Learning Multi-View Spatial Reasoning from Cross-View Relations

Suchae Jeong*[1,2] Jaehwi Song*[2,3] Haeone Lee[1,2] Hanna Kim[1] Jian Kim[4] Dongjun Lee[1]
Dong Kyu Shin[5] Changyeon Kim[1] Dongyoon Hahm[1] Woogyeol Jin[1] Juheon Choi[1] Kimin Lee[1,2]
[1]KAIST [2]Config [3]Hanyang University [4]Yonsei University [5]Seoul National University

https://cross-view-relations.github.io

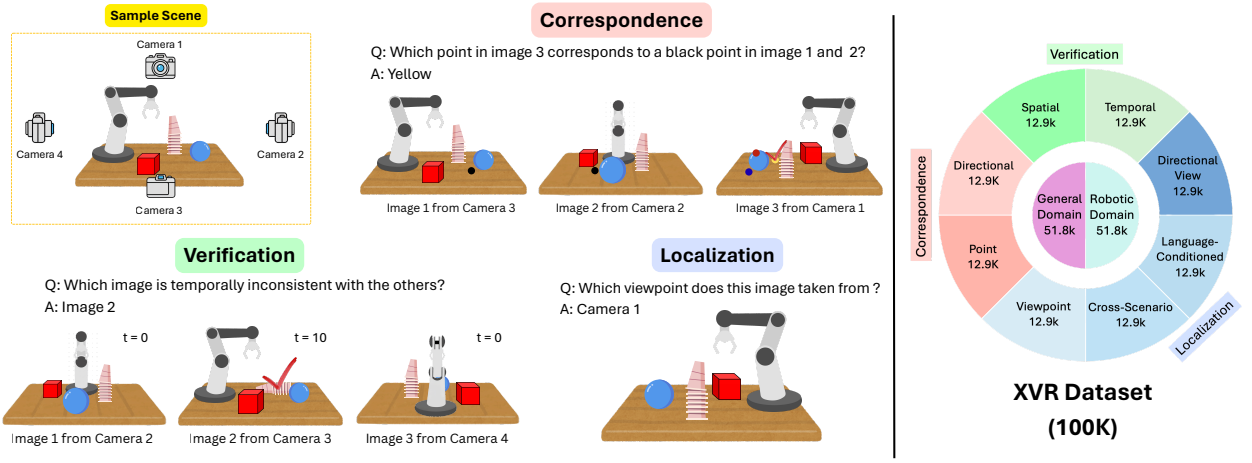

Figure 1. Overview of the Cross-View Relations (XVR). The illustration highlights how multi-view images relate across viewpoints: linking spatial relations (Correspondence), checking cross-view consistency (Verification), and inferring the camera viewpoint (Localization). All XVR dataset samples are derived from real images.

Abstract

Vision-language models (VLMs) have achieved impressive results on single-view vision tasks, but lack the multi-view spatial reasoning capabilities essential for embodied AI systems to understand 3D environments and manipulate objects across different viewpoints. In this work, we introduce Cross-View Relations (XVR), a large-scale dataset designed to teach VLMs spatial reasoning across multiple views. XVR comprises 100K vision-question-answer samples derived from 18K diverse 3D scenes and 70K robotic manipulation trajectories, spanning three fundamental spatial reasoning tasks: Correspondence (matching objects across views), Verification (validating spatial relationships), and Localization (identifying object positions). VLMs fine-tuned on XVR achieve substantial improvements on established multi-view and robotic spatial reasoning benchmarks (MindCube and RoboSpatial). When integrated as backbones in Vision-Language-Action models, XVR-trained representations improve success rates on RoboCasa. Our results demonstrate that explicit training on cross-view spatial relations significantly enhances multi-view reasoning and transfers effectively to real-world robotic manipulation.

1. Introduction

Vision-Language Models (VLMs) have demonstrated strong performance on visual understanding tasks, such as optical character recognition [12, 27, 30, 36], image captioning [32, 48, 57], and video understanding [3, 14, 59, 67]. Recent work has extended these capabilities to spatial reasoning [8, 9, 13, 68], enabling models to reason about object locations, relations, and motion within visual scenes.

However, existing spatial reasoning research has focused almost exclusively on single-view settings. Most VQA datasets and spatial reasoning benchmarks [9, 17, 24, 35, 39, 54, 55, 70] provide only a single viewpoint, which suffers from limited spatial information and frequent occlusions. This is particularly problematic given that multi-camera setups have become standard in robotics applications [11, 16, 19, 26, 29, 37, 38, 41, 42, 49, 52, 56], where understanding geometric relationships between viewpoints is essential for tasks such as manipulation and navigation. While recent work has introduced multi-view datasets [18, 65, 66], these focus primarily on identifying what objects appear in each view, rather than understanding how different viewpoints relate geometrically. With-

*Equal contribution

Dataset	Split	# Imgs / sample	Domain	# Images	# QAs
3DSRBench-real	Eval	1.00	General	2.1K	2.1K
All-Angles-Bench	Eval	4–5	General	450	2.1K
MMSI-Bench	Eval	2.55	General, Robotic	2K	1K
SpatialVLM	Train, Eval	1.00	General	10M	2B
RoboSpatial	Train, Eval	1.00	General	1M	3M
MindCube	Train, Eval	3.37	General	3.2K	21K
MultiSPA	Train, Eval	1.85	General	1.1M	27M
XVR (Ours)	**Train, Eval**	**4.32**	**General, Robotic**	**447K**	**103K**

Table 1. Comparison of spatial reasoning datasets. XVR provides the highest mean images per sample among training datasets, with supervision spanning both general and robotic domains.

out explicit supervision on cross-view spatial relationships, VLMs often generate predictions that appear visually plausible within individual views but are spatially inconsistent across viewpoints.

To address this limitation, we introduce Cross-View Relations (XVR), a dataset of 100K multi-view VQA samples that provides explicit supervision on geometric relationships across viewpoints. Drawing inspiration from Structure-from-Motion (SfM) [46, 50], we design three reasoning primitives that capture how views relate geometrically: (i) Cross-view Correspondence: identifying matching elements across views, (ii) Geometric Consistency Verification: validating whether view relationships are geometrically plausible, and (iii) Relative Viewpoint Localization: reasoning about spatial relationships between camera perspectives (see Figure 1).

To construct XVR at scale, we leverage two complementary data sources. Calibrated multi-view captures (the general domain) provide dense geometric supervision with accurate camera parameters, enabling precise correspondence and consistency annotations. Robotic trajectories (the robotic domain) contribute temporal continuity and diverse viewpoint transitions from embodied interactions, enriching the dataset with dynamic perspective changes. Together, these sources provide the geometric precision and viewpoint diversity needed for comprehensive cross-view reasoning.

Evaluation across ten VLMs (both open-source [4, 34, 58, 63] and closed-source models [1, 2, 15, 45]) demonstrates substantial improvements: models trained with XVR achieve a 1.8× relative gain in accuracy on XVR-Eval (our internal benchmark) and show consistent improvements on external benchmarks including MindCube-Tiny [66] and RoboSpatial-Home [55]. Furthermore, when XVR-trained VLMs serve as backbones for Vision-Language-Action (VLA) models, they yield significant gains, improving manipulation success rates on simulated environments from RoboCasa [44] by an average of 13% absolute. This demonstrates that cross-view relation reasoning transfers effectively to real-world robotic control.

Our contributions are summarized as follows:

- We introduce XVR, a dataset with explicit supervision on cross-view relations for multi-view spatial reasoning.
- XVR contains 100K samples spanning two complementary domains, i.e., general scenes and robotic trajectories, organized into three task categories (Correspondence, Verification, and Localization) across eight specific tasks.
- We show that training on XVR improves performance on XVR-Eval, transfers to external multi-view and robotic spatial benchmarks, and enhances downstream VLA manipulation performance.

2. Related Work

Single-view Spatial Reasoning Spatial reasoning research has primarily focused on single-view settings. Early work established baselines on synthetic scenes [25] and extended them to real images with relational structure [21]. Subsequent studies exposed failures in directional reasoning [35], distance estimation [54], and frame-of-reference understanding [17, 24]. To address these limitations, recent methods inject 3D cues through large-scale supervision [9], augment features with depth and scene structure [13], or simulate viewpoint changes via abstract 3D proxies [31]. However, single-view observations provide limited spatial information and often suffer from occlusions. This motivates multi-view approaches where cross-view relations become essential.

Multi-view Spatial Reasoning Multi-view settings address single-view limitations by leveraging complementary viewpoints. Prior work transfers knowledge across views for improved QA [40, 69] and probes viewpoint robustness through relative direction, distance, and 6D pose [20, 33, 43]. Recent benchmarks evaluate multi-view understanding across diverse settings. AllAnglesBench [65] tests perspective-taking abilities. MindCube [66] assesses spatial reasoning from limited views. 3DSRBench [39] probes viewpoint robustness by varying camera poses. These benchmarks primarily focus on object properties within views or object-view grounding rather than cross-view rela-

tions. Large-scale datasets with explicit cross-view supervision remain limited. Recent works have made progress in multi-frame spatial reasoning: MultiSPA [62] provides large-scale training data for depth and visual correspondence, and MMSI-Bench [64] offers a human-curated evaluation benchmark for multi-image spatial intelligence. However, these works either lack explicit supervision on cross-view geometric relationships or do not cover both general and robotic domains. XVR addresses this gap by providing dense cross-view supervision across both domains, with an average of 4.32 images per sample.

Vision-Language-Action Models Recent VLA models map vision-language inputs directly to actions [5, 6, 28, 47, 53, 71]. To enhance spatial reasoning in VLA backbones, recent work injects robot-specific spatial signals [18, 55] and develops trajectory-grounded QA [10, 23, 51]. Methods like pi0.5 [22] demonstrate improved embodied reasoning through enhanced VLM backbones. XVR leverages robotic trajectories to construct datasets with explicit cross-view relation supervision. VLMs trained with XVR serve as improved backbones for VLA models, enhancing embodied manipulation performance.

3. Cross-View Relation Dataset

We introduce Cross-View Relation (XVR), a dataset for learning multi-view spatial reasoning through explicit cross-view relation supervision.

3.1. Task Categories

Multi-view spatial reasoning requires understanding how different viewpoints relate to each other geometrically. We organize XVR into the following three task categories:

- **Correspondence**: Identifying matching elements across views that represent the same physical entity. Tasks in this category teach models to link visual features across different viewpoints, forming the foundation for understanding shared scene geometry across views.
- **Verification**: Checking whether multi-view observations are geometrically or temporally consistent. Tasks in this category teach models to detect spatial or temporal inconsistencies, ensuring their understanding maintains coherence across views.
- **Localization**: Determining relative camera positions and which viewpoint corresponds to specific spatial conditions. This category captures how cameras relate to each other spatially and enables reasoning about relative viewpoints.

Together, these three categories provide structured supervision for learning cross-view relations, enabling robust multi-view spatial reasoning. We operationalize them through eight tasks. Figure 2 illustrates the three categories with representative examples.

Connection to Structure-from-Motion. Our categorization draws inspiration from Structure-from-Motion (SfM) [46, 50], a classical approach that integrates geometric information across multiple views to reconstruct 3D scenes. SfM operates through three key stages that directly inspired our categories: (i) identifying correspondences across views, (ii) verifying geometric consistency, and (iii) estimating camera poses. We adapt these stages into cross-view supervision for multi-view spatial reasoning.

3.2. Task Definitions

We instantiate the three categories through eight specific tasks.

Correspondence. *Point Correspondence* requires identifying which point across multiple views represents the same physical location in 3D space. This task evaluates whether models can match spatially aligned visual features under viewpoint changes. *Directional Correspondence* extends this to 3D orientation, requiring models to align directional arrows or vectors consistently across different camera projections. It tests reasoning about directional geometry beyond simple point matching.

Verification. *Spatial Verification* requires detecting correspondences that violate 3D spatial consistency among multiple views. By identifying geometrically inconsistent matches, this task measures the model's ability to enforce spatial coherence across perspectives. *Temporal Verification* requires identifying temporally inconsistent frames within a sequence. It assesses understanding of spatial-temporal structure by detecting frames that break temporal continuity.

Localization. *Viewpoint Localization* determines which camera view corresponds to a specific spatial position in the scene. This task evaluates whether models can infer relative viewpoint positions based on visual cues from multiple reference views. *Directional View Localization* identifies which camera view is located in a specific direction (e.g., left or right) relative to a reference camera. It evaluates directional awareness and relational reasoning between viewpoints. *Cross-Scenario Localization* requires matching corresponding viewpoints across structurally similar but distinct scenes. This task examines the generalization of viewpoint reasoning under scene-level variations. *Language-Conditioned Localization* selects the camera view that best matches a natural language spatial description. It integrates linguistic spatial cues (e.g., wrist-mounted camera) with geometric reasoning to identify corresponding visual perspectives.

Figure 2. **Overview of the question–answer (QA) structure in XVR.** The figure shows representative examples from eight task types across correspondence, verification, and localization categories, demonstrating the consistent QA format used throughout the dataset. Each category is color-coded: red for Correspondence (Point, Directional), green for Verification (Spatial, Temporal), and blue for Localization (Viewpoint, Directional View, Cross-Scenario, Language-Conditioned).

3.3. Data Generation Pipeline

To instantiate the eight tasks, we develop a unified generation framework (denoted as $\mathcal{G}$). This framework operationalizes our cross-view relation categories by structuring raw multi-view data to concrete question-answer (QA) pairs. As formalized in the supplementary material (Eq. 5), our framework is defined as $\mathcal{G} : (\mathcal{I}, \mathcal{P}, X, \mathcal{T}, \mathcal{M}) \rightarrow (\mathcal{Q}, \mathcal{A})$, where inputs comprise images ($\mathcal{I}$), camera parameters ($\mathcal{P}$),

3D geometry (X), temporal indices ($\mathcal{T}$), and metadata ($\mathcal{M}$).

The generation process differs based on data source characteristics. We describe two primary pipelines: the general domain pipeline, which leverages explicit 3D geometric information, and the robotic domain pipeline, which utilizes spatio-temporal metadata from robotic trajectories.

General domain. For tasks leveraging explicit 3D geometry (*Point Correspondence*, *Directional Correspondence*, *Spatial Verification*, and *Viewpoint Localization*), we employ a 3D-to-2D projection approach. We sample 3D primitives—points for correspondence tasks, camera positions for localization tasks—that are visible across multiple views. Using camera parameters from $\mathcal{P}$, we project these primitives onto available views and construct reference-target QA pairs. To create challenging questions, we generate spatially separated distractors for multiple-choice options, ensuring models must perform genuine cross-view reasoning rather than relying on low-level visual cues.

Robotic domain. For tasks utilizing robotic trajectories (*Temporal Verification*, *Directional View Localization*, *Cross-Scenario Localization*, and *Language-Conditioned Localization*), we sample from spatio-temporal metadata $\mathcal{M}$ and temporal indices $\mathcal{T}$. A critical quality control step ensures generated questions are perceptually meaningful: for Temporal Verification, we employ SSIM-based filtering [60] combined with action-based heuristics to verify that temporal differences produce visually distinguishable scene changes. This filtering prevents trivial questions where images are perceptually identical despite different timestamps.

All tasks follow a consistent reference-target QA structure where multiple reference views provide context and models must identify correct answers through cross-view reasoning. Complete task formalization is provided in Table 3. Further details on the generation pipeline are provided in Appendix 9 with an illustration in Figure 6.

3.4. Data Sources and Curation

We construct XVR using the following specific sources. These sources provide geometric richness from calibrated multi-view captures and realistic embodied dynamics from robotic trajectories, forming a balanced foundation for multi-view spatial reasoning.

General Domain. General domain data provides dense geometric supervision with accurate camera calibration, essential for geometry-based task generation. We adopt WildRGB-D [61] as our primary source, which contains multi-view RGB-D captures of diverse scenes with calibrated camera parameters. To ensure reliable geometric grounding and high-quality QA generation, we retain only samples with sufficiently dense point clouds (at least 1M points), guaranteeing robust 3D-to-2D projection and visibility analysis.

Robotic Domain. Robotic domain data provides temporal continuity and viewpoint diversity observed during manipulation tasks. We leverage OXE [47] and AgiBot-World [7] datasets as primary sources. Given the variable quality in raw robotic data, we apply strict filtering criteria to ensure task validity: (1) We include only sequences providing at least three distinct camera views to enable meaningful multi-view reasoning. Among publicly available datasets within the OXE suite, only DROID [26], MobileAloha [19], RoboSet [29], and FMB [38] satisfy this requirement. (2) We exclude sequences with inconsistent or ambiguous camera identifiers, as these compromise metadata-based localization task accuracy. (3) We retain only trajectories lasting at least 20 seconds with sufficient motion dynamics, measured by end-effector displacement, ensuring perceptually meaningful temporal variations for verification tasks. Further details on data sources and distribution are provided in Appendix 11.

4. Experiments

We conduct three complementary experiments to thoroughly evaluate the impact of XVR on multi-view spatial reasoning. First, we benchmark models on our proposed XVR-Eval suite (Sec. 4.1). Second, we evaluate models on external spatial benchmarks (Sec. 4.2). Finally, we examine embodied transfer by integrating XVR-trained backbones into a Vision-Language-Action (VLA) model (Sec. 4.3).

4.1. Benchmarking on XVR-Eval

Setup. To evaluate cross-view relation reasoning, we construct XVR-Eval, which consists of 1,866 held-out samples constructed from data sources unseen during XVR creation. Specifically, we include new sources: MobileAloha trajectories and WildRGB-D boat category scenes in XVR-Eval. We refer readers to Appendix 12 for statistics of XVR-Eval.

Using XVR-Eval, we test both open-source VLMs, such as Eagle2-2B [34], Paligemma-3B [4], InternVL-3.5-4B [58], and Qwen3-VL-Instruct (2B and 4B variants)[63], and closed models: Claude-4.5-Sonnet[2], GPT-5 [45], Gemini-2.5-Flash, Gemini-2.5-Pro [15], and Gemini-Robotics-ER-1.5 [1]. To verify the benefits of our XVR dataset, we fine-tune Qwen3-VL-Instruct (2B) on our XVR dataset and denote it as Qwen3-VL-2B-XVR. We also report a human baseline established from nine researchers with at least four years of higher education, collecting 795 annotations across all tasks.

Main results. Table 2 shows that most open-source models perform near chance level, while closed-source models achieve substantially higher performance yet still fall short

Model	Correspondence		Verification		Localization				Overall
	Point	Directional	Spatial	Temporal	Viewpoint	Directional View	Cross-scenario	Language-conditioned	
Closed-source Models									
Claude-4.5-Sonnet	68.94	24.24	52.65	51.76	23.65	**63.35**	71.95	57.01	51.18
GPT-5	**83.33**	**32.20**	**68.56**	**65.29**	**38.59**	60.63	**80.54**	**67.87**	**61.74**
Gemini-2.5-flash	78.03	31.44	60.61	56.47	14.52	57.47	66.06	56.11	52.36
Gemini-2.5-Pro	74.24	26.14	56.06	50.59	24.48	52.94	60.18	48.42	49.04
Gemini-Robotics-ER-1.5	76.89	22.35	50.00	51.76	6.22	53.85	66.06	56.11	47.48
Open-source Models									
Eagle2-2B	20.45	23.86	20.08	31.18	0.41	14.48	27.60	0.00	16.99
paligemma2-3b	2.65	4.55	23.11	35.29	6.64	30.77	11.76	33.48	17.36
InternVL-3.5-4B	34.09	25.00	24.62	49.41	4.15	52.04	37.10	41.18	32.32
Qwen3-VL-2B-Instruct	46.59	26.14	23.11	45.29	19.50	47.06	41.63	51.58	36.82
Qwen3-VL-4B-Instruct	57.95	29.55	48.11	**51.76**	10.37	53.39	60.63	52.94	45.02
Qwen3-VL-2B-XVR (Ours)	**94.32**	**53.79**	**84.85**	41.18	**57.68**	**68.33**	**70.14**	**63.35**	**68.06**
Baseline									
Random	20.00	25.00	22.22	33.33	33.33	50.00	33.33	50.00	32.64
Human	**92.31**	**67.11**	**88.46**	**77.08**	**64.94**	**92.08**	**87.74**	**93.48**	**83.85**

Table 2. Performance comparison on **XVR-Eval** (%). Results include closed-source models, open-source models (zero-shot and + XVR), and baselines.

of human baselines, indicating significant room for improvement. Our model, Qwen3-VL-2B-XVR, achieves a 1.8× improvement over its base model and ranks first among all evaluated models, surpassing both open-source and closed-source alternatives. Notably, Qwen3-VL-2B-XVR exceeds human performance on Point Correspondence, demonstrating that targeted supervision on cross-view relations can substantially improve spatial reasoning capabilities.

Task-specific patterns. Our analysis reveals two key findings. First, geometric reasoning tasks benefit substantially from XVR training. Point Correspondence and Spatial Verification show dramatic improvements, with Spatial Verification surpassing even GPT-5. Localization tasks demonstrate consistent gains, with Viewpoint Localization approaching human-level performance. These results validate that cross-view supervision enables models to perform geometric consistency checking, precise point matching, and camera-relative reasoning.

Second, Temporal Verification declines after XVR training, the only task showing this pattern. This reveals a trade-off: since most XVR tasks emphasize spatial reasoning at synchronized time points, training biases the model toward geometric structure at the expense of temporal sensitivity.

Closed-source model analysis. Despite their scale, closed-source models reveal task-specific limitations. GPT-5 exhibits large within-category variance: it excels at Point Correspondence but struggles with Directional Correspondence, despite both testing correspondence reasoning. Similarly, GPT-5 handles Spatial Verification well but fails at Viewpoint Localization.

Gemini-Robotics-ER-1.5 achieves the lowest accuracy among closed-source models. Its Viewpoint Localization accuracy (6.22%) falls below random guessing (22.22%), indicating minimal camera-relative reasoning capability. Even robotics-specialized training does not develop view-view relation reasoning without explicit supervision.

Gemini-2.5-Flash outperforms Gemini-2.5-Pro despite smaller scale. This shows that model capacity alone does not improve spatial reasoning. After XVR training, Qwen3-VL-2B surpasses all closed-source models, demonstrating that explicit supervision on view relations outweighs scale.

Human baseline comparison. XVR-trained models achieve super-human performance on Point Correspondence and Spatial Verification. However, gaps remain on Directional Correspondence and Temporal Verification, where human performance exceeds model performance by over 10 and 35 percentage points, respectively. Models excel at precise geometric calculations while humans handle ambiguous orientations and temporal dynamics better.

4.2. Evaluation on External Benchmarks

We test on two external benchmarks not used during XVR creation. MindCube-Tiny [66] evaluates scene imagination from limited viewpoints through three subtasks: *Around* (object identification under assumed camera motion), *Rotation* (spatial understanding from 360-degree viewpoints), and *Among* (object localization from alternative camera views). RoboSpatial-Home [55] evaluates spatial understanding for robotic manipulation through three subtasks, of which we

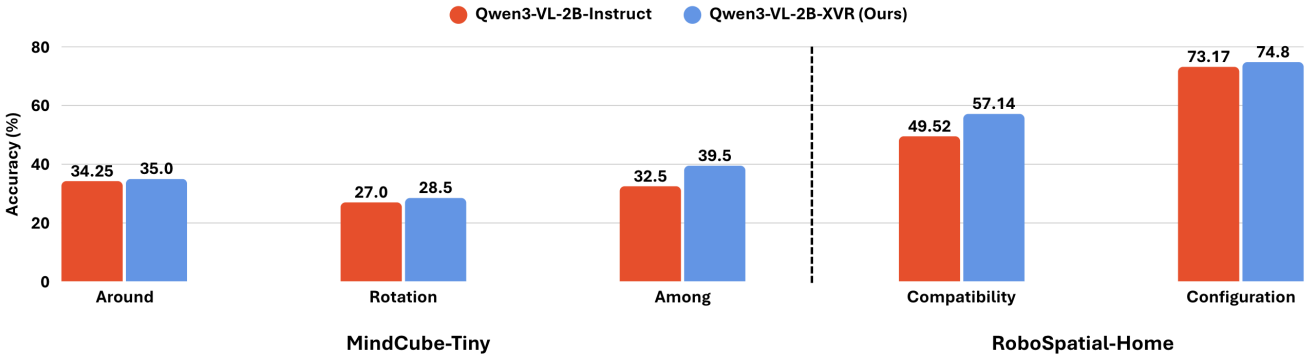

Figure 3. Generalization to external spatial benchmarks (MindCube-Tiny and RoboSpatial-Home). Training on XVR improves Qwen3-VL-2B across all tasks, with the largest gains in Compatibility (+7.6%) and Among (+7.0%).

evaluate two: *Compatibility* (spatial fit assessment) and *Configuration* (object-object spatial relations). We exclude the Context subtask as all evaluated models score 0. We compare baseline Qwen3-VL-2B against the XVR-trained variant.

Figure 3 shows that XVR training improves performance across subtasks in both benchmarks, though improvement magnitude varies systematically across tasks.

Transfer patterns. Tasks aligned with XVR's training distribution show substantial improvements. MindCube Among requires object localization from alternative camera views, directly matching XVR's multi-view training. RoboSpatial Compatibility and Configuration improve despite testing object-object spatial reasoning, suggesting that cross-view relation training builds 3D representations that transfer more broadly.

Tasks requiring camera motion understanding show minimal improvements. MindCube Around and Rotation involve continuous camera movement patterns absent from XVR's training distribution. XVR consists of 50% static multi-view scenes and 50% robotic trajectories that emphasize static camera configurations during manipulation. The limited transfer to motion-based tasks aligns with our temporal reasoning limitations on XVR-Eval.

Distribution shift. The improvements occur despite substantial distribution shifts. MindCube uses outside-looking-inward camera configurations, absent from XVR training data which focuses on inside-looking-outward setups. RoboSpatial evaluates single-view spatial reasoning while XVR trains on multi-view relations. These cross-domain improvements validate that cross-view relation reasoning captures general spatial principles rather than dataset-specific patterns.

Despite training exclusively on cross-view relation tasks, XVR-trained models show improvements on object-object spatial reasoning and partially on object-view reasoning

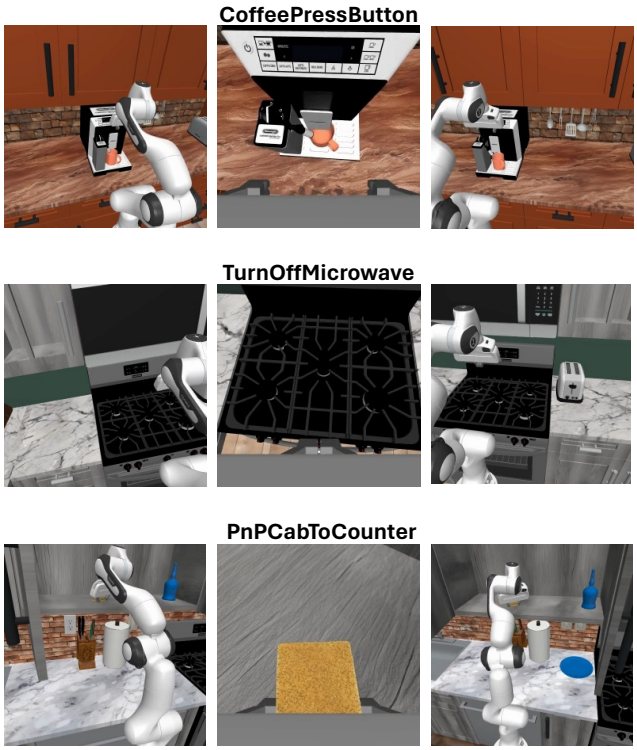

Figure 4. Visualization of the three manipulation tasks and their camera-view configurations used for VLA transfer evaluation.

across external benchmarks. This demonstrates that cross-view relation supervision provides a foundation for certain aspects of broader spatial reasoning, particularly those involving geometric relationships. Detailed task-by-task analysis is provided in Appendix 13.

4.3. Transfer to Vision-Language-Action Models

To investigate the benefits of XVR on embodied tasks, we extend VLMs trained on XVR into Vision-Language-Action (VLA) models. Specifically, we add a diffusion action head

to VLM representations following the architecture design of GR00T-N1.5 VLA [5]. Using the NVIDIA GR00T-X-Embodiment-Sim dataset from the RoboCasa simulator [44], we train VLAs to control a Franka Emika arm performing various manipulation tasks. We compare a VLA model based on Qwen3-VL-2B-Instruct against one based on our VLM, Qwen3-VL-2B-XVR, and report average success rates across 1,000 rollouts.

We evaluate three manipulation scenarios that require different forms of cross-view spatial reasoning. *CoffeePressButton* involves locating and pressing a small button that is visible only from the wrist camera due to occlusion, testing precise relative distance estimation under partial observability. *TurnOffMicrowave* presents the opposite visibility pattern—the control panel is clearly observed from the left and right cameras but occluded from the wrist view—requiring spatial disambiguation among multiple similar buttons across complementary viewpoints. *PnPCabToCounter* requires grasping one of 64 randomly selected object categories and placing it on the counter, testing generalizable multi-view pose estimation across diverse objects.

Figure 5 shows that our models consistently improve manipulation performance across all three tasks, with the largest gains on *TurnOffMicrowave*, where cross-view spatial disambiguation is most critical.

These improvements arise from the specific cross-view relation capabilities learned during XVR fine-tuning. Correspondence tasks teach point-level alignment across views, enabling view-consistent 3D representations that support accurate relative distance estimation. Localization tasks provide explicit camera-pose understanding, improving the integration of complementary viewpoints under partial observability. Verification tasks strengthen geometric consistency checking across views, supporting robust pose estimation for diverse object categories. The substantial gains on tasks requiring partial observability, spatial disambiguation, and cross-view generalization demonstrate that cross-view relation supervision enhances the geometric understanding necessary for downstream VLA manipulation.

5. Conclusion

We introduce XVR, a dataset for learning multi-view spatial reasoning from cross-view relations. Unlike existing multi-view datasets that emphasize objects within individual views, XVR provides explicit supervision on geometric relationships between views themselves. XVR comprises 100k samples from calibrated multi-view captures and robotic trajectories, organized into three reasoning categories: Correspondence, Verification, and Localization. We also introduce XVR-Eval, a 1,866-sample benchmark for systematic evaluation. Models trained on XVR demonstrate substantial improvements on XVR-Eval and consistent gains on external multi-view and robotic spatial benchmarks. When integrated into Vision-Language-Action models, XVR-trained backbones improve manipulation success rates on embodied tasks. These results demonstrate that explicit supervision on cross-view relations enhances multi-view spatial reasoning and transfers effectively to embodied manipulation.

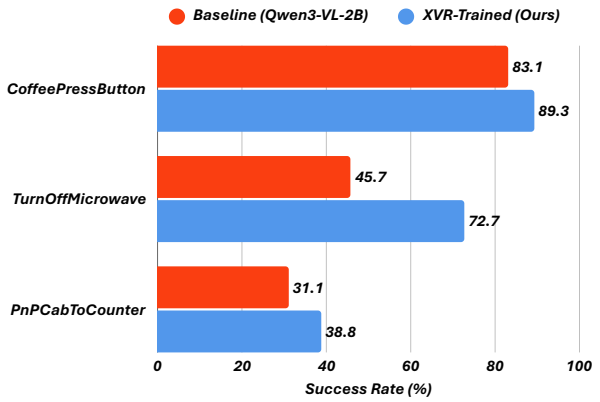

Figure 5. **Transfer to Embodied Tasks: RoboCasa VLA Performance.** Fine-tuning on XVR improves Qwen3-VL-2B performance on RoboCasa manipulation tasks, showing effective transfer of spatial reasoning skills to robotic action prediction.

This work enables more robust perception for robotic systems that rely on multi-camera setups. Beyond robotics, the approach has broader implications for applications requiring spatial understanding across multiple viewpoints, including autonomous navigation and AR/VR systems where maintaining geometric consistency is essential.

6. Limitation

Our work has two main limitations. First, we observe a limitation in temporal reasoning. Performance on Temporal Verification declines after XVR training, and models show minimal improvements on tasks involving dynamic camera movements. XVR emphasizes geometric consistency across static multi-view configurations, which reduces sensitivity to temporal dynamics. This trade-off improves structural stability across views at the cost of temporal flexibility. Future work could extend cross-view relation reasoning to explicitly incorporate temporal relationships, enabling models to understand both static spatial configurations and dynamic camera movements.

Second, our VLA transfer evaluation is conducted only in a simulation environment. While simulation provides controlled conditions for systematic analysis, it cannot fully capture the complexities of physical execution. Extending XVR-trained models to real robot platforms would offer a more comprehensive assessment of how cross-view relation reasoning transfers to real-world manipulation, and we view this as an important direction for future work.

Acknowledgments

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2019-II190075, Artificial Intelligence Graduate School Program(KAIST)); and by the Institute of Information & Communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (RS-2025-02304967, AI Star Fellowship(KAIST)). This research was also conducted as part of the Sovereign AI Foundation Model Project(Data Track), organized by the Ministry of Science and ICT(MSIT) and supported by the National Information Society Agency(NIA), S.Korea. (Grant No. 2026-AIData-WII01).

References

[1] Abbas Abdolmaleki, Saminda Abeyruwan, Joshua Ainslie, Jean-Baptiste Alayrac, Montserrat Gonzalez Arenas, Ashwin Balakrishna, Nathan Batchelor, Alex Bewley, Jeff Bingham, Michael Bloesch, et al. Gemini robotics 1.5: Pushing the frontier of generalist robots with advanced embodied reasoning, thinking, and motion transfer. *arXiv preprint arXiv:2510.03342*, 2025. 2, 5

[2] Anthropic. https://www.anthropic.com/news/claude-sonnet-4-5, 2025. 2, 5

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 1

[4] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2, 5

[5] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025. 3, 8

[6] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π0: A vision-language-action flow model for general robot control. *arXiv preprint ARXIV.2410.24164*, 2024. 3

[7] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025. 5, 6

[8] Wenxiao Cai, Iaroslav Ponomarenko, Jianhao Yuan, Xiaoqi Li, Wankou Yang, Hao Dong, and Bo Zhao. Spatialbot: Precise spatial understanding with vision language models. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9490–9498. IEEE, 2025. 1

[9] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1, 2

[10] Kaiyuan Chen, Shuangyu Xie, Zehan Ma, Pannag R Sanketi, and Ken Goldberg. Robo2vlm: Visual question answering from large-scale in-the-wild robot manipulation datasets. *arXiv preprint arXiv:2505.15517*, 2025. 3

[11] Lawrence Yunliang Chen, Simeon Adebola, and Ken Goldberg. Berkeley UR5 demonstration dataset. https://sites.google.com/view/berkeley-ur5/home. 1

[12] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. *arXiv preprint arXiv:2305.18565*, 2023. 1

[13] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37: 135062–135093, 2024. 1, 2

[14] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1

[15] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 5

[16] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. *arXiv preprint arXiv:1910.11215*, 2019. 1

[17] Mengfei Du, Binhao Wu, Zejun Li, Xuan-Jing Huang, and Zhongyu Wei. Embspatial-bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–355, 2024. 1, 2

[18] Zhiyuan Feng, Zhaolu Kang, Qijie Wang, Zhiying Du, Jiongrui Yan, Shubin Shi, Chengbo Yuan, Huizhi Liang, Yu Deng, Qixiu Li, et al. Seeing across views: Benchmarking spatial reasoning of vision-language models in robotic scenes. *arXiv preprint arXiv:2510.19400*, 2025. 1, 3

[19] Zipeng Fu, Tony Z Zhao, and Chelsea Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024. 1, 5, 6, 7

[20] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2023. 2

[21] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 2

[22] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. pi0.5: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025. 3

[23] Yuheng Ji, Huajie Tan, Jiayu Shi, Xiaoshuai Hao, Yuan Zhang, Hengyuan Zhang, Pengwei Wang, Mengdi Zhao, Yao Mu, Pengju An, et al. Robobrain: A unified brain model for robotic manipulation from abstract to concrete. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1724–1734, 2025. 3

[24] Mengdi Jia, Zekun Qi, Shaochen Zhang, Wenyao Zhang, Xinqiang Yu, Jiawei He, He Wang, and Li Yi. Omnispatial: Towards comprehensive spatial reasoning benchmark for vision language models. *arXiv preprint arXiv:2506.03135*, 2025. 1, 2

[25] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[26] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024. 1, 5, 6

[27] Geewook Kim, Teakgyu Hong, Moonbin Yim, JeongYeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022. 1

[28] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024. 3

[29] Vikash Kumar, Rutav Shah, Gaoyue Zhou, Vincent Moens, Vittorio Caggiano, Abhishek Gupta, and Aravind Rajeswaran. Robohive: A unified framework for robot learning. *Advances in Neural Information Processing Systems*, 36:44323–44340, 2023. 1, 5, 6

[30] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 1

[31] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. *arXiv preprint arXiv:2504.17207*, 2025. 2

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[33] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[34] Zhiqi Li, Guo Chen, Shilong Liu, Shihao Wang, Vibashan VS, Yishen Ji, Shiyi Lan, Hao Zhang, Yilin Zhao, Subhashree Radhakrishnan, et al. Eagle 2: Building post-training data strategies from scratch for frontier vision-language models. *arXiv preprint arXiv:2501.14818*, 2025. 2, 5

[35] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 1, 2

[36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llavanext: Improved reasoning, ocr, and world knowledge, 2024. 1

[37] Jianlan Luo, Charles Xu, Xinyang Geng, Gilbert Feng, Kuan Fang, Liam Tan, Stefan Schaal, and Sergey Levine. Multi-stage cable routing through hierarchical imitation learning. *URL https://arxiv. org/abs/2307.08927*, 22, 2023. 1

[38] Jianlan Luo, Charles Xu, Fangchen Liu, Liam Tan, Zipeng Lin, Jeffrey Wu, Pieter Abbeel, and Sergey Levine. Fmb: a functional manipulation benchmark for generalizable robotic learning. *The International Journal of Robotics Research*, 44 (4):592–606, 2025. 1, 5, 6

[39] Wufei Ma, Haoyu Chen, Guofeng Zhang, Yu-Cheng Chou, Jieneng Chen, Celso de Melo, and Alan Yuille. 3dsrbench: A comprehensive 3d spatial reasoning benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6924–6934, 2025. 1, 2

[40] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. *arXiv preprint arXiv:2210.07474*, 2022. 2

[41] Tatsuya Matsushima, Hiroki Furuta, Yusuke Iwasawa, and Yutaka Matsuo. Weblab xarm dataset, 2023. 1

[42] Peter Mitrano and Dmitry Berenson. Conq hose manipulation dataset, v1.15.0. https://sites.google.com/view/conq-hose-manipulation-dataset, 2024. 1

[43] Wentao Mo, Qingchao Chen, Yuxin Peng, Siyuan Huang, and Yang Liu. Advancing 3d scene understanding with mv-scanqa multi-view reasoning evaluation and tripalign pre-training dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12973–12980, 2025. 2

[44] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024. 2, 8

[45] OpenAI. https://openai.com/gpt-5/, 2025. 2, 5

[46] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion*. *Acta Numerica*, 26:305–364, 2017. 2, 3

[47] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024. 3, 5

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[49] Amrita Sawhney, Steven Lee, Kevin Zhang, Manuela Veloso, and Oliver Kroemer. Playing with food: Learning food item representations through interactive exploration. In *International Symposium on Experimental Robotics*, pages 309–322. Springer, 2020. 1

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 3

[51] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 645–652. IEEE, 2024. 3

[52] Haochen Shi, Huazhe Xu, Samuel Clarke, Yunzhu Li, and Jiajun Wu. Robocook: Long-horizon elasto-plastic object manipulation with diverse tools. *arXiv preprint arXiv:2306.14447*, 2023. 1

[53] Lucy Xiaoyang Shi, Brian Ichter, Michael Equi, Liyiming Ke, Karl Pertsch, Quan Vuong, James Tanner, Anna Walling, Haohuan Wang, Niccolo Fusai, et al. Hi robot: Open-ended instruction following with hierarchical vision-language-action models. *arXiv preprint arXiv:2502.19417*, 2025. 3

[54] Fatemeh Shiri, Xiao-Yu Guo, Mona Far, Xin Yu, Reza Haf, and Yuan-Fang Li. An empirical analysis on spatial reasoning capabilities of large multimodal models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21440–21455, 2024. 1, 2

[55] Chan Hee Song, Valts Blukis, Jonathan Tremblay, Stephen Tyree, Yu Su, and Stan Birchfield. Robospatial: Teaching spatial understanding to 2d and 3d vision-language models for robotics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15768–15780, 2025. 1, 2, 3, 6

[56] Chen Wang, Linxi Fan, Jiankai Sun, Ruohan Zhang, Li Fei-Fei, Danfei Xu, Yuke Zhu, and Anima Anandkumar. Mimicplay: Long-horizon imitation learning by watching human play. *arXiv preprint arXiv:2302.12422*, 2023. 1

[57] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19175–19186, 2023. 1

[58] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 5

[59] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European Conference on Computer Vision*, pages 396–416. Springer, 2024. 1

[60] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5, 4

[61] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22378–22389, 2024. 5, 6, 7

[62] Runsen Xu, Weiyao Wang, Hao Tang, Xingyu Chen, Xiaodong Wang, Fu-Jen Chu, Dahua Lin, Matt Feiszli, and Kevin J Liang. Multi-spatialmllm: Multi-frame spatial understanding with multi-modal large language models. *arXiv preprint arXiv:2505.17015*, 2025. 3

[63] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2, 5

[64] Sihan Yang, Runsen Xu, Yiman Xie, Sizhe Yang, Mo Li, Jingli Lin, Chenming Zhu, Xiaochen Chen, Haodong Duan, Xiangyu Yue, et al. Mmsi-bench: A benchmark for multi-image spatial intelligence. *arXiv preprint arXiv:2505.23764*, 2025. 3

[65] Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. *arXiv preprint arXiv:2504.15280*, 2025. 1, 2

[66] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, et al. Spatial mental modeling from limited views. In *Structural Priors for Vision Workshop at ICCV'25*, 2025. 1, 2, 6

[67] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in neural information processing systems*, 34:23634–23651, 2021. 1

[68] Jirong Zha, Yuxuan Fan, Xiao Yang, Chen Gao, and Xinlei Chen. How to enable llm with 3d capacity? a survey of spatial reasoning in llm. *arXiv preprint arXiv:2504.05786*, 2025. 1

[69] Weichen Zhang, Zile Zhou, Zhiheng Zheng, Chen Gao, Jinqiang Cui, Yong Li, Xinlei Chen, and Xiao-Ping Zhang.

Open3dvqa: A benchmark for comprehensive spatial reasoning with multimodal large language model in open space. *arXiv preprint arXiv:2503.11094*, 2025. 2

[70] Zheyuan Zhang, Fengyuan Hu, Jayjun Lee, Freda Shi, Parisa Kordjamshidi, Joyce Chai, and Ziqiao Ma. Do vision-language models represent space and how? evaluating spatial frame of reference under ambiguities. *arXiv preprint arXiv:2410.17385*, 2024. 1

[71] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pages 2165–2183. PMLR, 2023. 3