Aligning Cross-View Visual Geometries in LVLMs Through Human-Like Reasoning Learning

Yuming Qiao[1*], Liang Luo[1*], Dan Meng[1†], Yifan Yang[1,2], Qingyuan Wang[3,4], Juntuo Wang[1,5], Yuwei Zhang[1], Ru Zhen[3], Yanhao Zhang[3], Haonan Lu[3], Xudong Zhang[1]

[1]OPPO Research Institute
[2]Shanghai Normal University
[3]OPPO AI Center
[4]Xidian University
[5]Brown University
qym.2021@tsinghua.org.cn, luoliang1@oppo.com, mengdan90@163.com

Abstract

Spatial understanding is a critical capability for LVLMs (Large Vision-Language Models) to advance embodied AI applications. Existing works primarily focus on enhancing spatial understanding within a single frame, i.e., injecting 3D spatial concepts into LVLMs under single coordinate system. However, such improvements struggle in real-world tasks that require consistent cross-view spatial reasoning. In this paper, we propose CVVG-Reasoner(Cross-View Visual Geometries) that lifts single-frame spatial comprehension to unified cross-view spatial understanding by mimicking human-like cross-view reasoning mechanisms. First, we introduce MV3DSR(Multi-View 3D Spatial Reasoning), a scalable pipeline for cross-view spatial reasoning data generation, and construct MV3DSR-Dataset, a large-scale dataset with diverse 3D cross-view reasoning tasks. Based on MV3DSR, we propose MV3DSR-Bench, a comprehensive benchmark for evaluating cross-view spatial reasoning capabilities. Second, we design a three-stage training strategy: the first two stages progressively equip the model with (1) fundamental spatial knowledge and (2) human-like cross-view reasoning patterns, while the final stage employs reinforcement learning to further boost its performance. Extensive experiments demonstrate that our CVVG-Reasoner significantly outperforms existing 3D LLMs(Large Language Models) and advanced LVLMs in cross-view tasks while maintaining robust performance on out-of-domain data. Ablations further reveal that injecting human-like reasoning patterns yields 44% performance gain, validating the effectiveness of our design.

Introduction

In recent years, Large Vision-Language Models(LVLMs) have advanced rapidly, demonstrating remarkable performance across a wide range of visual tasks and significantly expanding the practical applications of AI assistants(Li et al. 2024; Lin et al. 2024; Chen et al. 2024b; Abdin et al. 2024; Bai et al. 2025; Team et al. 2025; Zhang et al. 2024a). However, most existing LVLMs remain constrained to 2D image understanding, leading to suboptimal performance in real-world interactive scenarios such as navigation and robotics(Nahavandi et al. 2025; Zeng et al. 2023). To bridge this gap, equipping LVLMs with human-like spatial reasoning capabilities enabling seamless 2D to 3D spatial understanding has emerged as a critical research topic.

Existing works mainly attribute LVLMs' limited spatial reasoning to the scarcity of 3D domain-specific datasets (Cai et al. 2024; Cheng et al. 2024a; Hong et al. 2023; Zheng, Huang, and Wang 2025), focusing on 3D data construction to enhance spatial understanding (Chen et al. 2024a; Ma et al. 2025b). However, these datasets typically feature single camera coordinate systems, easy to scale but restrictive for cross-view reasoning. Moving beyond static single-view comprehension, cross-view spatial understanding emerges as a critical capability for real-world deployment.

Recent works have evaluated multi-view understanding in LVLMs(Wu et al. 2025; Xu et al. 2025; Yang et al. 2025), revealing persistent challenges in cross-view spatial reasoning despite extensive 3D training data(Yeh et al. 2025; Li et al. 2025b). This highlights that cross-view reasoning in scenes with multiple camera coordinate systems presents inherent limitations. To address this, we propose a paradigm inspired by human cognition. Our framework incorporates bio-inspired spatial reasoning mechanisms, enhancing cross-view geometry perception and coordinate alignment without compromising model generalizability.

Analyzing human cognition in multi-view spatial reasoning(Burgess 2008; Denis and Loomis 2007), we decompose cross-view geometric thinking into two levels: low-level and high-level. At the **low-level**, humans infer 3D spatial information within single-view contexts—a foundational capability aligned with prior single-coordinate spatial understanding works. At the **high-level**, humans integrate multi-view observations to achieve consistent spatial reasoning. Inspired by this, we propose to enhance LVLMs through: (1) **Single-view spatial understanding (low-level)**: extracting basic spatial features from individual views. (2) **Cross-view spatial reasoning (high-level)**: aligning cross-coordinate spatial features via human-like geometric reasoning.

[*]These authors contributed equally.
[†]Corresponding Author

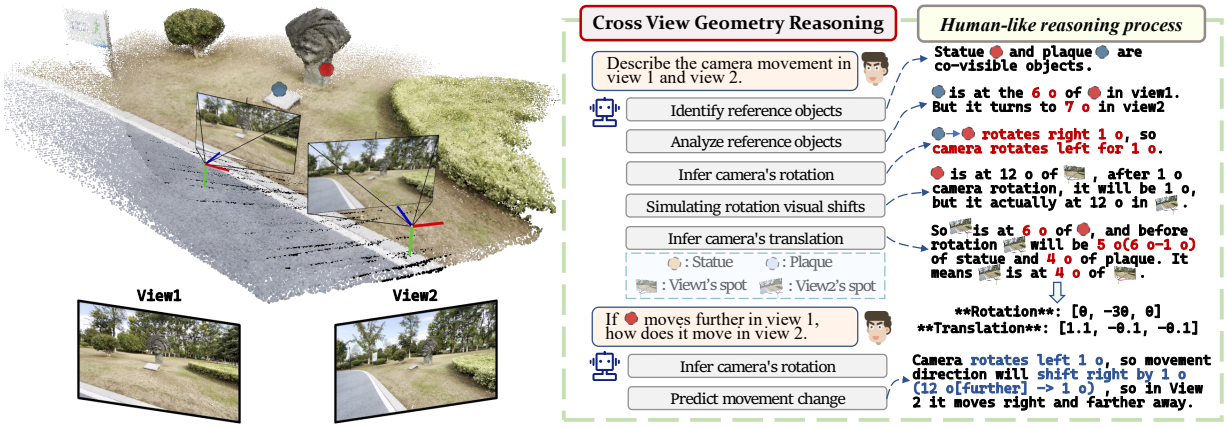

Figure 1: Our CVVG-Reasoner aligns cross-view geometric information to enable complex spatial reasoning across viewpoints.

Motivated by this, we first propose **MV3DSR**, a novel framework for generating both single-view and cross-view 3D spatial reasoning data from arbitrary continuous videos. Next, we introduce a three-stage training strategy: (1) Single-view fine-tuning to establish foundational spatial perception. (2) Cross-view alignment to incorporate cross-view human-like reasoning patterns into trained models. (3) Reinforcement learning to further boost model performance, resulting in our **CVVG-Reasoner** as shown in Fig. 1. For evaluation, we introduce MV3DSR-Bench, a comprehensive benchmark with 4,500 questions covering single-view and cross-view 3D understanding tasks across diverse indoor/outdoor scenes, ensuring robust assessment. Our contributions can be summarized as follows:

- We propose a scalable pipeline for constructing cross-view spatial reasoning datasets, enabling efficient data generation from arbitrary continuous wild videos.
- We introduce CVVG-Reasoner, a novel reasoning-augmented LVLM that mimics human-like thinking to solve complex cross-view spatial reasoning tasks.
- Our CVVG-Reasoner delivers strong results across multiple benchmarks, enhancing both single-view and cross-view spatial reasoning. Experiments confirm the effectiveness of our human-inspired reasoning, data framework, and training strategy.

Related Work

Explicit 3D Spatial Understanding. A direct way to enhance LVLMs' spatial reasoning is to integrate external 3D foundation models for explicit 3D priors. For example, SpatialPIN(Cheng et al. 2024b) reconstructs 3D scenes using 2D segmentation, depth estimation, and camera parameters, then prompts LVLMs with these priors. SpatialScore(Wu et al. 2025) employs SpatialAgent, which dynamically selects 3D tools via Plan-Execute and ReAct mechanisms for spatial reasoning. While these methods improve performance, they depend heavily on external 3D models, lacking end-to-end implicit reasoning. This limits their generalizability when precise 3D priors are unavailable. Therefore, injecting 3D knowledge into model weights to achieve end-to-end spatial reasoning represents a promising direction for enhancing both model generalizability and efficiency.

Single-view Spatial LVLMs. Single-view spatial LVLMs improve spatial understanding by injecting single-coordinate 3D spatial concepts into LVLMs. For instance, SpatialVLM(Chen et al. 2024a) employs tagging, grounding, segmentation, and depth estimation to construct 3D datasets and train 3D LVLM. SpatialRGPT (Cheng et al. 2024a) further employs region features to improve object localization. SpatialBot(Cai et al. 2024) incorporates depth modality to boost spatial reasoning. SpatialLLM(Ma et al. 2025b) identifies 3D-informed data as key for spatial understanding. SpatialReasoner(Ma et al. 2025a) combines explicit 3D representations with reinforcement learning. While these methods advance single-view reasoning, they lack cross-view spatial reasoning capabilities.

Cross-view Understanding Evaluation. The lack of cross-view spatial reasoning in LVLMs remains a major limitation. Benchmarks like STI-Bench(Li et al. 2025b) and ViewSpatial-Bench(Li et al. 2025a) reveal that even top-performing LVLMs struggle with tasks like object motion prediction and multi-view localization. All-Angles Bench(Yeh et al. 2025), using annotated EGO4D-EXO(Grauman et al. 2024) and EgoHumans(Khirodkar et al. 2023) data, further shows their weakness in cross-coordinate reasoning (e.g., object manipulation and camera pose prediction), highlighting the need for improvement. To overcome these limitations, we first propose MV3DSR, a scalable pipeline for generating large-scale cross-coordinate spatial reasoning datasets with human-like logic. Based on this, we develop CVVG-Reasoner, a novel LVLM fine-tuned with cross-view cognitive patterns, significantly improving spatial understanding across spatial coordinates.

Methods

CVVG-Reasoner is a powerful LVLM equipped with advanced cross-view spatial reasoning capabilities, enabling it to tackle complex cross-coordinate reasoning tasks through human-like cognitive processes. The key to achieving this

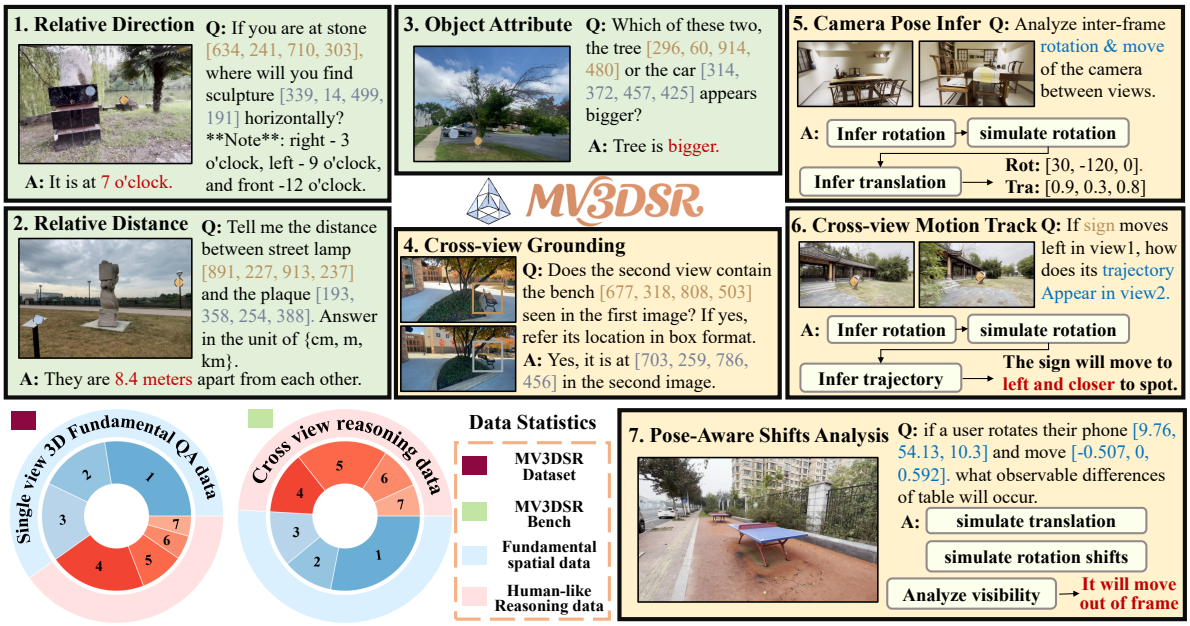

Figure 2: Task definition and data statics of our proposed **MV3DSR-Dataset** and **MV3DSR-Bench**.

lies in two aspects: (1) an effective cross-view reasoning dataset, and (2) a well-designed task paradigm that enables the model to learn cross-spatial coordinate reasoning logic. In this section, we elaborate on three crucial components: (1) We explain how to design dataset tasks to facilitate the model's learning of cross-view spatial reasoning logic. The motivation and task definition of our MV3DSR-Dataset are detailed in Dataset section. (2) We present our automated pipeline for constructing cross-view spatial understanding datasets in dataset pipeline section. (3) We detail the training strategy of CVVG-Reasoner in model training section, demonstrating how to effectively utilize the cross-view spatial reasoning dataset for model training.

MV3DSR-Dataset

Cross-view spatial reasoning is an extremely challenging task. We first examined the process of humans : individuals initially observe the environmental context and objects from distinct perspectives, then select reference targets through matching common objects across perspectives to conduct reasoning about spatial perspective transformations. Accordingly, we decompose the reasoning process into two modes: the low-level thinking mode and the high-level thinking mode, leading us to construct the MV3DSR Dataset comprising two corresponding components. Fig. 2 illustrates the overall task definition and data distribution. Taking cross-view camera pose reasoning as an example (shown in Fig. 3). In low-level thinking, humans first separately understand and analyze the 3D spatial information from two views. In high-level thinking, humans correlate 3D spatial information across views, thereby progressively inferring camera rotation and translation. The reasoning proceeds as follows: (i) **establishing reference points through cross-view matching of co-occurring objects**. Then, since

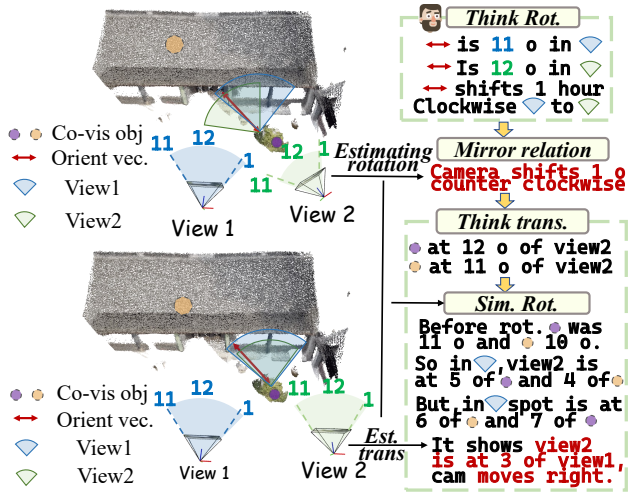

Figure 3: Illustration of human cognitive process in estimating cross-view camera movements.

rotation and translation have compounded effects, they must be analyzed separately.This leads to (ii) **estimating rotation from relative positional changes**, and (iii) **deriving translation after eliminating the effects of rotation**. As in Fig. 3, the 11→12 o'clock shift of co-occurring objects signals 1-hour CCW yaw. Then reverse-simulate the yaw to isolate translation: in View2 the spot sits at 5 o'clock to the building and 6 o'clock to the plant; pre-rotation these bearings were 4 o'clock and 5 o'clock. The intersection of the 4 o'clock (building) and 5 o'clock (plant) rays shows the camera itself slid rightward—i.e., toward 3 o'clock from View1.

Task Definitions. By analyzing the aforementioned cog-

nitive process, we systematically categorize the tasks included in the dataset into two major classes: (1) Single-View 3D Fundamental Data (Inspired by Low-Level Thinking), designed to enhance the model's ability to extract 3D information from individual images. (2) Cross-View Human-Like Reasoning Data (Sourced from High-Level Thinking). Aimed at improving the model's geometric reasoning capabilities, including reference object identification, joint analysis of cross-view geometric transformations, and viewpoint change inference based on visual geometric cues. Fig. 2 shows samples of these two major categories, along with their seven subcategories: (1) ***Relative Direction:*** Predicts horizontal and vertical 12-clock orientations (30° resolution) for fine-grained inter-object spatial relationships and coarse directional labels (e.g., left, right, up, down) for coarse inter-object spatial relationships. (2) ***Relative Distance:*** Estimates absolute distances between objects or object-to-camera spot for translation reasoning. (3) ***Object Attribute:*** Analyzes physical dimensions/comparisons to strengthen single-view 3D reasoning. (4) ***Cross-View Grounding.*** Matches and localizes co-occurring objects across views via bounding box prediction, which is a critical step in high-level thinking. (5) ***Camera Pose Estimation.*** Infers relative camera movement (coarse: directions; fine: degrees/units). (6) ***Cross-View Motion Tracking.*** Projects object motion trajectories between views using inferred spatial transforms. (7) ***Pose-Aware Shifts Analysis.*** Simulates object-level visual changes (position/occlusion/scale) from given camera poses alone without actually observe the altered view. This task represents a higher-order cross-view spatial reasoning challenge, requiring the model to abstractly infer object-level visual changes purely through geometric reasoning. In all cross-view tasks, camera pose is defined relative to the camera coordinate system. For a more detailed task definition and construction methodology, please refer to Appendix A-1.

MV3DSR Data Generation Pipeline

As illustrated in Fig. 4, our proposed MV3DSR framework automatically generates spatial QA data for all task types in the MV3DSR-Dataset, taking arbitrary continuous wild videos as input. The MV3DSR data generation pipeline consists of three components: (1) 2D & 3D Context Annotation. (2) 2D & 3D Context Alignment. (3) Dataset Generation.

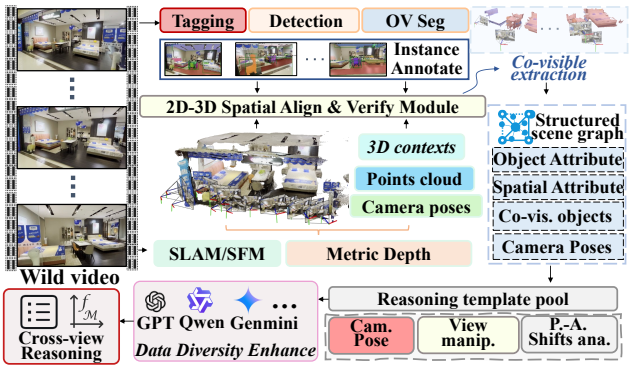

Figure 4: **MV3DSR** data generation pipeline

2D & 3D Annotation. In **2D Annotation,** we adopt the 2D object annotation pipeline from SpatialVLM(Chen et al. 2024a) and SpatialRGPT(Cheng et al. 2024a). Specifically, we employ RAM (Recognize Anything Model)(Zhang et al. 2024b) for object tagging in each video frame. Applying Grounding Dino(Liu et al. 2024) coupled with SAM(Kirillov et al. 2023) to annotate bounding boxes (C_{2d}^{box}) and instance masks (C_{2d}^{mask}) for all detected objects. In **3D Annotation,** we leverage off-the-shelf 3D expert models to reconstruct the 3D scene from the input video. Specifically, we employ VGGT(Wang et al. 2025) to estimate per-frame camera extrinsics (E_i), intrinsics (K_i), and 3d scene point clouds (P_{3d}). Then UniDepth(Piccinelli et al. 2024) is applied to convert relative depth into metric scale, enabling real-world distance measurements in the reconstructed 3D space.

2D & 3D contexts alignment. Building upon the frame-wise 2D annotations (C_{2d}^{box}, C_{2d}^{mask}) and 3D scene information (E_i, K_i, P_{3d}), we propose a 2D-3D alignment module to establish correspondences between 2D and 3D annotations, ultimately generating a cross-view scene graph. We first **select image pairs via 3D points IoU.** Given the camera intrinsics (K_i, K_j) and extrinsics (E_i, E_j)(world-to-camera matrices) for frames F_i and F_j, we first project the 3D point cloud P_{3d} into both camera coordinate systems and compute their 2D pixel coordinates (I_i, I_j) using the intrinsic K_i and K_j as shown in Eq. 1.

$$\mathbf{p}_{3d}^{F_i} = (\mathbf{E}_i) \begin{bmatrix} \mathbf{p_{3d}} \\ 1 \end{bmatrix}, I_i = \frac{\mathbf{K}_i}{\mathbf{p}_{3d}^{F_i}[2]} p_{3d}^{F_i}. \quad (1)$$

Then we extract the subsets of P_{3d} visible in F_i and F_j via judging $0 < \{I_i, I_j\} < \{u, v\}$, u and v are width and height of the image. Visible points are denoted as P_i and P_j, respectively. Image pairs are selected by computing the Intersection-over-Union (IoU) between P_i and P_j, image pairs with IoU > 0.3 are selected as valid samples for training data generation. We then employ a similar approach to align the 2D annotations to the 3D space within selected image pairs. We identify co-occurring objects across views by computing the IoU of projected 3D points within corresponding object masks among views, following a similar formulation to Eq. 1. Object pairs exhibiting an IoU greater than 0.4 between their mask regions are designated as co-occurring objects, with the union of their projected 3D points between image pairs serving as the 3D contexts. For non-co-occurring objects, we utilize the single-view projected 3D points as their 3D contexts. Then the parsed data is stored as a **structured cross-view scene graph,** containing Object attributes, Inter-object spatial relationships, Camera poses (extrinsics/intrinsics), and Metric-scale 3D information. For more detailed 2D-3D alignment & verification procedures, please refer to Appendix A-4.

Dataset Generation. For fundamental 3D data, we randomly sample a single frame from the scene graph and compute the 3D distances between objects, as well as their horizontal and vertical clock-relative orientations, to construct relative distance and relative direction data. Additionally, we derive object attributes data(e.g., width and height) from their 3D points. For cross-view grounding

data, we extract co-occurring and non-co-occurring objects from the scene graph to generate two types of data: (1) queries where the target object exists in the second view, and (2) queries where it does not. To generate more complex tasks—such as camera pose estimation, cross-view motion tracking, and pose-aware analysis, we construct advanced cross-view spatial reasoning data based on human-like cognitive patterns. These patterns are derived from the analysis in task definition section and Fig. 3. Based on these patterns, we have constructed human-like cross-view reasoning templates, with detailed implementations provided in Appendix A-3. Since template-generated data may lack diversity, we further employ a strong LLM to rewrite the QA pairs, enhancing the variability and richness of the training dataset.

CVVG-Reasoner Training

We collect video data from DL3DV(Ling et al. 2024), then generate a 300K training dataset using our proposed MV3DSR pipeline. Due to inevitable noise in the annotation pipeline—particularly from the 2D annotation stage (e.g., errors introduced by the tag model and grounding model), we employ five STEM-trained annotators to clean the generated scene graphs. To minimize workload, they are instructed to only remove incorrect annotations without correcting or adding new ones. Thus, the final dataset consists of 100K high-quality samples derived from manually verified scene graphs and 200K unverified samples.

SFT stage. We adopt a coarse-to-fine training strategy. We first train the model on single-view 3D fundamental data (e.g., object attributes, relative distances/directions) and cross-view grounding data to establish foundational 3D spatial perception capabilities. Then fintune the pretrained model (now equipped with basic spatial understanding) using human-inspired cross-view spatial reasoning data, teaching it to leverage low-level spatial cues for high-level, human-like cross-view geometric reasoning.

RL stage. After supervised fine-tuning (SFT), the model acquires foundational spatial perception and high-level cross-view reasoning capabilities. To further enhance its performance, we employ GRPO(Guo et al. 2025) training with task-specific reward functions. For coarse-level tasks (e.g., relative direction, object attributes), we directly reward binary correctness of output orientations/attributes. While for multi-output reasoning tasks (e.g., camera pose estimation, pose-aware shifts analysis, cross-view manipulation), we design a composite reward R_{mul} as shown in Eq.2, $N_{correct}$ is the number of correct answer, N_{error} is the number of wrong answer, N_{miss} is number of missed answers.

$$R_{mul} = N_{correct} - \alpha * N_{error} - \beta * N_{miss} \tag{2}$$

For quantitative output tasks, we leverage task-specified rewards. Relative distance reward R_{dist} and relative direction reward R_{dir} are shown in Eq.3. $\hat{d}$ is predicted distance, d is ground truth. $\Delta\theta$ is angular error with clock unit(1h=30°).

$$R_{\text{dist}}=\begin{cases}1.0, & \hat{d}\in[0.8d,1.25d]\\ 0.5, & \hat{d}\in[0.5d,2.0d]\\ 0.0, & \text{otherwise}\end{cases}, R_{\text{dir}}=\begin{cases}1.0, & \Delta\theta\le 30^\circ\\ 0.0, & \text{otherwise}\end{cases} \tag{3}$$

Visibility reward R_{vis} and IoU reward R_{IoU} of cross-view grounding task are shown in Eq.4.

$$R_{\text{vis}}=\begin{cases}1.0, & \text{correct}\\ 0.0, & \text{wrong}\end{cases}, R_{\text{IoU}}=\begin{cases}\frac{B_{est}\cap B_{GT}}{B_{est}\cup B_{GT}}, & \text{correct}\\ 0.0, & \text{wrong}\end{cases} \tag{4}$$

Rotation reward $\mathbf{R_{rot}}$ and translation reward $\mathbf{R_{tra}}$ in fine-level camera pose tasks are shown in Eq.5-Eq.7. $\mathbf{R}_1$ and $\mathbf{R}_2$ are rotation matrices derived from the predicted and ground-truth Euler angles, respectively, while $\mathbf{t}_1$ and $\mathbf{t}_2$ denote the predicted and ground-truth translation vectors.

$$\mathbf{R}_{\text{rel}} = \mathbf{R}_1^T\mathbf{R}_2,\ \mathbf{t}_{\text{rel}} = \mathbf{R}_1^T(\mathbf{t}_2-\mathbf{t}_1) \tag{5}$$

$$\text{tr}(\mathbf{R}_{\text{rel}}) = r_{11}+r_{22}+r_{33},\ \theta = \arccos\left(\frac{\text{tr}(\mathbf{R}_{\text{rel}})-1}{2}\right) \tag{6}$$

$$\mathbf{R_{rot}} = 1-\frac{\theta}{\pi},\ \mathbf{R_{tra}} = 1-\frac{\|\mathbf{t}_{\text{rel}}\|_2}{\|\mathbf{t}_{\text{rel}}\|_2+1.5} \tag{7}$$

Experiment

To rigorously evaluate CVVG-Reasoner's cross-view spatial reasoning capabilities, we construct the MV3DSR-Bench, containing 4,500 samples from 500 held-out videos. The benchmark maintains a 1:1 ratio between single-view and cross-view tasks, with all non-quantitative tasks framed as single/multiple-choice questions (contrasting with training's free-form QA). It is notable that single-choice tasks have random baseline while multiple-choice tasks exhaustive correct combinations with no random baseline exists. Meanwhile, evaluation uses precise 1° rotation annotations (vs. training's 30° discretization) for camera pose estimation.

Implementation Details. Qwen2.5VL exhibits stronger visual grounding capabilities compared to other open-source LVLMs. It can localize objects with absolute bounding boxes in images of arbitrary resolutions, which is crucial for cross-view grounding and object identification. Therefore, we adopt Qwen2.5VL-7B as our base model. During the supervised fine-tuning (SFT) stage, we employ a cosine learning rate schedule with an initial learning rate of 1×10^{-5} and AdamW optimizer to perform full-parameter fine-tuning on the MV3DSR-Dataset for one epoch. Due to computational constraints, the reinforcement learning (RL) stage is conducted on a subset of the MV3DSR-Dataset containing 50K samples, with 8 rollouts per instruction, and trained for one epoch. All experiments are conducted on a single node with 8×A100 80GB GPUs. The SFT stage takes 17 hours, while the RL stage requires 69 hours.

Baselines. We validate our approach using Qwen2.5VL-7B/72B as baselines, alongside widely adopted open-source LVLMs (InternVL3-38B, LLaVA-Video/OneVision-8B) and 3D-specialized models (SpatialRGPT/LLM/Reasoner). Furthermore, we test against commercial models including Gemini 1.5-flash, Claude 3.5-sonnet, and GPT-4o-20241120 to demonstrate that even state-of-the-art LVLMs and 3D-specialized LVLMs exhibit significant limitations in cross-view spatial reasoning tasks. Since many of these models tend to refuse answering specific questions, we employ additional prompts to encourage responses. We also use

Models	*Single-view spatial understanding*						*Cross-view spatial reasoning*							
	Obj Att.	Rel loc.			Rel Dir.		Grounding		Cam. pose			Mot.	P.-A. shifts	
	Ql.↑	th1↑	th2↑	Qt.↓	Ql.↑	Qt.↑	Ql.↑	IoU↑	Ql.↑	RPE↓	RTE↓	Ql.↑	Ql.↑	Qt.↑
Open Source Models														
QwenVL-2.5	0.629	0.159	0.351	6.053	0.726	0.376	0.369	0.076	0.093	1.226	12.509	0.149	0.084	0.236
7B & 72B	0.691	0.183	0.423	2.641	0.797	0.223	0.572	0.327	0.131	0.957	1.871	0.311	0.05	0.172
InternVL-2.5	0.716	0.214	0.445	3.112	0.699	0.21	0.235	0.28	0.059	0.936	3.256	0.122	0.005	0.196
LLaVA video	0.659	0.128	0.266	9.232	0.663	0.239	0.172	0.038	0.013	0.93	1.993	0.228	0	0.075
LLaVA oneV	0.652	0.025	0.05	/	0.683	0.197	0.182	0.016	0.034	1.828	5.728	0.173	0	0.072
Closed Source Commercial Models														
Claude 3.5	0.563	0.095	0.163	11.45	0.577	0.386	0.244	0.01	0.089	0.932	2.66	0.199	0.029	0.184
GPT4o	0.723	/	/	/	0.761	0.48	0.363	0.103	0.11	0.981	2.958	0.31	0.094	0.124
Genmini 1.5	0.733	0.155	0.311	3.803	0.606	0.385	0.254	0.06	0.053	0.967	7.027	0.07	0.032	0.199
3D Specialized Models														
SpatialLLM	0.672	0.14	0.276	6.73	0.69	0.29	0.253	0.08	0.078	0.817	3.054	0.231	0.039	0.217
SpatialReason	0.686	0.257	0.42	4.909	0.719	0.116	0.29	0.042	0.105	0.747	3.01	0.206	0.015	0.207
SpatialRGPT	0.68	0.165	0.373	5.762	0.594	0.497	0.215	0.071	0.053	1.053	1.336	0.25	0.04	0.164
Ours	**0.804**	**0.487**	**0.765**	**0.9**	**0.885**	**0.907**	**0.751**	**0.45**	**0.69**	**0.419**	**0.979**	**0.773**	**0.349**	**0.588**

Table 1: Quantitative experiments on MV3DSR-Bench. Our CVVG-Reasoner achieves remarkable improvements across both single-view and cross-view spatial reasoning tasks.

Models	*STI-Bench*			*All-Angles-Bench*					
	Cam. pose↑	RPE↓	RTE↓	Att. Ide.↑	Rel. Dist.↑	Cam. pose↑	RPE↓	RTE↓	Manipul.↑
QwenVL-2.5-7B	0.108	1.147	5.184	0.627	0.549	0.224	1.845	7.308	0.145
QwenVL-2.5-72B	0.167	1.293	1.555	**0.765**	0.607	0.295	1.414	1.475	0.152
InternVL2.5-38B	0.126	1.052	6.834	0.756	0.507	0.313	1.965	3.202	0.147
LLaVA video 8B	0.103	1.289	1.724	0.655	0.443	0.115	1.595	0.977	0.166
GPT4o-20241120	0.111	1.663	2.913	0.668	0.512	0.358	2.591	12.17	0.263
Genmini 1.5-flash	0.173	1.655	6.276	0.684	0.589	0.438	1.623	2.862	0.22
Claude 3.5-sonnet	0.163	0.965	6.954	0.632	0.553	0.358	1.754	3.537	0.197
SpatialLLM	0.152	1.183	1.267	0.453	0.452	0.284	1.712	1.687	0.28
SpatialReasoner	0.185	0.884	1.084	0.35	0.456	0.309	1.861	2.011	0.293
CVVG-Reasoner	**0.571**	**0.787**	**0.924**	0.742	**0.718**	**0.632**	**0.893**	**0.855**	**0.445**

Table 2: Quantitative experiments on STI-Bench and All-Angles-Bench. Our CVVG-Reasoner shows strong generalization capabilities, achieving robust performance on out-of-domain benchmarks.

Qwen2.5-32B to post-process these models' outputs to ensure compatibility with the evaluation framework.

Quantitative Comparison

We evaluated the performance of our model on MV3DSR-Bench, with the results presented in Table 1. Here, Ql. denotes qualitative results (for single-choice questions, this directly represents accuracy; for multiple-choice questions, it reflects a composite score calculated based on correct, incorrect, and missing options; and for cross-view grounding tasks, it indicates the accuracy of object visibility judgment). Qt. represents quantitative results (for relative distance tasks, it measures the relative error between predicted and ground-truth distances; for relative direction tasks, it reports the accuracy of predicted clock orientations; RPE is the relative rotation error in radians between rotation matrices; and RTE denotes the L2 norm of relative translation error). The results demonstrate that our CVVG-Reasoner achieves marginal improvements across all tasks, with particularly significant gains in cross-view spatial reasoning. For instance, in camera pose estimation, our model outperforms the second-best approach by 81% and 44% in coarse-level and fine-level predictions, respectively. In motion tracking, it achieves a 59% improvement. For the Pose-Aware Shifts Analysis, visibility prediction (Ql.) improves by 73%, while the multiple-choice score for visual shifts of objects increases by 58%. In Fig. 5, we visualize the performance of CVVG-Reasoner and other state-of-the-art LVLMs on cross-view spatial reasoning tasks. The results show that CVVG-Reasoner exhibits robust cross-view geometric reasoning capabilities.

Generalization of CVVG-Reasoner

In addition to our in-domain MV3DSR-Bench, we evaluated the generalization capability of CVVG-Reasoner on two ex-

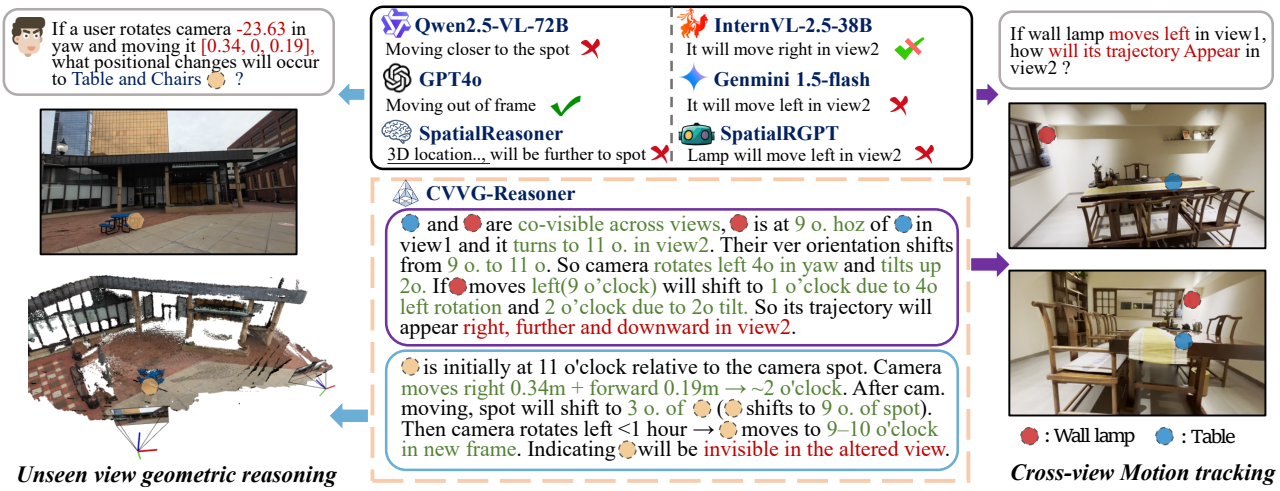

Figure 5: Qualitative cases of CVVG-Reasoner compared to other state-of-the-art models.

ternal benchmarks: STI-Bench and All-Angles-Bench. STI-Bench is a video understanding benchmark, includes camera pose prediction task between two given video frames. However, in this benchmark, poses are directly represented and predicted as RT matrices, which is highly unintuitive for LVLMs. To address this, we reformulated the camera pose data in STI-Bench into a more interpretable format, decomposing the RT matrices into three-axis Euler angles and translation vectors, then we calculated the relative rotation and translation of the queried frame with respect to the initial frame (treated as the world coordinate system). All-Angles-Bench is specifically designed to assess cross-view spatial reasoning, which exhibits strong alignment with our goal. We evaluate CVVG-Reasoner across multiple tasks including Attribute Identification (similar to our cross-view grounding but without bounding box outputs), Relative Distance (measuring changes in an object's distance relative to camera spot across views), Motion Tracking and Camera Pose Estimation. Furthermore, we leveraged the camera pose annotations from Ego-Exo4D to further assess the accuracy of predicted rotation angles and translation matrices. The evaluation results are summarized in Table 2. The results demonstrate that our CVVG-Reasoner exhibits strong generalization capabilities across both benchmarks. Despite differences in question formulations and answer choices compared to the training phase, the model maintains robust cross-view reasoning performance, further validating its adaptability to diverse evaluation settings.

Ablation Study

We posit that the enhanced cross-view spatial reasoning capability of CVVG-Reasoner stems primarily from its human-like geometric reasoning paradigm. To validate the importance of this reasoning process, we conduct ablation studies in this section, with results detailed in Table 3. SFT w/ HR indicates the model supervised fine-tuned exclusively on data containing human-like reasoning (HR) processes. SFT+RL w/ HR is the SFT model further refined via reinforcement learning (RL) while retaining reasoning-augmented data. While SFT+RL w/o HR is the model trained identically but without reasoning process, directly supervised on cross-view task ground truth alone. It can be observed that for cross-view grounding tasks, the inclusion of reasoning processes has negligible impact on performance. But in complex cross-view reasoning tasks (e.g., camera pose estimation), the SFT+RL w/o HR model underperforms even the SFT w/ HR baseline. This regression underscores the critical role of human-like reasoning in boosting performance for cross-view spatial understanding tasks.

Method	CV. Groun.	Cam. pose	Mot.	P.-A. shifts
SFT w/ reasoning	0.331	0.392	0.484	0.385
SFT+RL w/ reasoning	0.45	**0.69**	**0.773**	**0.588**
SFT+RL w/o reasoning	**0.46**	0.384	0.571	0.512

Table 3: Analysis of gain from human-like reasoning patterns and reinforcement learning.

Conclusion

In this paper, We present CVVG-Reasoner, a powerful LVLM for cross-view spatial understanding. Inspired by human cognitive processes, we decompose reasoning into low-level (perceptual) and high-level (conceptual) components. To facilitate training, we propose the MV3DSR data generation pipeline. Our MV3DSR pipeline automatically generates training data for both levels. Leveraging MV3DSR-Dataset, we adopt a three-stage training strategy to progressively equip the model with fundamental spatial understanding and advanced cross-view reasoning capabilities. Extensive experiments on MV3DSR-Bench and external benchmarks show CVVG-Reasoner achieves strong reasoning performance and generalization ability. Furthermore, ablations confirm human-like reasoning patterns enhance complex cross-view task performance.

References

Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R. J.; Javaheripi, M.; Kauffmann, P.; et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Burgess, N. 2008. Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, 1124(1): 77–97.

Cai, W.; Ponomarenko, I.; Yuan, J.; Li, X.; Yang, W.; Dong, H.; and Zhao, B. 2024. Spatialbot: Precise spatial understanding with vision language models. *arXiv preprint arXiv:2406.13642*.

Chen, B.; Xu, Z.; Kirmani, S.; Ichter, B.; Sadigh, D.; Guibas, L.; and Xia, F. 2024a. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14455–14465.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Cheng, A.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024a. SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Cheng, A.; Yin, H.; Fu, Y.; Guo, Q.; Yang, R.; Kautz, J.; Wang, X.; and Liu, S. 2024b. SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Denis, M.; and Loomis, J. M. 2007. Perspectives on human spatial cognition: memory, navigation, and environmental learning. *Psychological Research*, 71(3): 235–239.

Grauman, K.; Westbury, A.; Torresani, L.; Kitani, K.; Malik, J.; Afouras, T.; Ashutosh, K.; Baiyya, V.; Bansal, S.; Boote, B.; et al. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19383–19400.

Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Hong, Y.; Zhen, H.; Chen, P.; Zheng, S.; Du, Y.; Chen, Z.; and Gan, C. 2023. 3D-LLM: Injecting the 3D World into Large Language Models. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Khirodkar, R.; Bansal, A.; Ma, L.; Newcombe, R.; Vo, M.; and Kitani, K. 2023. Ego-humans: An ego-centric 3d multi-human benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19807–19819.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*.

Li, D.; Li, H.; Wang, Z.; Yan, Y.; Zhang, H.; Chen, S.; Hou, G.; Jiang, S.; Zhang, W.; Shen, Y.; et al. 2025a. ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models. *arXiv preprint arXiv:2505.21500*.

Li, Y.; Zhang, Y.; Lin, T.; Liu, X.; Cai, W.; Liu, Z.; and Zhao, B. 2025b. Sti-bench: Are mllms ready for precise spatial-temporal world understanding? *arXiv preprint arXiv:2503.23765*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2024. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5971–5984.

Ling, L.; Sheng, Y.; Tu, Z.; Zhao, W.; Xin, C.; Wan, K.; Yu, L.; Guo, Q.; Yu, Z.; Lu, Y.; et al. 2024. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22160–22169.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, 38–55. Springer.

Ma, W.; Chou, Y.-C.; Liu, Q.; Wang, X.; de Melo, C.; Xie, J.; and Yuille, A. 2025a. Spatialreasoner: Towards explicit and generalizable 3d spatial reasoning. *arXiv preprint arXiv:2504.20024*.

Ma, W.; Ye, L.; de Melo, C. M.; Yuille, A.; and Chen, J. 2025b. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17249–17260.

Nahavandi, S.; Alizadehsani, R.; Nahavandi, D.; Mohamed, S.; Mohajer, N.; Rokonuzzaman, M.; and Hossain, I. 2025. A comprehensive review on autonomous navigation. *ACM Computing Surveys*, 57(9): 1–67.

Piccinelli, L.; Yang, Y.-H.; Sakaridis, C.; Segu, M.; Li, S.; Van Gool, L.; and Yu, F. 2024. UniDepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10106–10116.

Team, K.; Du, A.; Yin, B.; Xing, B.; Qu, B.; Wang, B.; Chen, C.; Zhang, C.; Du, C.; Wei, C.; et al. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wu, H.; Huang, X.; Chen, Y.; Zhang, Y.; Wang, Y.; and Xie, W. 2025. SpatialScore: Towards Unified Evaluation for Multimodal Spatial Understanding. *arXiv preprint arXiv:2505.17012*.

Xu, R.; Wang, W.; Tang, H.; Chen, X.; Wang, X.; Chu, F.-J.; Lin, D.; Feiszli, M.; and Liang, K. J. 2025. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*.

Yang, J.; Yang, S.; Gupta, A. W.; Han, R.; Fei-Fei, L.; and Xie, S. 2025. Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 10632–10643. Computer Vision Foundation / IEEE.

Yeh, C.-H.; Wang, C.; Tong, S.; Cheng, T.-Y.; Wang, R.; Chu, T.; Zhai, Y.; Chen, Y.; Gao, S.; and Ma, Y. 2025. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*.

Zeng, F.; Gan, W.; Wang, Y.; Liu, N.; and Yu, P. S. 2023. Large language models for robotics: A survey. *arXiv preprint arXiv:2311.07226*.

Zhang, J.; Huang, J.; Jin, S.; and Lu, S. 2024a. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5625–5644.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2024b. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1724–1732.

Zheng, D.; Huang, S.; and Wang, L. 2025. Video-3d llm: Learning position-aware video representation for 3d scene understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8995–9006.