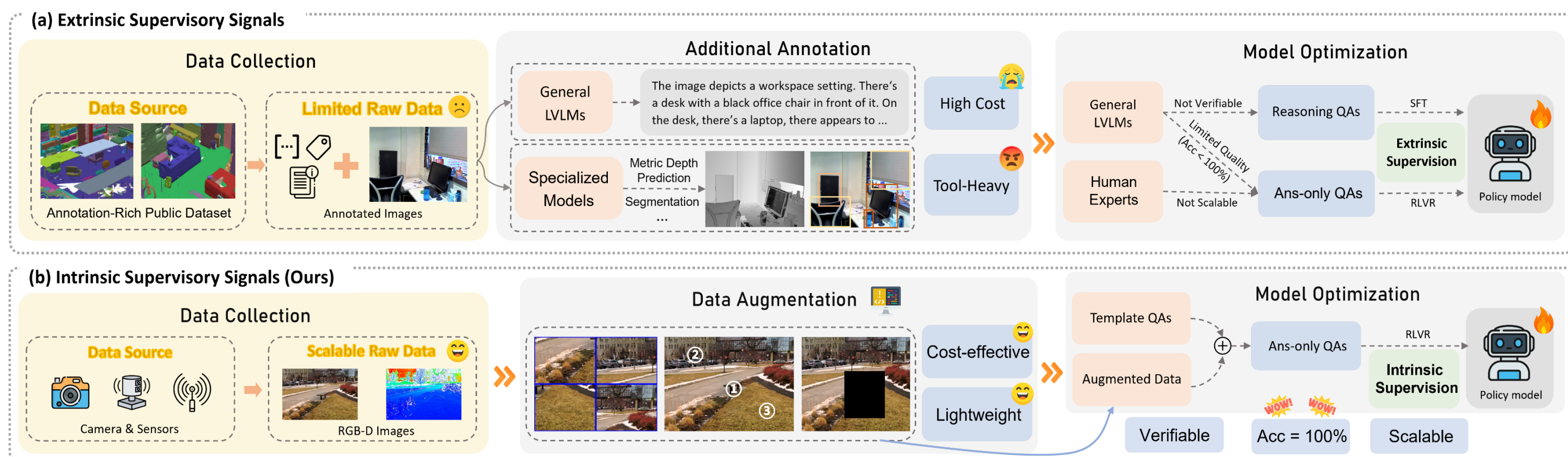
Yuhong Liu^{1,2} Beichen Zhang^{1,3} Yuhang Zang^{1*} Yuhang Cao¹ Long Xing¹

Xiaoyi Dong¹ Haodong Duan¹ Dahua Lin^{1,3} Jiaqi Wang^{1,4*}

¹Shanghai AI Laboratory ²Shanghai Jiao Tong University ³The Chinese University of Hong Kong ⁴Shanghai Innovation Institute

Motivation

Spatial-SSRL is a novel **self-supervised RL paradigm** aimed at enhancing LVLM spatial understanding.



Prior approaches for spatial understanding rely heavily on annotation of external tools, which inherit errors in training data from these tools and incur additional cost!

Advantages:

Highly Scalable: We use ordinary raw RGB and RGB-D images instead of richly-annotated public datasets or manual labels for data curation.

Cost-effective: We avoid the need for human labels or API calls of proprietary models throughout the entire pipeline.

Lightweight: Spatial-SSRL is completely tool-free and can easily integrate additional tasks.

Naturally Verifiable: Intrinsic supervisory signals determined by pretext objectives are naturally verifiable, aligning Spatial-SSRL well with the powerful RLVR paradigm.

We welcome you to join Spatial-SSRL with effective pretext tasks to further strengthen the (spatial) capabilities of LVLMs!

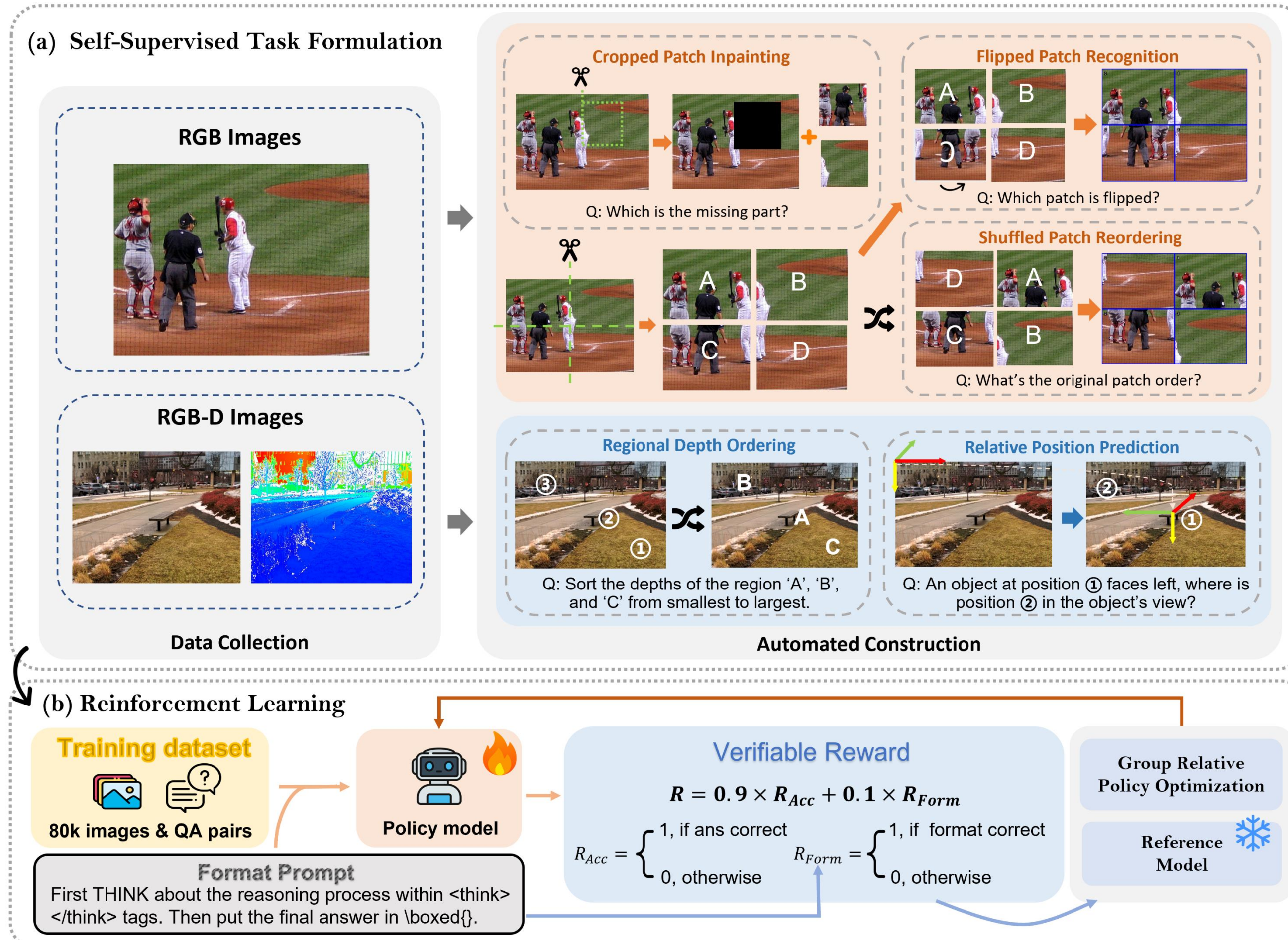
Method

Stage 1: Task Formulation

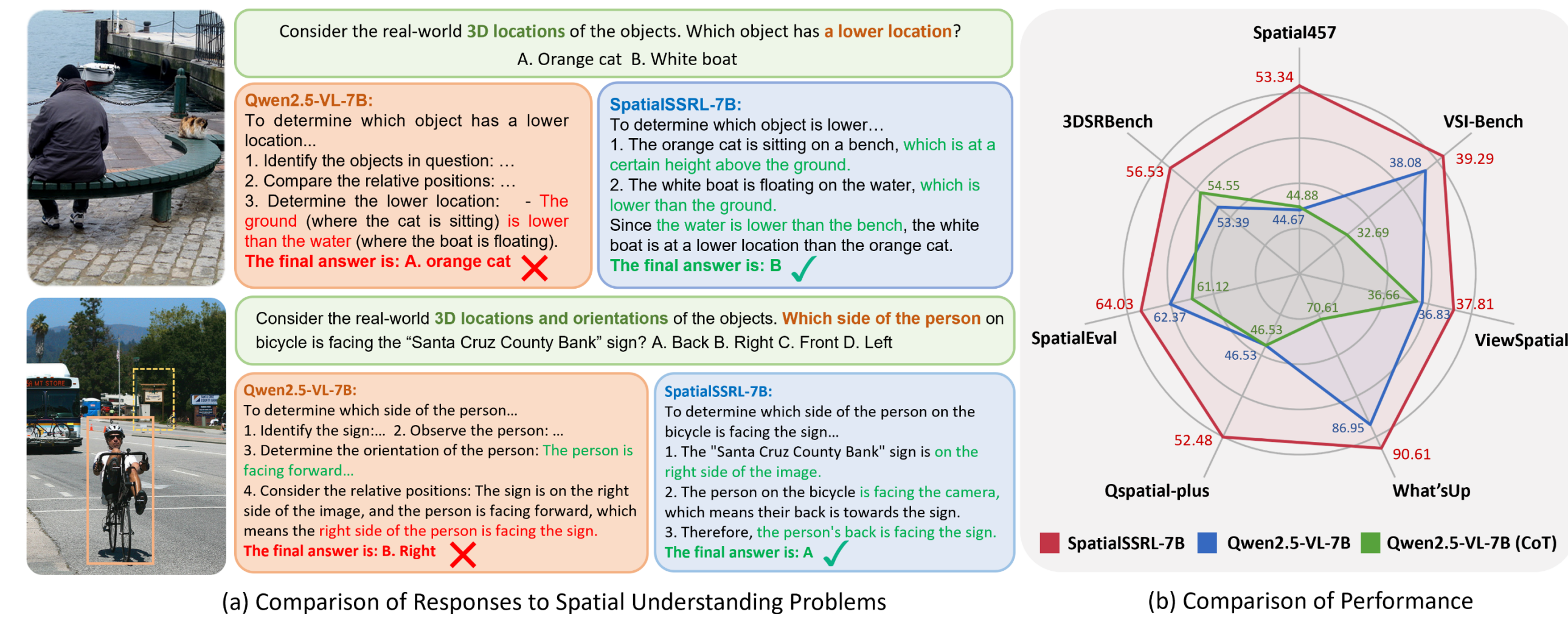
- Depth-free Tasks: Cropped Patch Inpainting, Shuffled Patch Reordering, Flipped Patch Recognition
- Depth-based tasks: Regional Depth Ordering, 3D Relative Position Prediction

Stage 2: LVLM Post-Training

- SFT Coldstart (~3.6k samples)
- GRPO Training on Spatial-SSRL-81k dataset



Experiment



Perspective Transformation



Q: Consider the real-world 3D locations and orientations of the objects. If I stand at the recreational vehicle's position facing where it is facing, is the dog in front of me or behind me? A. behind B. in front of

Qwen2.5-VL-7B:

<think>\n\nTo determine if the dog is in front of or behind the recreational vehicle (RV) when standing at its position, we need to consider the orientation and relative positions of the RV and the dog.\n\n1. The RV is positioned with its front facing towards the left side of the image.\n2. The dog is standing on the ground in the foreground, closer to the viewer than the RV.\n3. If you were to stand at the RV's position and face forward, the dog would be in your line of sight and closer to you than the RV.\n\nGiven this information, the dog is in front of the RV when you are standing at its position.\n\nThe final answer is: \boxed{B}"

Spatial-SSRL-7B:

<think>\n\nTo determine the position of the dog relative to the recreational vehicle, I need to consider the orientation of the RV and the dog's position in the image.\n\n1. The recreational vehicle is facing away from the camera, as indicated by the direction of its wheels and the orientation of the trailer.\n2. The dog is positioned in the foreground, closer to the camera than the RV.\n3. Since the dog is in front of the RV from the camera's perspective, it is behind the RV from the RV's perspective.\n\nTherefore, the dog is behind the recreational vehicle if I stand at the RV's position facing where it is facing.\n\n</think>\n\n\boxed{A}

Further Information

Arxiv Paper: <https://arxiv.org/abs/2510.27606>

Homepage: <https://github.com/InternLM/Spatial-SSRL>

Email: liuyuhong1@pjlab.org.cn