ROTATE: A Synthetic Pipeline for Spatial Relational Data with Systematic Variations

Anonymous ACL submission

Abstract

Spatial perception is a crucial component of intelligence and plays a vital role in understanding the physical world. Current multimodal large language models (MLLMs) exhibit promising spatial perception abilities. However, existing datasets are limited to absolute spatial perspectives or small-scale, unstructured collections lacking systematic variations in spatial relationships. This hinders models in understanding how rotation affects spatial relations. To address this issue, we propose ROTATE, a novel pipeline for synthesizing spatial relation datasets, and create the ROTATE dataset with 48K synthetic images, 608K captions, and 250K QA pairs covering both relative and absolute spatial perspectives. To further enhance the model's understanding of rotation, we propose a novel task: Spatial Difference Generation. In this task, the model must identify and generate both commonalities and differences in spatial relationships between paired images. Experimental results show that through three-stage training, the ROTATE dataset significantly improves the model's ability to comprehend spatial relationships from both relative and absolute perspectives. Furthermore, incorporating the Spatial Difference Generation task during training yields additional improvements in rotation comprehension and increases response consistency. Dataset and code will be published after the paper is published.

1 Introduction

Spatial awareness is an innate core capability for all visually enabled organisms, which underpins their movement and decision-making in complex environments. Consider a house cat hunting a mouse in a room: the predator does not simply pounce toward the prey's current position but rather anticipates the escape route based on the mouse's body orientation. This hunting strategy vividly demonstrates two complementary perceptual perspectives: the absolute perspective allows the cat to perceive its surroundings for path planning, while the relative perspective enables it to infer potential movement trajectories based on the prey's orientation. As shown in Figure 1, absolute perspective spatial relationships are centered on the observer and generally do not account for the influence of the orientation of an object. In contrast, spatial relationships in the relative perspective are centered on a specific object in the field of view, where the orientation of the object significantly impacts the perceived spatial relationships. This dual-perspective mechanism holds critical application value for systems that require environmental perception for navigation, such as autonomous driving systems. For example, the absolute perspective provides autonomous vehicles with ego-centric perception of road geometry, lane markings, and obstacle positions, while the relative perspective plays an indispensable role in understanding other traffic participants' behavioral intentions. The system can predict pedestrian crossing intentions from their body orientation, anticipate lane changes by detecting adjacent vehicles' wheel angles, and determine right-of-way at intersections by analyzing approaching vehicles' heading angles.

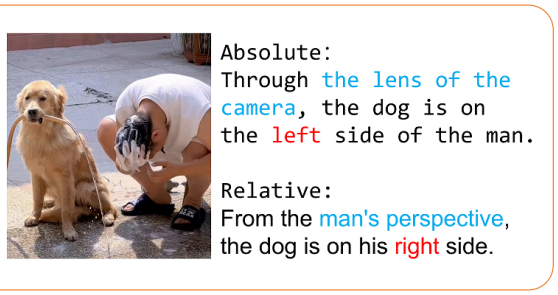

Figure 1: An example of absolute perspective and relative perspective.

Existing multimodal spatial perception datasets, such as VG(Krishna et al., 2017), GQA(Hudson and Manning, 2019), and VSR(Liu et al., 2023), rely on manual annotations and lack automated

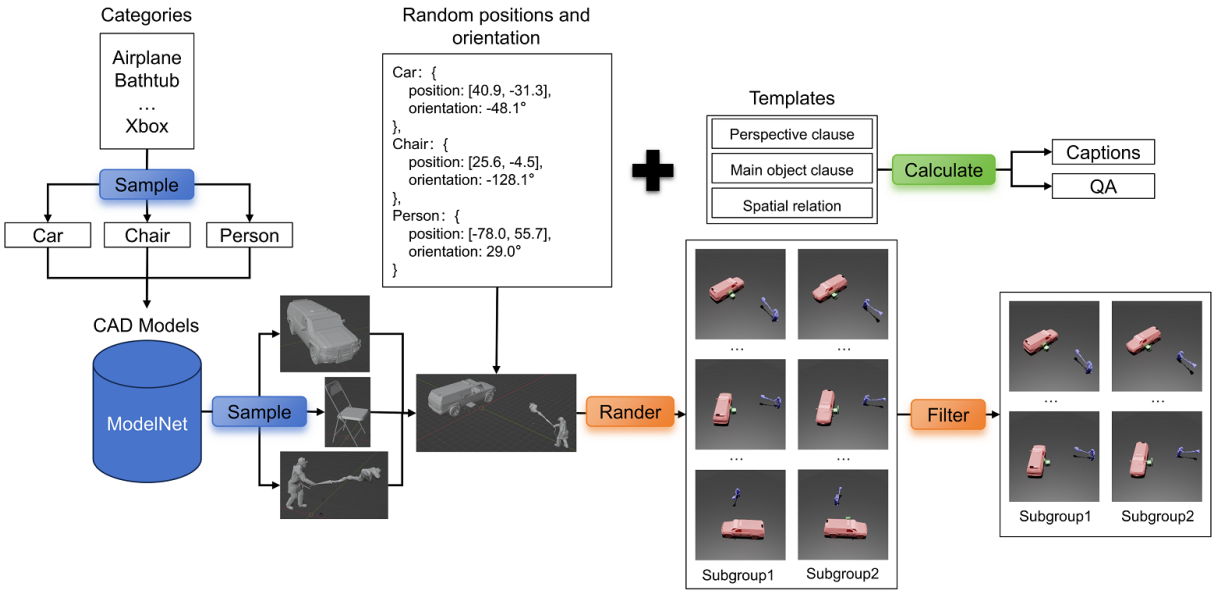

Figure 2: Overview of ROTATE pipeline.

construction methods, making it difficult to scale them up. Some datasets, such as SpatialVLM(Chen et al., 2024a) and SpatialRGPT(Cheng et al., 2024), adopt automated approaches to build spatially relevant datasets. However, in either computer vision or multimodal learning, determining the orientation of objects from real-world images remains a significant challenge. As a result, these datasets lack spatial relationships from a relative perspective. Since these datasets extract spatial relationships from real-world images, neither object orientation nor camera angles can be manually controlled, leading to a lack of systematic spatial variation. Synthetic image-based datasets, such as CLEVR(Johnson et al., 2017), typically generate images using geometric symmetric shapes or primitives, where objects inherently lack the notion of directionality.

To address these challenges, we propose the ROTATE pipeline. As illustrated in Figure 2, the ROTATE pipeline uses Blender(Community, 2018) to render CAD models from the ModelNet(Wu et al., 2015) dataset in images containing various spatial relationships by configuring different orientations and positions. Each image is paired with a textual description or QA pair generated using carefully designed templates. The ROTATE pipeline renders images in groups of 8 pairs (16 images total). Each group consists of two subgroups of images captured by eight cameras from eight different viewpoints. Crucially, each corresponding image pair contains identical objects with identical layouts, with only the orientations of all objects reversed between the paired images. Using this pipeline, we construct the ROTATE dataset, comprising 48K synthetic images, with 608K captions (covering both relative and absolute perspectives) and 250K QA pairs. To further enhance the consistency of the answer, we introduce a novel task: **S**patial **D**ifference **G**eneration (SDG). The SDG task requires models to identify and describe similarities and differences in spatial relationships between image pairs from specified perspectives.

The experimental results demonstrate that while spatial perception from relative perspectives remains a significant challenge for current multimodal models, three-stage training with ROTATE-synthesized data can substantially enhance this capability. In particular, without introducing additional data, our proposed SDG task significantly enhances both spatial perception accuracy and the consistency of their spatial reasoning. The contributions of this work can be summarized as follows:

- We propose the ROTATE pipeline, which enables large-scale generation of spatial perception datasets with systematic spatial variations. Based on this pipeline, we construct the ROTATE dataset.

- We introduce a novel task: **S**patial **D**ifference **G**eneration (SDG), which significantly improves spatial reasoning capabilities and improves the consistency of the answer for spatial perception questions.

2 Related Work

Prior work has extensively explored spatial perception. The VG(Krishna et al., 2017) and GQA(Hudson and Manning, 2019) datasets construct spatially aware QA pairs through manually annotated scene graphs, while VSR(Liu et al., 2023) provides human-labeled captions for three reference frames. Spatial-MM(Shiri et al., 2024) and MM-Vet(Yu et al., 2024) collect limited images and design spatial perception benchmarks based on them. Since these datasets are based on human annotation, they inherently incorporate both absolute and relative perspectives. However, their scalability is constrained by this manual annotation process. The CLEVR(Johnson et al., 2017) dataset generates synthetic images using geometric primitives. Due to the symmetric nature of most primitives lacking directional concepts, CLEVR excludes relative perspectives. SpatialVLM(Chen et al., 2024a) and SpatialRGPT(Cheng et al., 2024) employ existing tools to extract 3D scene graphs from web-collected images, then generate QA pairs from these graphs. However, since object orientation detection remains challenging in real images, these datasets only consider absolute spatial relationships. With the exception of CLEVR, all the aforementioned datasets source images from the Web, resulting in limited systematic variation in spatial relationships. What's Up(Kamath et al., 2023) dataset attempted to address this by manually arranging objects to capture systematic spatial variations through photography. However, constrained by labor-intensive processes and predefined spatial relationship categories, What's Up remains limited in scale and lacks relative spatial relationship data.

3 ROTATE

3.1 Pipeline

The construction of relative-perspective images requires accurate knowledge of the orientations of objects. Since current techniques cannot reliably extract object orientations from real-world images, generating relative spatial relationships from natural imagery remains infeasible. To address this fundamental limitation, we turn to 3D rendering technology. By systematically adjusting the orientation and position of 3D models, we achieve controlled variations in spatial relationships. Crucially, known object orientations enable precise computation of relative spatial relationships between any pair of objects. For reliable orientation information, we select the ModelNet(Wu et al., 2015) dataset, where all 3D models are pre-aligned with consistent orientation baselines. This fundamental property allows the deterministic calculation of both absolute and relative spatial perspectives.

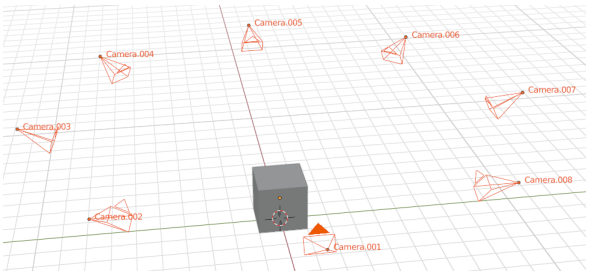

Figure 3: Schematic diagram of 8 cameras.

Render: Figure 2 illustrates the detailed workflow of the ROTATE pipeline. During the rendering of each image group, ROTATE first selects 2-3 object categories from the ModelNet dataset and chooses one 3D model per category for rendering. We classify ModelNet objects into three distinct types: (1) objects with inherent directional properties (e.g., humans, vehicles), (2) objects that humans typically use in fixed orientations (e.g., chairs), and (3) objects without directional concepts (e.g., flower pots) - with the first two categories collectively referred to as "directional objects." To ensure that every image can generate relative perspective data, ROTATE always includes at least one directional object in each scene. The pipeline then randomly generates each object's position and orientation while performing collision detection to prevent bounding-box overlaps between objects. After sequentially applying different colored materials to each model, ROTATE renders two image subgroups from the 8 camera angles shown in Figure 3: the first subgroup maintains the originally generated orientations, while the second subgroup applies an additional 180 degree rotation to all object orientations.

Filter: We implemented a rigorous filtering process to eliminate images with object occlusion or recognition ambiguity. Specifically, we used the InternVL2-26B model as our filtering mechanism. For each candidate image, we systematically queried the model about the presence of every individual object in the scene. Only images in which InternVL2 confidently confirmed the presence of all objects were retained in the final dataset.

Captions and QA pairs: ROTATE employs carefully designed templates to generate captions

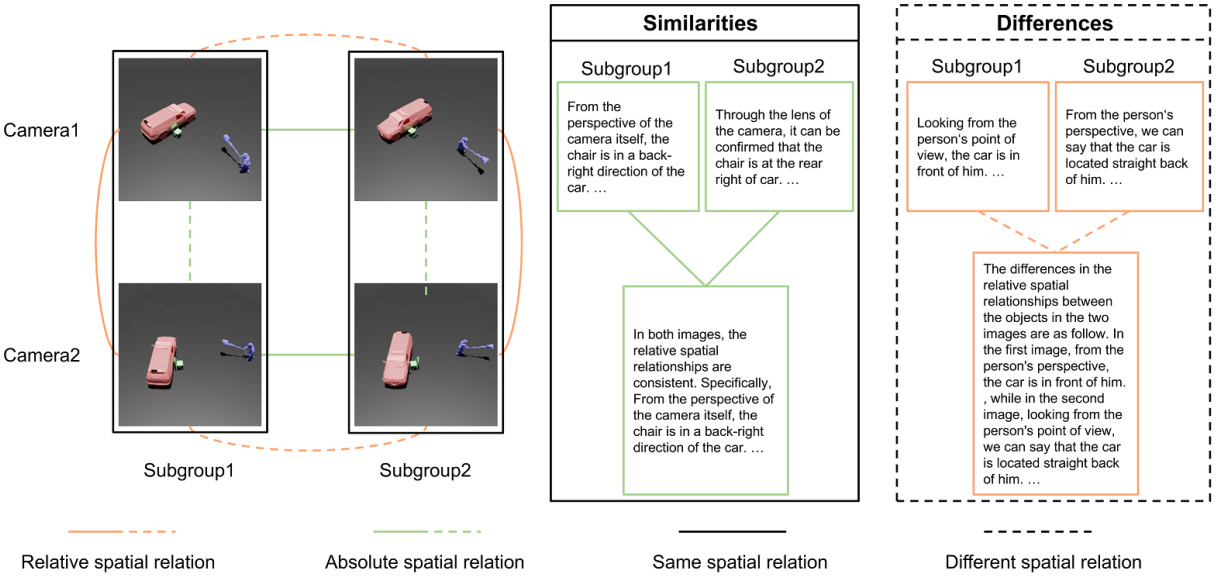

Figure 4: Examples of Spatial Difference Generation data.

and question-answer pairs for each image. To enhance linguistic diversity, we decompose spatial relation descriptions into three components: the perspective clause (relative or absolute; see Appendix A), main object clause (added only in relative perspectives for objects that humans typically use in fixed orientations; Appendix B), and spatial relation (selected from Appendix C based on task) and recombine them variably. We define two types of spatial relation: cardinal directions (front, back, left, right, and their diagonals) and clock directions (12 hour positions). These underpin two QA task templates: judgment tasks (verifying spatial correctness) and prediction tasks (identifying precise spatial relations).

3.2 Dataeset

Using the ROTATE pipeline, we initially synthesized 10,000 image groups. After filtering, the final dataset comprises 48K images, 608K captions, and 250K QA pairs. For caption generation, we used the first 9,000 groups, divided into training and validation sets. The remaining 1,000 groups were allocated for the construction of QA pairs, divided into training, validation, and test sets. As demonstrated in Table 1 and Appendix D, the textual distributions in the ROTATE dataset are well balanced, a direct result of our carefully designed caption and QA generation rules. However, we note two important filtering effects: (1) While object counts were randomly selected during rendering, three-object scenes exhibited higher occlusion rates than two-object scenes, leading to disproportionate filtering of three-object images. (2) Three-object images could still generate questions about two-object relationships, resulting in significantly more two-object questions, as shown in Table 2's comparison with other multimodal spatial relation datasets.

3.3 Spatial Difference Generation

To further enhance the model's comprehension of rotation and improve its consistency in spatial understanding when only orientation or camera angles differ, we propose the Spatial Difference Generation (SDG) task. The SDG task requires the model to process two similar images as input and identify both similarities and differences in their spatial relationships, either from relative or absolute perspectives. Specifically, the model must first explicitly state whether the relationships are similar or different and then generate fine-grained comparisons of spatial relationships at the object level. Figure 4 illustrates the methodology for constructing the task dataset. To minimize data collection efforts, we reuse caption data from the ROTATE dataset. In ROTATE, images within the same subgroup share identical relative spatial relationships, while images from different subgroups but captured from the same camera viewpoint share identical absolute spatial relationships. By using pairs of images from either the same subgroup or the same camera viewpoint as input, we systematically organize their similarities and differences into structured spatial-difference captions.

ROTATE		Modality		Viewpoint		Task		Relation Type		Object Numbers	
Component	Split	Images	Text	Relative	Absolute	Judgement	Prediction	Cardinal	Clock	2	3
Captions	train	41K	595K	50%	50%	-	-	100%	100%	100%	32.43%
	val	1K	13K	50%	50%	-	-	100%	100%	100%	34.35%
QA	train	5.6K	225K	51.47%	48.53%	66.67%	33.33%	50%	50%	81.26%	18.74%
	val	300	12K	52.63%	47.37%	66.67%	33.33%	50%	50%	80.67%	19.33%
	test	312	13K	53.87%	46.13%	66.67%	33.33%	50%	50%	81.23%	18.77%
Images	-	-	-	-	-	-	-	-	-	67.37%	32.63%

Table 1: ROTATE dataset distribution. To comprehensively characterize spatial relationships in images, each caption incorporates both types of spatial relations. Crucially, even when an image contains three objects, it inherently allows describing spatial relationships between two objects - therefore, every caption contains a section describe the spatial relationship between paired objects.

Dataset	Images	Captions	QA	Symmetric Image	Relative Viewpoint	Systematic Variation
VG*	108K	5M	1.7M	✗	✗	✗
GQA*	113K	-	22M	✗	✗	✗
VSR	6940	11K	-	✗	✓	✗
Spatial-MM	3.1K	-	3.1K	✗	✓	✗
MM-Vet*	187	-	205	✗	✗	✗
SpatialVLM	10M	-	2B	✗	✗	✗
SpatialRGPT	1M	-	8.7M	✗	✗	✗
What's up	820	820	-	✗	✗	✓
CLEVR	100K	-	1M	✓	✗	✗
ROTATE	48K	608K	250K	✓	✓	✓

Table 2: Comparison of existing multimodal spatial relationship datasets. Dataset with "*" indicates that only a portion of the dataset is related to spatial relationships

3.4 Training Strategy

As illustrated in Figure 5, we adopt a carefully designed three-stage training approach using the ROTATE dataset to progressively improve the spatial understanding of the model and the consistency of the answers across multimodal inputs. In the first stage, we train the entire model using image captions, which provide rich descriptions of object arrangements and spatial relationships. This stage enables the model to develop a foundational understanding of how objects interact within a scene. To maintain parameter efficiency while ensuring effective learning, we employ LoRA(Hu et al., 2021) for both the language model and the visual encoder, allowing them to adapt to the multimodal task with minimal training parameters. Meanwhile, the MLP layer, which bridges the visual and linguistic modalities, undergoes a full fine-tuning to better align cross-modal representations. The second stage focuses on improving the spatial consistency of the model between images that share identical objects but differ in their arrangement. To achieve this, we train the model with the proposed SDG task. In particular, to minimize additional data overhead, we repurpose the original captions from the first stage into SDG-compatible formats through automated rule-based transformations. During this phase, we freeze the visual encoder to stabilize the training and only update the LM and MLP layers, ensuring that the model hones its spatial differentiation skills without overfitting to low-level visual features. Finally, in the third stage, we further fine-tune the model in QA pairs using the same training protocol as in the second stage. This phase adapts the model to downstream spatial QA tasks, reinforcing its ability to generate accurate and coherent responses grounded in visual-spatial understanding.

Direction	1	2	3	4	5	6	7	8	9	10	11	12
Front	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓
Front-Right	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Right	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗
Back-Right	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
Back	✗	✗	✗	✗	✓	✓	✓	✗	✗	✗	✗	✗
Back-Left	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗
Left	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✗
Front-Left	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗

Table 3: Same direction under different spatial relationship types.

4 Experiement

4.1 Configurations

Models: We trained the InternVL2-8B model using our proposed three-stage training strategy as a baseline. To the best of our knowledge, there are currently no viable training methods for relative spatial relationships currently exist. Therefore, for comparison, we fine-tune the InternVL-8B model using only the ROTATE QA training set. Furthermore, we evaluated zero-shot performance across models of varying scales, including open-source models (InternVL2(Chen et al., 2024b), Llama3.2(AI@Meta, 2024), LlavaNext(Liu et al., 2024), Qwen2.5-VL(Bai et al., 2025)) and proprietary models (GPT-4o(OpenAI et al., 2024)). Appendix E shows the prompt for the zero-shot exper-

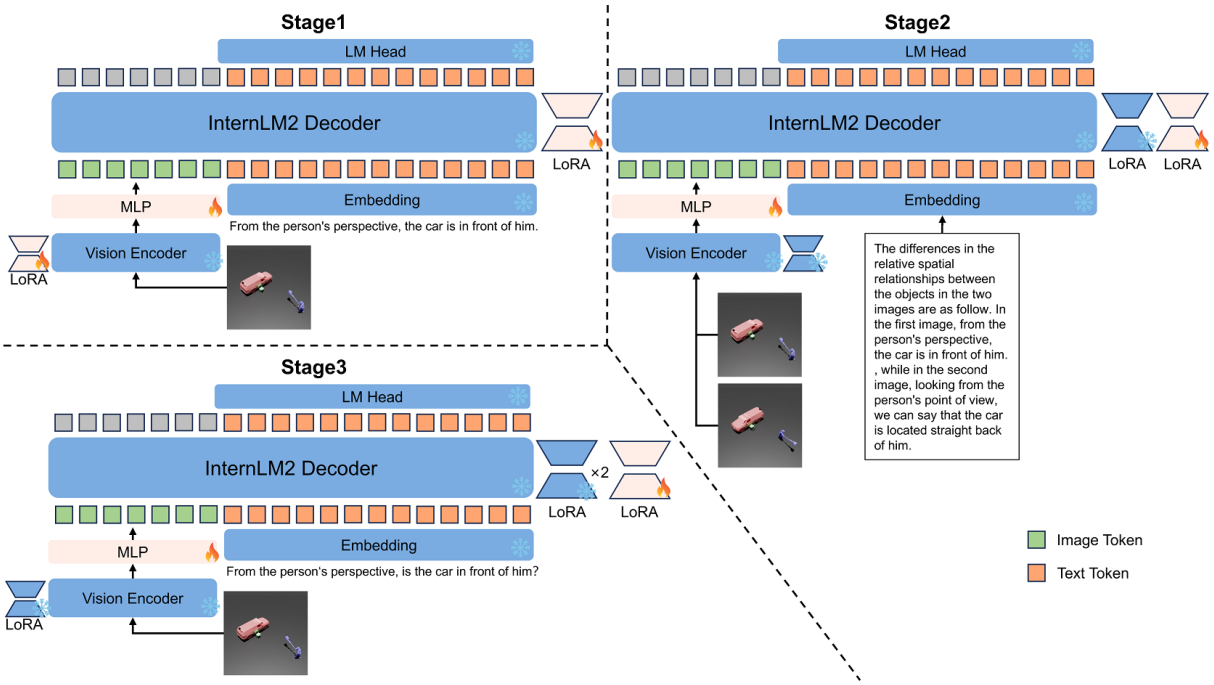

Figure 5: Three stage training strategy. In the first stage, the model is trained on captions to learn spatial relationships across perspectives. During the second stage, we freeze the visual encoder and employ the SDG task to improve rotation understanding. Finally, the model is fine-tuned using QA pairs.

iment. All experiments were conducted using four A100 80GB GPUs. The complete training hyperparameters for the three-stage training strategy are provided in Appendix F.

Metric: We employ accuracy as the primary metric to evaluate the performance of the model in different tasks. Furthermore, to assess whether the model has genuinely acquired spatial knowledge, we introduce two consistency metrics: **R**elationship **T**ype **C**onsistency (RTC) and **R**otation **C**onsistency(RC). Relation-type consistency measures the model's ability to identify equivalent directions using different spatial relation types (e.g., "12 o'clock direction" and "front" represent the same direction). Specifically, we evenly divide the circle into 12 and 8 equal sectors, respectively, representing directional ranges through angular intervals for different types of relationship. We define two directions as identical when their angular intervals under a given relationship type have a non-empty intersection. Table 3 shows the same direction under different types of spatial relation. This metric is calculated as the proportion of question pairs that differ only in the required type of spatial relation, where the model provides consistent directional answers. Rotation consistency evaluates response invariance across varying camera viewpoints or object orientations, reflecting the model's understanding of rotational transformations. We quantify this using Normalized Entropy (NE), where inconsistent responses indicate higher disorder (greater entropy). The normalization process eliminates potential biases caused by varying the sizes of image groups in the evaluation. The Normalized Entropy is calculated as Equation 1 when N questions should yield identical answers, where M denotes the total number of distinct possible answers, and n_i represents the occurrence count of the i^{th} answer:

$$NE = \frac{\sum_{i=1}^{M} -\frac{n_i}{N} log_2 \frac{n_i}{N}}{log_2 N} \quad (1)$$

To handle cases where the model generates responses outside our predefined answer set, we treat each such unique response as a distinct category to prevent irrelevant answers from lowering the NE metric.

4.2 Main Results

Table 4 presents the performance of different models and training approaches on the ROTATE QA test set. Current multimodal models show limited capability in spatial relationship understanding, where even substantial increases in model scale yield negligible accuracy improvements. In particular, models fine-tuned solely on the ROTATE QA

Model	Task↑		Viewpoint↑		Relation Type↑		Total
	Judgement	Prediction	Relative	Absolute	Cardinal	Clock	
GPT-4o	53.43	35.08	38.40	57.73	47.74	46.89	47.32
InternVL2-8B	52.33	19.77	40.25	42.91	44.97	37.99	41.48
InternVL2-76B	53.35	28.57	39.96	51.08	48.80	41.38	45.09
Llama3.2-11B	50.29	8.52	34.79	38.21	37.33	35.40	36.37
Llama3.2-90B	52.12	24.46	37.41	49.33	46.36	39.45	42.90
LlavaNext-13B	50.00	11.95	36.77	37.96	39.37	35.26	37.32
LlavaNext-34B	53.64	14.33	37.85	43.67	43.42	37.66	40.54
Qwen2.5-VL-7B	51.77	16.64	38.33	42.09	44.96	35.17	40.06
Qwen2.5-VL-72B	51.27	29.34	34.26	55.30	42.67	45.25	43.96
Finetune	56.24	34.31	40.83	58.40	51.87	46.00	48.93
Ours	80.53	51.24	52.57	92.02	71.96	69.58	70.77

Table 4: Accuracy of multimodal models on ROTATE QA test set. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and pink backgrounds.

Model	RTC↑	RC$\times 10^{-2}$ ↓		
		Relative	Absolute	Total
GPT-4o	35.72	55.08	54.39	54.75
InternVL2-8B	11.48	57.68	54.24	56.06
InternVL2-76B	10.53	45.31	42.69	44.08
Llama3.2-11B	12.38	83.07	71.78	77.75
Llama3.2-90B	28.80	53.52	44.30	51.17
LlavaNext-13B	9.35	65.89	64.77	65.36
LlavaNext-34B	13.52	28.52	26.46	27.55
Qwen2.5-VL-7B	6.69	68.10	64.33	66.32
Qwen2.5-VL-72B	34.30	32.16	24.85	28.72
Finetune	44.02	57.68	50.88	54.48
Ours	64.90	51.30	11.70	32.64

Table 5: Consistency of multimodal models on ROTATE QA test set, where RTC represents relationship type consistency and NE represents Normalized Entropy. For each column, the highest, the second, and the third highest figures are highlighted by green, orange and pink backgrounds.

training set show only marginal improvements over GPT-4o. In contrast, our three-stage training strategy achieves a remarkable 21.84% performance gain compared to the standalone finetuning. We comprehensively evaluate the spatial reasoning capabilities of existing multimodal models through four key dimensions: task, viewpoint, spatial relation, and consistency.

Task: Unexpectedly, current multimodal models perform poorly in both judgment and prediction tasks. For the judgment task, even after fine-tuning, the accuracy stays just above 50%, close to random guessing. For the prediction task, the fine-tuned InternVL2-8B model improves significantly but still falls slightly short of GPT-4o, demonstrating the inherent difficulty of the ROTATE dataset. However, our three-stage training enables even smaller models to achieve substantially higher prediction accuracy.

Viewpoint: Current models consistently underperform in relative perspective tasks compared to absolute perspective tasks, regardless of training. Fine-tuning improves performance primarily on absolute perspective tasks, with minimal gains for relative perspective understanding. However, our three-stage training yields significant improvements for both perspectives. Despite this advancement, there remains a substantial performance gap between relative and absolute perspectives, confirming that understanding relative spatial relationships remains a challenge in the field.

Spatial Relation: The inherent data scarcity of clock direction expressions (compared to cardinal directions) in daily use leads to systematically weaker performance on clock direction tasks in all models. Although fine-tuning improves both types of representations, it shows limited efficacy in bridging this performance gap, reducing the disparity in InternVL2-8B from 6.98% to just 5.87%. In contrast, our three-stage training demonstrates significantly stronger generalization ability, narrowing the gap to 2.38% and proving particularly effective in learning low-frequency spatial patterns.

Consistency: We evaluate the consistency by examining the agreement of the answer (regardless of correctness) to assess whether the modes

establish conceptual connections. Table 5 presents the consistency of the type of relation(RTC) and the consistency of rotation(RC). The RTC metric reveals that most models, except GPT-4o, Qwen2.5-VL-72B, and Llama3.2-90B, fail to recognize the relationship between the two spatial representation methods. Even these three large-scale models show limited alignment capability. Training significantly improves this understanding, enabling better representation alignment. Unexpectedly, fine-tuning only marginally boosts relative spatial consistency while failing to enhance rotation consistency in relative perspectives. However, our three-stage training effectively improves rotation consistency, though gains in relative perspectives remain substantially lower than in absolute perspectives, highlighting the persistent challenge of relative spatial reasoning. InternVL2-76B, LLava-Next-34B, and Qwen2.5-VL-72B demonstrate exceptionally high rotation consistency scores. However, statistical analysis of their responses reveals severe answer biases that artificially inflate these metrics. Under relative perspective conditions, InternVL2-76B produces "left-back" responses 54% of the time; LLaVA-Next-34B shows a 50.3% bias toward "right" answers, with this preference increasing to 89.9% for "10 o'clock" responses in clock-direction tasks; Qwen2.5-VL-72B generates "front"/"back" responses in less than 1% of cases. These extreme biases elevate rotation consistency metrics without indicating a genuine understanding of how rotation affects spatial relationships.

4.3 Ablation

Strategy	Accuracy	Consistency	
		RTC↑	RC$\times 10^{-2}$ ↓
0-shot	41.48	11.48	56.06
S3	48.93	44.02	54.48
S1+S3	56.64	42.46	54.61
Ours	70.77	64.90	32.64

Table 6: The ablation experiment results of the three-stage training strategy. S3: only uses the third-stage training model, that is, only uses question answer pairs for fine-tuning. S1+S3: the first and third stages, which involve training with caption first and then fine-tuning with question answer pairs.

As evidenced by Table 11, the incorporation of captions during training improves the accuracy of the model, but does not improve consistency. The subsequent introduction of the SDG task yields significant improvements in both accuracy and consistency metrics. This demonstrates that caption-based training solely boosts predictive accuracy without fostering understanding of either (1) the relationships between different spatial relation types or (2) rotation's impact on spatial relationships. Crucially, the SDG task achieves these advanced comprehension capabilities without requiring additional training data. This phenomenon potentially reveals a fundamental limitation in current multimodal model training paradigms: Although large-scale caption training effectively enhances multimodal comprehension and facilitates knowledge acquisition, it does not allow models to systematically organize these discrete knowledge components into a unified cognitive framework. In contrast, the SDG task artificially constructs inter-knowledge relationships through comparative learning, thereby enabling the model to establish an integrated knowledge system. Remarkably, this architectural improvement achieves substantial performance gains without requiring additional training data. In addition, contrastive learning also introduces comparative information. As detailed in Appendix G, we conducted experiments replacing the SDG task with contrastive learning. Our findings reveal that contrastive learning is not suitable for current generative multimodal models and fails to deliver performance improvements. This limitation may be closely tied to the fundamental differences in task objectives between contrastive learning and next token prediction task, and the computational resource requirements involved.

5 Conclusion

In this paper, we present ROTATE, a novel pipeline for batch synthesis of multimodal spatial perception data. Using this pipeline, we construct the ROTATE dataset and propose a new Spatial Difference Generation task(SDG). Our experimental results demonstrate that while spatial understanding from relative perspectives remains challenging for current multimodal models, our three-stage training protocol using the ROTATE dataset yields significant improvements in spatial perception capabilities.Furthermore, the SDG task enhances the model's spatial perception capability without requiring additional data, while also improving response consistency across varying viewpoints and different spatial representation methods.

Limitations

Due to constraints in our rendering capabilities, synthesized images lack photorealism and could be improved in several aspects, such as incorporating more realistic material textures and detailed background environments. Furthermore, existing spatial perception datasets suffer from three critical limitations: (1) their small scale, (2) inconsistent definitions of spatial relationships across datasets, and (3) the lack of explicit annotations distinguishing between relative and absolute perspectives. Consequently, evaluations on these datasets provide a limited reference value to assess model performance.

References

AI@Meta. 2024. Llama 3 model card.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923.

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

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, and 1 others. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821.

An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. 2024. Spatialrgpt: Grounded spatial reasoning in vision-language models. In NeurIPS.

Blender Online Community. 2018. Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. Preprint, arXiv:2106.09685.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 6693–6702.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), page 1988–1997. IEEE.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. What's "up" with vision-language models? investigating their struggle with spatial reasoning. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 9161–9175, Singapore. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73.

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

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. Preprint, arXiv:2410.21276.

Fatemeh Shiri, Xiao-Yu Guo, Mona Golestan Far, Xin Yu, Reza Haf, and Yuan-Fang Li. 2024. 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, Miami, Florida, USA. Association for Computational Linguistics.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1912–1920.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. MM-vet: Evaluating large multimodal models for integrated capabilities. In Proceedings of the 41st International Conference on Machine Learning, volume 235 of Proceedings of Machine Learning Research, pages 57730–57754. PMLR.

A Template for Viewpoint

Listing 1: Template for relative viewpoint.

```
'relative_prefix': [
    'from [A]\'s perspective',
    'looking from [A]\'s point of view',
    'from [A]\'s viewpoint',
    'from [A]\'s point of view',
    'looking from [A]\'s perspective',
    'according to [A]\'s view',
    'in [A]\'s line of sight',
    'through [A]\'s perspective',
    'from [A]\'s vantage point',
    'aligned with [A]\'s gaze',
    'following [A]\'s line of sight',
    'in the direction [A] is facing',
    'through the lens of [A]\'s
       perspective',
    'with [A]\'s forward focus',
    'seen as if standing in [A]\'s
       position',
    'relative to [A]\'s front',
],
```

Listing 2: Template for absolute viewpoint.

```
'absolute_prefix': [
    'from the perspective of the camera
       itself',
    'in terms of a purely objective
       angle',
    'through the lens of the camera',
    'from the absolute viewpoint that
       the camera records',
    'viewed directly from an absolute
       standpoint',
    'seen directly from this position',
    'seen directly from the image\'s
       perspective',
    'viewed precisely as the image shows
       ',
    'as observed from the angle
       presented in the image',
    'in view as the image frames it'
],
```

B Template for Main Object

Listing 3: Template for main object.

```
'used_directed_prefix': [
    'if someone is [V] [A]',
    'while someone is [V] [A]',
    'as a person [V] [A]',
    'when a person is [V] [A]',
    'if a person is [V] [A]'
],
```

C Template for Spatial Relationship

Listing 4: Example template for captions with two objects.

```
'directed': {
    'location': {
        'left': ['[B] is on the left
           side of [A]', ],
        'right': ['we can confirm [B] is
            positioned on [A]\'s right
           side'],
    }
    'clock': ['[B] is located at [A]\'s
       [X] o\'clock', ]
}
'used_directed': {
    'location': {
        'left': ['[B] is on his left
           side', ],
        'right': ['[B] is positioned to
           his right', ],
    }
    'clock': ['[B] is positioned at [X]
       o\'clock', ]
}
```

Listing 5: Example template for qa pairs with three objects.

```
'true_or_false': {
    'question': {
        'left': ['is [B] to the left of
           [C]', ],
        'clock': ['is [B] located at the
            [X] o\'clock position
           relative to [C]', ],
    }
    'answer': {
        'left': {
            'true': ['[B] is indeed to
               the left of [C]', ],
            'false': ['[B] is not on the
                left side of [C]', ]},
        'clock': {
            'true': ['[B] is indeed at [
               X] o\'clock position',
               ],
            'false': ['[B] is not at the
                [X] o\'clock position',
                ]},
    }
}
'predict': {
    'question': {
        'location': ['what direction is
           [B] relative to [C]', ],
        'clock': ['what is the
           approximate clock position
           of [B] relative to [C]', ]
    },
    'answer': {
        'left': ['[B] is on the left
           side of [C]', ],
        'clock': ['[B] would appear to
           be around [X] o\'clock
           relative to [C]', ],
    }
}
```

D Answer Distribution

QA Split	Yes	No
train	50%	50%
val	50%	50%
test	50%	50%

Table 7: Distribution of answers for the judgment task.

QA Split	Left	Right	Front-Left	Front-Right
train	12.95%	12.93%	11.16%	11.80%
val	11.21%	10.92%	10.20%	16.92%
test	10.50%	10.46%	10.18%	11.30%
QA Split	**Front**	**Back**	**Back-Left**	**Back-Right**
train	14.07%	14.20%	11.80%	11.08%
val	11.70%	11.50%	17.50%	10.05%
test	17.37%	17.46%	11.16%	11.58%

Table 8: Distribution of answers for cardinal directions.

QA Split	1	2	3	4
train	9.02%	7.11%	8.89%	8.55%
val	13.10%	6.67%	8.26%	6.77%
test	12.28%	5.51%	5.74%	9.10%
QA Split	**5**	**6**	**7**	**8**
train	6.76%	9.64%	9.10%	7.10%
val	6.72%	7.93%	13.68%	6.62%
test	7.56%	10.69%	11.90%	5.79%
QA Split	**9**	**10**	**11**	**12**
train	8.92%	8.55%	6.78%	9.58%
val	8.60%	6.96%	6.62%	8.07%
test	5.84%	8.40%	6.63%	10.55%

Table 9: Distribution of answers for clock directions.

E Zero-shot prompt

Listing 6: Prompt for zero-shot experiments.

```
[IMAGE] I will ask you a question. If
the question requires a yes/no judgment,
 answer strictly "Yes" or "No". For
short-answer questions: if the answer is
 a direction, use only one of these:
front, back, left, right, front-left,
back-left, front-right, back-right. If
the answer is a clock position, use the
format "[X] o'clock" where [X] is a
number (1-12) or its English word (e.g.,
 "2 o'clock" or "two o'clock"). Provide
only the answer without explanations.
Question: [QUESTION]
Answer:
```

F Experimental Hyperparameters

	Stage1	Stage2	Stage3
Batch size	64	64	32
Micro batch size	4	1	4
Max token num	1024	1536	512
Total step	5000	5000	10000
Optimizer	AdamW		
LR schedule	Linear warmup cosine decay		
LR	4e-5		
Min LR	2e-5		
Warm up start LR	3e-5		
Weight decay	0.05		
Warm up step	100		

Table 10: Experiemental hyperparameters for each stage. The hyperparameters that are the same in all three stages will only be displayed once. LR: Learning Rate

G Contrastive Learning

Strategy	Accuracy	Consistency	
		RTC↑	RC$\times 10^{-2}$ ↓
0-shot	41.48	11.48	53.93
S3	48.93	44.02	54.48
S1+S3	56.64	42.46	54.48
contrastive	55.12	41.79	53.86
Ours	70.77	64.90	32.51

Table 11: The ablation experiment results of the three-stage training strategy with comparative learning. S3: only uses the third-stage training model, that is, only uses question answer pairs for fine-tuning. S1+S3: the first and third stages, which involve training with caption first and then fine-tuning with question answer pairs.

We also experimented with replacing the second-stage SDG task with contrastive learning, investigating whether this approach could improve the model's consistency in spatial relationship understanding. As shown on the left side of Figure 4, we leverage the inherent correlations in spatial relationships among image groups within the ROTATE dataset to construct positive and negative samples for our contrastive learning task, where each data instance consists of four images (similar to those in Figure 4) representing two subgroups captured from two different camera angles. For relative perspective conditions, positive samples are images from the same subgroup sharing identical relative spatial relationships, while negative

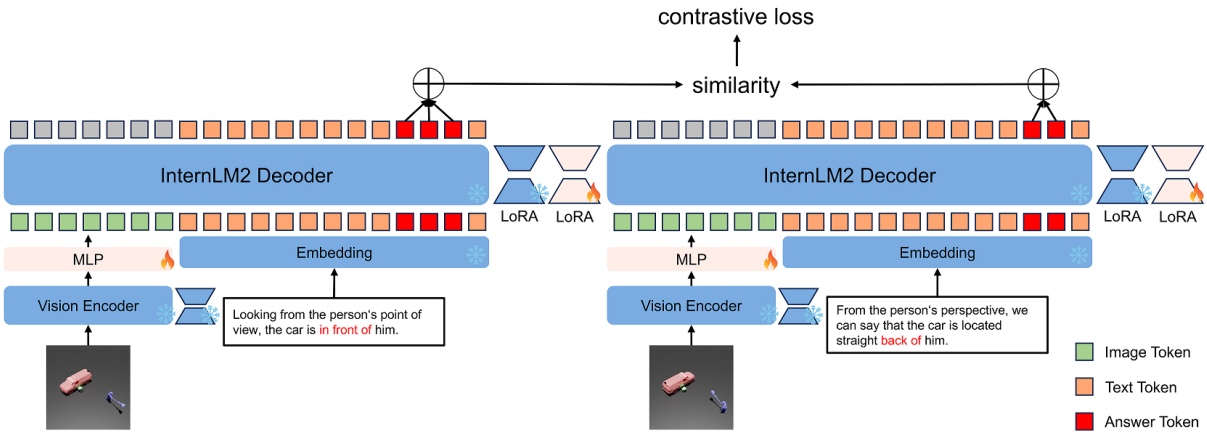

Figure 6: Comparative Learning of LVLM.

samples include both images from the same camera angle but different subgroups (relative perspective) and the image's own absolute perspective relationships; correspondingly, for absolute perspective conditions, positive samples are images rendered from the same camera angle, with negative samples consisting of images from the same subgroup but different camera angles (absolute perspective) and the image's own relative perspective relationships. To our surprise, contrastive learning fails to enhance the model's spatial perception capabilities, improving only its consistency in relative spatial relationships. This phenomenon may stem from conflicting gradient update directions between contrastive learning and the QA task optimization objectives. Our experiments reveal that contrastive learning solely affects the convergence speed of the loss function in the third training stage, with nearly identical final loss values and validation performance regardless of its inclusion.