Learning to Reason across Viewpoints: Multi-Viewpoint Spatial Reasoning for Multimodal Large Language Models

Xiaojian Huang[1], Jie Zhao[2*], Xin Liu[2], Jin Xu[1], Zhihong Zhang[1], Zhuodong Luo[1], Xuejin Chen[1*]
[1]University of Science and Technology of China, China
[2]Huawei Technologies Co., Ltd., China
[1]{xiaojianhuang, xj01spirit, zhangzhihong, coolzd, xjchen99}@mail.ustc.edu.cn
[2]{zhaojie104, liuxin572}@huawei.com

***Abstract*—Multimodal large language models (MLLMs) have achieved remarkable progress in visual understanding and reasoning, yet they still struggle with spatial reasoning—especially when it requires reasoning across viewpoints. To address this limitation, we propose a scalable data generation pipeline that combines a text-to-image generative model with a scene generator. The pipeline enables annotation-free construction of high-quality multi-viewpoint spatial reasoning data, spanning diverse task formulations and viewpoint types. Using this pipeline, we build MVSR-Dataset, a dataset of 205K question–answer (QA) pairs covering both single-image and multi-image settings. We further introduce MVSR-Bench, a manually verified benchmark that provides a more comprehensive evaluation of multi-viewpoint spatial reasoning than prior benchmarks. Fine-tuning on MVSR-Dataset leads to strong performance across multiple multi-image spatial reasoning benchmarks.**

***Index Terms*—multimodal large language models, multi-viewpoint spatial reasoning**

I. Introduction

Multimodal large language models (MLLMs) have achieved remarkable progress in visual understanding and reasoning tasks such as visual question answering (VQA), grounding, and image captioning [1]–[4]. Despite these advances, spatial reasoning remains challenging, particularly when it requires reasoning across viewpoints. While recent MLLMs have shown promising progress in egocentric spatial understanding, i.e., reasoning about object relations from the capturing camera's perspective, they still struggle to reason from alternative viewpoints in real-world 3D scenes. In particular, multi-viewpoint spatial reasoning requires models to adopt the perspective of an oriented entity, such as a person or an animal, and infer spatial relations (e.g., left, right, front, and back) among surrounding objects relative to that entity's facing direction rather than the camera view. This limitation substantially hinders the practical deployment of MLLMs in robotics, AR/VR, and autonomous driving.

Recent work has introduced a variety of datasets and benchmarks to improve and assess spatial reasoning in multi-modal models [5]–[13]. However, most existing efforts remain largely camera-centric and provide limited coverage of spatial reasoning under diverse viewpoints [5], [6], [9], [10]. Although several studies have begun to explore multi-viewpoint spatial reasoning [11]–[15], important gaps persist. First, some works study multi-viewpoint reasoning only in single-image settings [12], [14], without extending to multi-image inputs that require cross-view integration. Second, benchmarks such as SPAR [11] and ViewSpatial-Bench [13] consider multi-viewpoint reasoning in multi-image settings, yet they overlook richer viewpoint types in single-image scenarios (e.g., animal- and vehicle-centric views) and cover a limited range of challenging multi-image task formulations. Third, high-quality training data for multi-viewpoint spatial reasoning remains scarce. While SpatialReasoner [12] and ViewSpatial-Bench [13] both propose data generation pipelines, they exhibit limitations. SpatialReasoner constructs its data by synthesizing question–answer pairs based on estimations from multiple vision foundation models, followed by manual verification. However, this pipeline is prone to error accumulation across models. Furthermore, since the underlying orientation estimation models are trained exclusively on data with clean backgrounds and complete objects, they struggle to generalize to scenarios involving occlusions. In contrast, ViewSpatial-Bench [13] constructs high-quality samples based on ScanNet [16], but its reliance on 3D ground-truth annotations significantly limits scalability and practical extensibility. Together, these limitations motivate the need for a scalable, high-quality data generation pipeline and a more comprehensive benchmark to advance multi-viewpoint spatial reasoning in MLLMs.

In this work, we introduce a unified framework for multi-viewpoint spatial reasoning. Our framework defines eight task types spanning both single-image and multi-image settings and covers a diverse set of viewpoints. In particular, the multi-image tasks explicitly require joint reasoning over cross-view object association and camera localization. Building on this task taxonomy, we develop a unified data generation pipeline with two complementary branches: (1) a Single Image Branch that leverages text-to-image models to synthesize single-image

* Corresponding authors: Jie Zhao (zhaojie104@huawei.com) and Xuejin Chen (xjchen99@ustc.edu.cn).

data for viewpoint-conditioned spatial reasoning, and (2) a Multi-view Scene Branch that uses a scene generator to produce multi-image data that enables joint reasoning over cross-view object association and camera localization. Building on this pipeline, we construct MVSR-Dataset, a dataset of 205K QA pairs spanning both single-image and multi-image settings, and introduce MVSR-Bench, a high-quality benchmark of 2,650 QA pairs that enables a more comprehensive evaluation of multi-viewpoint spatial reasoning than existing benchmarks. All QA pairs in our MVSR-Bench are manually verified to ensure their reliability and annotation quality across all tasks. Fine-tuning Qwen3-VL-8B-Instruct [17] on MVSR-Dataset yields strong gains on MVSR-Bench and improves performance on multiple existing multi-view spatial reasoning benchmarks. Together, our dataset, benchmark, and scalable pipeline provide a solid foundation for advancing multi-viewpoint spatial reasoning in MLLMs.

Our contributions can be summarized as follows:

- We propose a unified pipeline for generating both single-image and multi-image VQA data for viewpoint-conditioned multi-view spatial reasoning.
- Based on this pipeline, we construct **MVSR-Dataset**, a dataset of 205K QA pairs covering both single-image and multi-image settings, and introduce **MVSR-Bench**, a comprehensive benchmark for evaluating multi-view spatial reasoning in existing models.
- By fine-tuning Qwen3-VL-8B-Instruct [17] on **MVSR-Dataset**, we achieve strong performance across multiple benchmarks for multi-view spatial reasoning.

The code and benchmark are available at https://github.com/hxiaoj/MVSR.

II. Related Work

Datasets for Spatial Reasoning. Existing MLLMs remain limited in 3D spatial reasoning, largely due to the scarcity of high-quality and diverse training data. Prior works either generate single-image spatial reasoning QA pairs from pseudo-annotations produced by vision foundation models [5], [6], [8], [12], [18], or leverage scene datasets with 3D ground-truth annotations to support multi-image reasoning [9]–[11], [19], [20]. However, most existing datasets focus on camera-centric reasoning and offer limited coverage of multi-viewpoint spatial reasoning. Recent efforts toward automated multi-viewpoint VQA data construction [12], [13] still rely on either vision-model estimates or 3D annotations, leading to trade-offs between quality and scalability. Our work addresses this gap with a scalable pipeline for generating high-quality multi-viewpoint spatial reasoning VQA data in both single-image and multi-image settings.

Benchmarks for Spatial Reasoning. Early benchmarks mainly evaluate egocentric spatial reasoning [6], [9], [10], [21]–[24]. Subsequent works extend evaluation to multi-viewpoint reasoning in single-image, video, and multi-image settings [11], [13]–[15], [25]. In contrast, our MVSR-Bench is specifically designed for multi-viewpoint spatial reasoning, with broader coverage of viewpoint types and task formulations for more comprehensive evaluation of MLLMs.

III. Our Method

This section presents our unified framework for multi-viewpoint spatial reasoning. We first define eight task types spanning single-image and multi-image settings, and then introduce two complementary data generation branches for these settings. Finally, we describe the construction of MVSR-Bench and the fine-tuning of MLLMs on MVSR-Dataset.

A. Multi-Viewpoint Spatial Reasoning Tasks

To enhance and evaluate multi-viewpoint spatial reasoning capabilities of various MLLM models, we design a comprehensive suite of eight task types, as summarized in Table I, spanning two evaluation settings. In the single-image setting, four tasks are introduced to assess a model's ability of spatial reasoning from animal-, person-, vehicle-, and camera-centric viewpoints. In the multi-image setting, we introduce four additional tasks that require not only viewpoint transformations but also explicit reasoning over camera positions and object relationships across views.

Specifically, the multi-image tasks are categorized into person-centric and camera-centric viewpoints. For the person-centric viewpoint, the model is required to reason from the perspective of a person located at one object and oriented toward another object. Unlike prior settings, we enforce that the two objects must appear in different images, thereby explicitly requiring cross-view reasoning. Moreover, we incorporate camera position reasoning by requiring the model to infer spatial relationships involving camera locations, such as reasoning about the position of one camera from either a person-centric or camera-centric viewpoint. Together, these designs impose stricter cross-view constraints and more complex viewpoint reasoning, resulting in a more challenging and realistic evaluation of multi-viewpoint spatial reasoning capabilities.

B. Single Image Branch

The single-image generation branch generates multi-viewpoint spatial reasoning QA data under the single-image setting, including text-to-image (T2I) prompt generation and spatial QA generation as illustrated above in Figure 1.

1) T2I Prompt Generation: First, we filter object categories from the Object365 [26] dataset using GPT-5 [27], removing those that do not exhibit a well-defined orientation, and generate descriptive texts for each remaining category. Next, we randomly sample a task from the single-image setting listed in Table I. Given the selected task, we sample a spatial configuration that specifies both relative object positions and object orientations. Specifically, relative positions are drawn from four canonical relations (left, right, front, and behind), while object orientations are selected from four discrete states: facing the camera, facing away from the camera, facing left, and facing right. Finally, we randomly select object descriptions from the generated pool and, conditioned on the

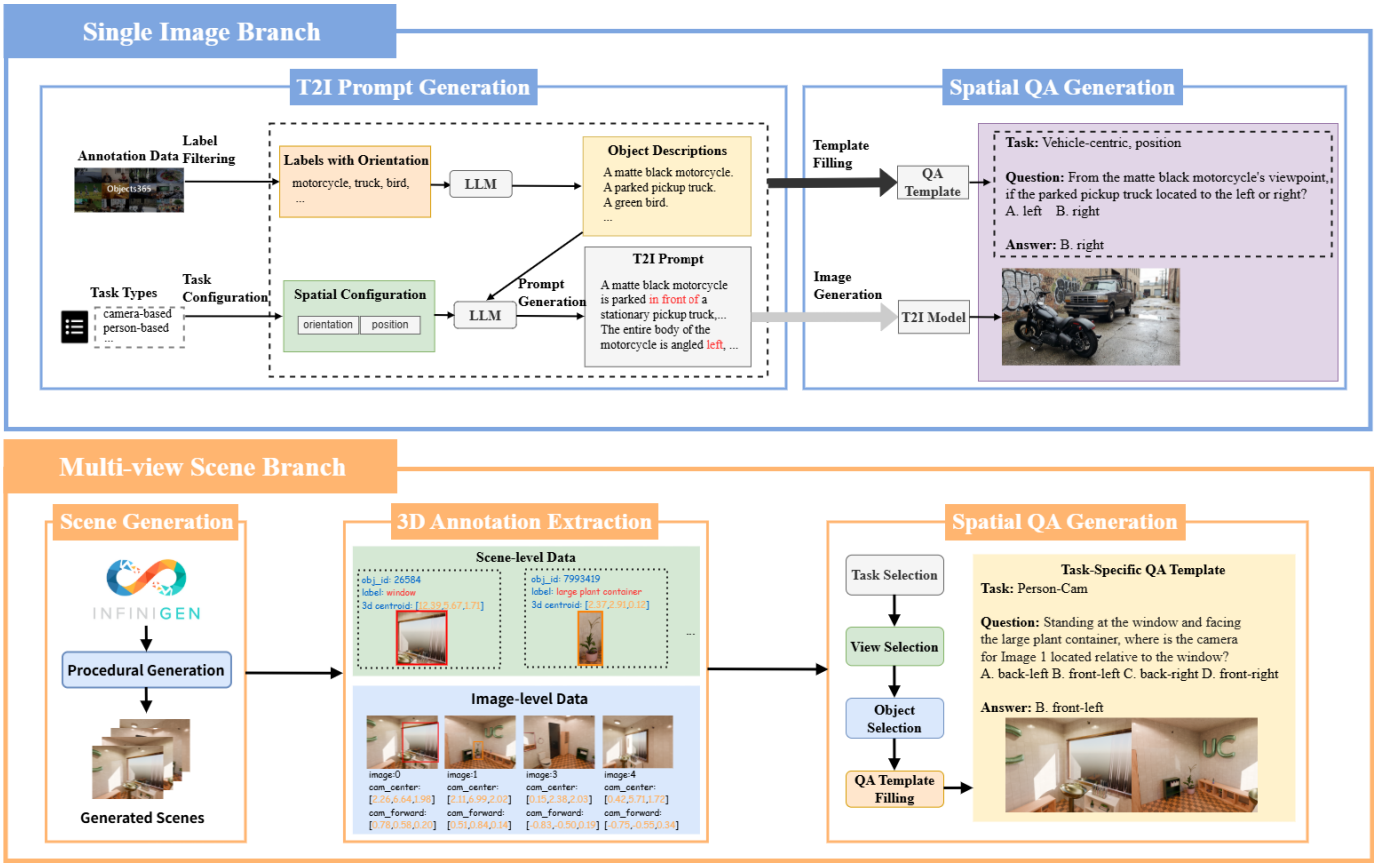

Fig. 1. **Our data generation pipeline.** The single image branch synthesizes single-image QA by sampling oriented objects and spatial configurations for text-to-image generation, while the multi-view scene branch generates multi-image QA by constructing 3D scenes with Infinigen and leveraging extracted scene annotations.

sampled spatial configuration, use GPT-5 [27] to generate the corresponding T2I prompts for subsequent image generation.

2) Spatial QA Generation: Based on the selected spatial configurations and object descriptions, we generate QA pairs through template-based instantiation. The templates are initially designed by human annotators, then expanded using GPT-5 [27] and manually curated to ensure high quality. In parallel, given the generated T2I prompts, we generate images that conform to the specified spatial configurations using Gemini-3-Pro-Image [28]. Together, this process yields the single-image subset of the final spatial QA dataset.

C. Multi-view Scene Branch

This section describes the generation of multi-viewpoint spatial reasoning data in multi-image setting, covering scene generation, 3D annotation extraction, and spatial QA generation as shown below in Figure 1.

1) Scene Generation: We generate 3D textured models of indoor scenes using Infinigen [29], a fully procedural framework that generates photorealistic 3D scenes by generating all assets from scratch via randomized mathematical rules, enabling unlimited variation without relying on external data. In total, we construct 439 indoor scenes spanning five types: bathroom, living room, kitchen, dining room, and bedroom. Each scene contains an average of 65 objects, resulting in complex and cluttered layouts. At the object level, we generate 15,592 unique object instances covering 95 object categories. The resulting diversity in scene layouts and object categories supports challenging spatial reasoning tasks.

To further obtain multiple images for spatial QA generation, we randomly sample camera viewpoints within the generated scenes and apply a filtering procedure to the resulting candidates. Specifically, we ensure that the overlap between viewpoints is neither excessively high nor excessively low, that cameras are not placed too close to objects, and that the selected viewpoints collectively provide sufficient coverage of each scene. Using this strategy, we generate an average of 10 frames per scene at a resolution of 640×480, which are used for subsequent QA generation.

2) 3D Annotation Extraction: At this stage, we extract 3D annotations from each scene and organize them into a unified data structure to facilitate subsequent multi-viewpoint spatial QA generation. This data structure consists of two components: scene-level data and image-level data. The scene-level data includes the 3D coordinates, object categories, and unique object identifiers for all objects in the scene. The image-level data records the image resolution, the 3D position and orientation of the capturing camera, as well as the visible

TABLE I
OVERVIEW OF MULTI-VIEWPOINT SPATIAL REASONING TASKS.

Setting	Tasks	Descriptions
Single-image	Animal-centric	Estimate an object's position and orientation from an animal-centric viewpoint.
	Person-centric	Estimate an object's position and orientation from a person-centric viewpoint.
	Vehicle-centric	Estimate an object's position and orientation from a vehicle-centric viewpoint.
	Camera-centric	Estimate an object's orientation from the camera viewpoint.
Multi-image	Cam–Cam	Estimate one camera's position in another camera's viewpoint; requires camera localization.
	Cam–Obj	Estimate object positions in one view from another camera's viewpoint; requires cross-view association.
	Person–Cam	Estimate another camera's position from a person-centric viewpoint; requires camera localization.
	Person–Obj	Estimate object positions in one view from a person-centric viewpoint in another; requires cross-view association.

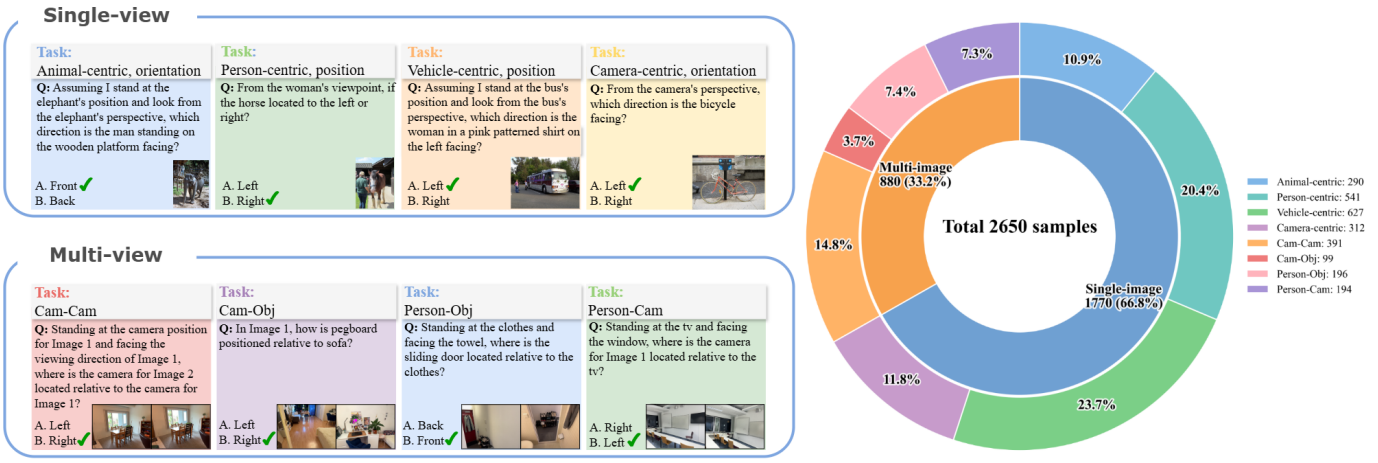

Fig. 2. Visualization (left) and statistical overview (right) of our MVSR-Bench.

objects in each image along with their corresponding 2D coordinates. This unified representation enables efficient and systematic construction of spatial QA pairs involving cameras and objects across different views.

3) Spatial QA Generation: Based on the extracted scene-level and image-level data, we procedurally generate the QA dataset. Specifically, we first perform task selection by randomly sampling a task type from the multi-image settings listed in Table I. Given the selected task type, we then select images that satisfy the corresponding task constraints. For instance, in the Person–Cam task, we choose image pairs with a sufficient degree of overlap while ensuring that at least two objects appear exclusively in each image. Next, objects are randomly sampled from the selected images, and QA pairs are generated through template-based instantiation. The QA templates are initially designed by human annotators according to each task type, and are subsequently expanded using GPT-5 [27] and manually filtered, thereby ensuring both diversity and quality of the generated QA data.

D. Construction of MVSR-Bench

Our MVSR-Bench is designed to evaluate multi-viewpoint spatial reasoning capabilities of MLLM models. It consists of eight tasks, as detailed in Table I, covering both single-image and multi-image settings. For the single-image setting, we manually select images that meet our annotation criteria from the validation and test sets of MS-COCO [35], as well as the test set of OpenImages [36]. Based on these images, we generate QA pairs for different task types using GPT-5 [27], followed by manual correction, resulting in a high-quality set of 1,770 QA pairs. For the multi-image setting, we extract 3D annotations from the ScanNet++ [37] validation set and select appropriate multi-image groups according to different task types. QA pairs are then generated via template-based instantiation and subsequently verified through manual inspection, yielding 880 high-quality QA pairs.

Ultimately, we construct MVSR-Bench, which comprises a total of 2,650 high-quality QA pairs spanning eight task types. Figure 2 illustrates the distribution of QA pairs across different tasks and representative examples from our MVSR-Bench.

E. Finetuning MLLM with MVSR-Dataset

Based on the proposed data generation pipeline, we generate 85k samples for single-image multi-viewpoint spatial reasoning and 120k samples for multi-image multi-viewpoint spatial reasoning, forming the MVSR-Dataset. We further fine-tune

TABLE II
ACCURACY OF DIFFERENT MODELS ON MVSR-BENCH (%). BEST RESULTS ARE IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED.

Model	Single-view Tasks				Multi-view Tasks				Overall
	Animal	Camera	Person	Vehicle	Cam–Cam	Cam–Obj	Person–Cam	Person–Obj	
Proprietary Models									
GPT-5.2 [27]	53.30	57.05	**50.83**	51.83	33.75	54.54	62.37	55.61	50.89
GPT-4o [30]	44.15	39.87	41.03	45.77	37.08	52.52	28.86	51.28	42.07
Gemini-2.5 Pro [31]	<u>54.48</u>	62.17	45.65	**59.33**	29.41	48.48	51.55	69.90	<u>51.73</u>
Open-source Models									
InternVL3-78B [32]	52.41	57.05	39.27	47.84	29.15	52.52	47.42	33.67	44.01
InternVL3-38B [32]	43.10	51.60	42.60	46.57	31.96	49.49	44.84	33.67	42.84
Qwen3-VL-32B-Instruct [17]	51.72	<u>64.42</u>	42.22	54.70	20.46	56.56	61.85	<u>79.08</u>	50.31
Qwen2.5-VL-72B-Instruct [33]	51.38	63.14	41.67	50.88	31.20	39.39	<u>65.98</u>	51.53	48.32
Qwen2.5-VL-32B-Instruct [33]	51.72	64.10	40.75	48.32	29.16	51.52	45.36	35.71	45.15
Qwen3-VL-8B-Instruct [17]	50.65	59.61	42.97	48.16	33.24	<u>59.59</u>	54.12	46.42	47.25
Qwen2.5-VL-7B-Instruct [33]	51.03	60.57	40.56	47.04	<u>44.24</u>	42.42	45.36	27.55	45.60
InternVL3-8B [32]	44.13	46.79	40.01	43.54	39.13	46.46	31.44	30.61	40.88
InternVL2.5-8B [34]	35.86	45.51	38.35	42.74	37.34	35.35	32.99	30.61	38.73
Qwen3-VL-4B-Instruct [17]	50.34	54.80	37.60	47.05	35.29	54.54	51.54	42.85	44.96
Qwen3-VL-8B-Instruct [17]+MVSR-Dataset	**61.61**	**83.01**	<u>47.60</u>	<u>56.87</u>	**52.42**	**82.82**	**87.11**	**90.81**	**63.61**

models on MVSR-Dataset to validate the effectiveness of the constructed dataset.

IV. EXPERIMENTS

In this section, we first describe the experimental setup. We then evaluate the effectiveness of MVSR-Dataset by fine-tuning models and comparing their performance with other leading models across multiple benchmarks for multi-viewpoint spatial reasoning. Finally, we provide a detailed analysis of data scaling effects and each component of the data generation pipeline.

A. Experimental settings

Baselines. To assess the effectiveness of MVSR-Dataset, we fine-tune Qwen3-VL-8B-Instruct [17] and benchmark it against representative closed-source models (GPT-5.2 [27], GPT-4o [30], and Gemini-2.5-Pro [31]), as well as strong open-source baselines, including Qwen2.5-VL [33], Qwen3-VL [17], InternVL-3 [32], and InternVL-2.5 [34].

Implementation details. MVSR-Dataset contains 205k samples, including 85k single-image and 120k multi-image samples. We fine-tune Qwen3-VL-8B-Instruct [17] for one epoch with a batch size of 128 and a learning rate of 1×10^{-5} on eight Ascend 910B NPUs.

Evaluation. We evaluate models on our MVSR-Bench, 3DSR-Bench [14], and ViewSpatial-Bench [13], and report the average accuracy across tasks.

B. Main Results

Fine-tuning models on our MVSR-Dataset consistently improves their performance across MVSR-Bench, 3DSRBench, and ViewSpatial-Bench. As shown in Table II, Qwen3-VL-8B-Instruct [17] fine-tuned on MVSR achieves 63.61% average accuracy on MVSR-Bench, outperforming all evaluated proprietary and open-source models by 11.88% over the second-best Gemini-2.5-Pro [31]. It also delivers the best performance in most task categories. On 3DSRBench [14], the fine-tuned model improves orientation accuracy by 13.53%, ranking second overall and narrowing the gap to GPT-5 [27] to 4.25% (Table III). On ViewSpatial-Bench [13], using Qwen2.5-VL-7B-Instruct [33] as the baseline, fine-tuning on MVSR-Dataset yields a 9.15% gain and achieves the best overall performance (Table IV). These results demonstrate the effectiveness and generalization of MVSR-Dataset across diverse multi-viewpoint spatial reasoning benchmarks.

TABLE III
PERFORMANCE OF DIFFERENT MODELS ON THE ORIENTATION TASK OF 3DSRBENCH (%). BEST RESULTS ARE IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED.

Model	Orientation
GPT-5 [27]	**62.73**
Gemini2.5-Pro [31]	53.90
Intern2.5VL-8B [34]	47.57
Qwen2.5-VL-7B-Instruct [33]	46.53
InternVL3-8B [32]	47.10
Qwen2.5-VL-3B-Instruct [33]	42.90
Qwen3-VL-8B-Instruct [17]	44.95
Qwen3-VL-8B-Instruct [17]+MVSR-Dataset	<u>58.48</u>

TABLE IV
PERFORMANCE OF DIFFERENT MODELS ON VIEWSPATIAL-BENCH (%). BEST RESULTS ARE IN BOLD AND SECOND-BEST RESULTS ARE UNDERLINED.

Model	Camera	Person	Overall
GPT-4o [30]	33.57	36.29	34.98
InternVL3-14B [32]	**47.09**	33.88	40.28
InternVL2.5-8B [34]	<u>46.48</u>	<u>40.20</u>	<u>43.24</u>
Qwen2.5-VL-7B [33]	40.56	33.37	36.85
Qwen2.5-VL-7B [33]+MVSR-Dataset	42.25	**49.75**	**46.00**

TABLE V
COMPARISON ON THE ORIENTATION TASK OF 3DSRBENCH (%). THE RESULTS SHOW THAT THE SINGLE-IMAGE DATA GENERATED BY OUR PIPELINE OUTPERFORMS THE MANUALLY VERIFIED REAL-WORLD DATASET.

Method	Orientation
Qwen3-VL-8B-Instruct [17]	44.38
+ Data from SpatialReasoner [12]	46.09
+ MVSR-Dataset(SV)	**48.00**

TABLE VI
COMPARISON ON VIEWSPATIAL-BENCH (%). THE RESULTS INDICATE THAT THE MULTI-IMAGE DATA GENERATED BY OUR PIPELINE ACHIEVES PERFORMANCE COMPARABLE TO THOSE USING MANUALLY ANNOTATED REAL-WORLD DATA.

Method	Cam-Cam	Cam-Obj	Person-Cam	Person-Obj	Overall
Qwen3-VL-8B-Instruct [17]	33.24	59.59	54.12	46.42	43.74
+ ScanNet [16]	**68.79**	**92.92**	76.28	**90.30**	**77.95**
+ MVSR-Dataset (MV)	52.69	87.88	**86.60**	88.27	72.05

C. Ablation Study

Comparison with Real-World Data. We further compare our generated data with real-world data by separately evaluating the Single Image and Multi-view Scene branches. For the single-image branch, we compare with SpatialReasoner [12] and randomly sample 24k examples from MVSR-Dataset for a fair comparison. As shown in Table V, fine-tuning Qwen3-VL-8B-Instruct [17] on our data yields a 1.91% gain over SpatialReasoner, validating the effectiveness of our single-image pipeline. For the multi-image branch, since no open-source real-world dataset is available, we construct a comparison set based on ScanNet [16] following ViewSpatial-Bench [13], and sample an equal number of examples from MVSR-Dataset. As shown in Table VI, both datasets provide comparable gains over the baseline, while our approach achieves similar improvements without using manual annotations, making it simpler and more scalable.

Impact of Data Scaling on Model Performance. To analyze the impact of data scale on model performance, we train the Qwen3-VL-8B-Instruct [17] using different amounts of data sampled from MVSR-Dataset and evaluate the performance of the fine-tuned Qwen3-VL-8B-Instruct on MVSR-Bench. As shown in Figure 3, the model performance consistently improves as the training data scale increases, indicating that multi-view spatial reasoning performance benefits from larger training datasets.

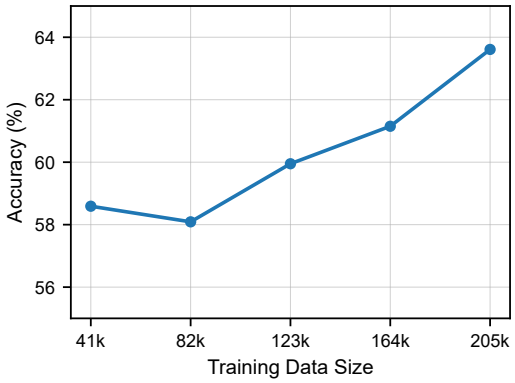

Fig. 3. Ablations with training data size.

V. CONCLUSION

In this work, we introduce a unified framework for multi-viewpoint spatial reasoning in multimodal large language models (MLLMs), addressing key limitations in existing methods that primarily focus on camera-centric reasoning. Our framework integrates two complementary data generation branches—Single Image and Multi-view Scene—to create both single-image and multi-image data for viewpoint-conditioned reasoning and joint reasoning over cross-view object associations and camera localization. We use this pipeline to construct the MVSR-Dataset and MVSR-Bench, a comprehensive benchmark for evaluating multi-viewpoint spatial reasoning. Experimental results, including fine-tuning Qwen3-VL-8B-Instruct [17] on the MVSR-Dataset, show significant improvements in performance across MVSR-Bench and existing benchmarks. These contributions provide a foundation for advancing spatial reasoning in MLLMs for applications in robotics, AR/VR, and autonomous driving, with future work focusing on expanding the framework for more complex tasks.

REFERENCES

[1] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 49 250–49 267, 2023.

[2] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu *et al.*, "LLaVA-OneVision: Easy Visual Task Transfer," *arXiv preprint arXiv:2408.03326*, 2024.

[3] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge *et al.*, "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[4] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu *et al.*, "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*, 2024, pp. 24 185–24 198.

[5] B. Chen, Z. Xu, S. Kirmani, B. Ichter, D. Sadigh, L. Guibas, and F. Xia, "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 455–14 465.

[6] A.-C. Cheng, H. Yin, Y. Fu, Q. Guo, R. Yang, J. Kautz, X. Wang, and S. Liu, "SpatialRGPT: Grounded Spatial Reasoning in Vision-Language Models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 135 062–135 093, 2024.

[7] D. Ko, S. Kim, Y. Suh, M. Yoon, M. Chandraker, H. J. Kim *et al.*, "ST-VLM: Kinematic Instruction Tuning for Spatio-Temporal Reasoning in Vision-Language Models," *arXiv preprint arXiv:2503.19355*, 2025.

[8] W. Ma, L. Ye, C. M. de Melo, A. Yuille, and J. Chen, "SpatialLLM: A Compound 3D-Informed Design towards Spatially-Intelligent Large Multimodal Models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 249–17 260.

[9] E. Daxberger, N. Wenzel, D. Griffiths, H. Gang, J. Lazarow, G. Kohavi, K. Kang, M. Eichner, Y. Yang, A. Dehghan *et al.*, "MM-Spatial: Exploring 3D Spatial Understanding in Multimodal LLMs," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 7395–7408.

[10] N. Deng, L. Gu, S. Ye, Y. He, Z. Chen, S. Li, H. Wang, X. Wei, T. Yang, M. Dou *et al.*, "InternSpatial: A Comprehensive Dataset for Spatial Reasoning in Vision-Language Models," *arXiv preprint arXiv:2506.18385*, 2025.

[11] J. Zhang, Y. Chen, Y. Zhou, Y. Xu, Z. Huang, J. Mei, J. Chen, Y.-J. Yuan, X. Cai, G. Huang *et al.*, "From Flatland to Space: Teaching Vision-Language Models to Perceive and Reason in 3D," *arXiv preprint arXiv:2503.22976*, 2025.

[12] W. Ma, Y.-C. Chou, Q. Liu, X. Wang, C. de Melo, J. Xie, and A. Yuille, "SpatialReasoner: Towards Explicit and Generalizable 3D Spatial Reasoning," *arXiv preprint arXiv:2504.20024*, 2025.

[13] D. Li, H. Li, Z. Wang, Y. Yan, H. Zhang, S. Chen, G. Hou, S. Jiang, W. Zhang, Y. Shen *et al.*, "ViewSpatial-Bench: Evaluating Multi-perspective Spatial Localization in Vision-Language Models," *URL https://arxiv. org/abs/2505.21500*, vol. 3, pp. 57–65.

[14] W. Ma, H. Chen, G. Zhang, Y.-C. Chou, J. Chen, C. de Melo, and A. Yuille, "3DSRBench: A Comprehensive 3D Spatial Reasoning Benchmark," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 6924–6934.

[15] J. Yang, S. Yang, A. W. Gupta, R. Han, L. Fei-Fei, and S. Xie, "Thinking in Space: How Multimodal Large Language Models See, Remember, and Recall Spaces," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10 632–10 643.

[16] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5828–5839.

[17] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-VL Technical Report," *arXiv preprint arXiv:2511.21631*, 2025. [Online]. Available: https://arxiv.org/abs/2511.21631

[18] X. Miao, H. Duan, Q. Qian, J. Wang, Y. Long, L. Shao, D. Zhao, R. Xu, and G. Zhang, "Towards Scalable Spatial Intelligence via 2D-to-3D Data Lifting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 945–959.

[19] A. Ray, J. Duan, R. Tan, D. Bashkirova, R. Hendrix, K. Ehsani, A. Kembhavi, B. A. Plummer, R. Krishna, K.-H. Zeng *et al.*, "SAT: Spatial Aptitude Training for Multimodal Language Models," *arXiv e-prints*, pp. arXiv–2412, 2024.

[20] W. Cai, I. Ponomarenko, J. Yuan, X. Li, W. Yang, H. Dong, and B. Zhao, "SpatialBot: Precise Spatial Understanding with Vision Language Models," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 9490–9498.

[21] P. Tong, E. Brown, P. Wu, S. Woo, A. J. V. IYER, S. C. Akula, S. Yang, J. Yang, M. Middepogu, Z. Wang *et al.*, "Cambrian-1: A Fully Open, Vision-Centric Exploration of Multimodal LLMs," *Advances in Neural Information Processing Systems*, vol. 37, pp. 87 310–87 356, 2024.

[22] X. Fu, Y. Hu, B. Li, Y. Feng, H. Wang, X. Lin, D. Roth, N. A. Smith, W.-C. Ma, and R. Krishna, "Blink: Multimodal Large Language Models Can See but Not Perceive," in *European Conference on Computer Vision*. Springer, 2024, pp. 148–166.

[23] Y.-H. Liao, R. Mahmood, S. Fidler, and D. Acuna, "Reasoning Paths with Reference Objects Elicit Quantitative Spatial Reasoning in Large Vision-Language Models," *arXiv preprint arXiv:2409.09788*, 2024.

[24] Y. Li, Y. Zhang, T. Lin, X. Liu, W. Cai, Z. Liu, and B. Zhao, "STI-Bench: Are MLLMs Ready for Precise Spatial-Temporal World Understanding?" *arXiv preprint arXiv:2503.23765*, 2025.

[25] S. Zhou, A. Vilesov, X. He, Z. Wan, S. Zhang, A. Nagachandra, D. Chang, D. Chen, X. E. Wang, and A. Kadambi, "VLM4D: Towards Spatiotemporal Awareness in Vision Language Models," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2025, pp. 8600–8612.

[26] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A Large-Scale, High-Quality Dataset for Object Detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8430–8439.

[27] OpenAI, "Introducing GPT-5," https://openai.com/index/introducing-gpt-5/, 2025, accessed: 2025-09-21.

[28] Google, "Gemini 3 Pro Image," https://deepmind.google/models/gemini-image/pro/, 2025, accessed: 2025-09-21.

[29] A. Raistrick, L. Lipson, Z. Ma, L. Mei, M. Wang, Y. Zuo, K. Kayan, H. Wen, B. Han, Y. Wang *et al.*, "Infinite Photorealistic Worlds using Procedural Generation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 12 630–12 641.

[30] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. J. Ostrow, A. Welihinda, A. Hayes, A. Radford, and et al., "GPT-4o System Card," *arXiv preprint arXiv:2410.21276*, 2024. [Online]. Available: https://arxiv.org/abs/2410.21276

[31] Google, "Gemini 2.5 Pro," https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro, 2025, accessed: 2025-10-01.

[32] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao *et al.*, "InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models," *arXiv preprint arXiv:2504.10479*, 2025.

[33] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang *et al.*, "Qwen2.5-VL Technical Report," *arXiv preprint arXiv:2502.13923*, 2025.

[34] Z. Chen, W. Wang, Y. Cao, Y. Liu, Z. Gao, E. Cui, J. Zhu, S. Ye, H. Tian, Z. Liu *et al.*, "Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[36] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov *et al.*, "The Open Images Dataset V4: Unified Image Classification, Object Detection, and Visual Relationship Detection at Scale," *International journal of computer vision*, vol. 128, no. 7, pp. 1956–1981, 2020.

[37] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "ScanNet++: A High-Fidelity Dataset of 3D Indoor Scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.