

Euclid’s Gift: Enhancing Spatial Perception and Reasoning in Vision-Language Models via Geometric Surrogate Tasks

Shijie Lian^{1,2,*}, Changti Wu^{3,2,*}, Laurence Tianruo Yang^{4,1,†}, Hang Yuan², Bin Yu², Lei Zhang³, Kai Chen^{5,†}

¹Huazhong University of Science and Technology ²Zhongguancun Academy ³East China Normal University ⁴Zhengzhou University ⁵Zhongguancun Institute of Artificial Intelligence

Abstract

Spatial intelligence spans a rich suite of abilities, including visualising and transforming shapes, mentally rotating objects, judging relational positions and containment, and estimating numerosity. However, it still remains a critical unresolved challenge for Multimodal Large Language Models (MLLMs). To fill this gap, we propose to treat Euclidean geometry problem-solving as a surrogate task. Specifically, we meticulously constructed a curated multimodal dataset, called Euclid30K, comprising approximately 30K plane and solid geometry problems. To enable the model to acquire and apply Euclidean principles from these geometry problems, we employed Group Relative Policy Optimization (GRPO) to finetune the Qwen2.5VL family and RoboBrain2.0 family, inspiring the models to identify shapes, count, and relate entities, and perform multi-step deductive reasoning using Euclidean principles. Our experiments demonstrate that the resulting models achieve substantial zero-shot gains across four spatial reasoning benchmarks (Super-CLEVR, Omni3DBench, VSI-Bench, and MindCube) without any task-specific adaptations. Notably, after training on the Euclid30K, the mean VSI-Bench accuracy of all evaluated models rose from 34.5% to 40.5%, improving by 5.5 percentage points. Among them, RoboBrain2.0-Euclid-7B achieves 49.6% accuracy, surpassing the previous state-of-the-art model, Spatial-MLLM. To our knowledge, this is the first systematic study showing that geometry-centric fine-tuning can confer vision-language models with broadly transferable spatial skills. Code and Euclid30K dataset can be found in [this](#).

1. Introduction

“The whole is greater than the part.”

— Euclid, *Elements I, Common Notion 5*

*These authors contributed equally

†Corresponding author

In recent years, multimodal large language models (MLLMs) [2, 41, 46] have achieved remarkable success across a broad range of vision-language tasks, from image captioning and visual question answering (VQA) to document understanding [4, 9, 28, 29, 37, 47]. State-of-the-art models like GPT-4o, Qwen2.5VL-series, and InternVL2.5-series now rival or even surpass human performance on certain benchmarks, especially tasks requiring advanced language understanding or mathematical reasoning [9, 17, 31]. For instance, model like GPT-4o [31], Claude-3.5-Sonnet [3], InternVL2-40B [17], and Qwen2.5VL-32B [9], have exceeded the average human score on the MathVista multimodal math benchmark [49], reflecting the rapid progress in integrating visual perception with high-level reasoning.

Despite recent progress, state-of-the-art MLLMs still fall short of genuine spatial intelligence [72, 78]. Spatial intelligence involves perceiving and mentally manipulating spatial relationships and spans several tasks, such as estimating quantity, interpreting spatial relations, and understanding geometric configurations [27, 72]. Nowadays, leading vision-language models (VLMs) still make occasional mistakes on tasks that young children solve with ease, such as determining object orientation or identifying which object is the nearest neighbor to a given object on its left [42, 53]. A recent evaluation on the Visual-Spatial Intelligence Benchmark (VSI-Bench) shows that more than 70% of the recorded errors arise from faulty spatial reasoning, not from deficiencies in visual recognition or language parsing [72]. This phenomenon is consistent with Moravec’s paradox [55], which suggests that high-level reasoning tasks are computationally simpler for VLM than low-level perceptual and sensorimotor skills. Closing this gap is essential for the next generation of VLMs [22].

Recent work on spatially aware VLMs, including Spatial-MLLM [69], SpaceVLM [13], RoboBrain [34], and RoboBrain2.0 [8], attempts to provide specially constructed spatial datasets to improve model performance. However, tasks in these spatial datasets usually cover only a subset of real-world spatial tasks and may not enhance the model’s overall spatial intelligence. For example, Spatial-MLLM

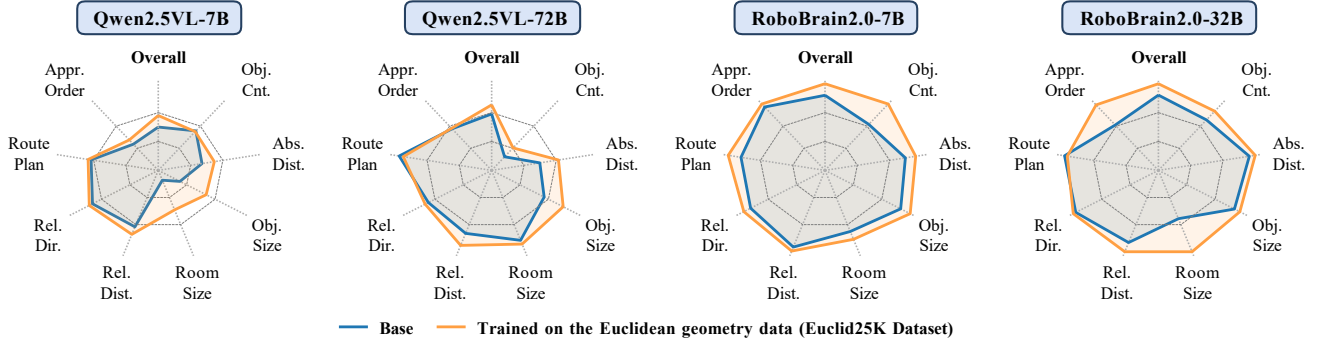


Figure 1. Performance gains on VSIBench after model training on Euclid30K, for more complete data please refer to Tab. 2.

collects data from ScanQA [7], SQA3D [51], and self-curated spatial QA data, and follows the eight tasks introduced in VSI-Bench [72] to build the Spatial-MLLM-120K dataset. The trained model therefore achieves state-of-the-art results on VSI-Bench, ScanQA, and SQA3D. However, its accuracy drops on the out-of-domain MindCube benchmark [74]. This highlights a critical challenge in the field: while fine-tuning on task-specific datasets can achieve high in-domain performance, it may lead to over-specialization and fail to cultivate a more fundamental, generalizable spatial intelligence. To bridge this gap, VLMs must learn from a broader and more foundational range of spatial phenomena, thus extending their capabilities beyond the limitations of any single dataset.

In order to develop generalized spatial skills beyond any single benchmark, in this work, we attempt to explore a novel training paradigm that incorporates solving geometric problems as a surrogate task for enhancing spatial intelligence in the VLM. Geometry compresses centuries of mathematical study into formal descriptions of space. Therefore, learning to solve planar and solid geometry problems forces the model to internalize the axioms and constraints of Euclidean geometry, and provides the model with stronger out-of-domain (OOD) generalization capabilities, because these principles are universal and independent of any single task. As shown in Fig. 1, these low-level geometry priors, like “Euclid’s Gift,” provide a principled foundation that supports zero-shot transfer to a wide range of downstream spatial tasks. This suggests that the abilities required to solve geometric problems, including recognizing shapes and configurations, inferring spatial relationships such as parallel, angular, and relative positions, calculating or measuring geometric elements, and performing multistep logical reasoning, are also required for spatial perception tasks, like object counting, relational localization, and size estimation [19, 72]. Thus, by using geometry as a surrogate task, we aim to instill foundational Euclidean priors that support a significant and critical subset of spatial intelligence, i.e., spatial perception and reasoning. [20, 57].

To operationalize geometry-centric training, we curated

and released Euclid30K, which contains 29,695 geometry VQAs collected from Geometry3K [48], MMK12 [54], and SolidGeo [67]. The answers to each question in Euclid30K were reformatted so that they could be directly recognized by the rule-based reward function in Group Relative Policy Optimization (GRPO) [62]. Euclid30K spans the full K–12 and Olympiad geometry syllabus, covering planar and solid reasoning, metric relations, and classical theorems.

Furthermore, in order to attribute the performance improvement strictly to the supervision of geometric knowledge rather than the interference of complex algorithms or data augmentation, we intentionally used the well-established GRPO framework and Euclid30K to train the models. Specifically, we used GRPO to fine-tune the Qwen2.5VL series (3B, 7B, 72B parameters) and the RoboBrain2.0 series (7B, 32B parameters). The resulting geometry-trained models deliver consistent gains on SuperCLEVR [42], Omni3D-Bench [53], VSI-Bench [72], and MindCube[74]. This suggests that the abstract geometric knowledge distilled from Euclid30K can be migrated to different spatial tasks and supports the model’s static spatial perception and reasoning capabilities, is an effective alternative for enhancing spatial intelligence in VLMs.

To our knowledge, this is the first work to use geometry problems as surrogate tasks to cultivate spatial intelligence in general-purpose VLMs. Unlike prior approaches that fine-tuned models on a single skill, our training paradigm uses geometric tasks to endow the model with a deeper, principle-driven understanding of space. This allows the model to go beyond the baseline in multiple spatial reasoning tasks without additional task-specific training. In summary, our contributions are as follows:

- We demonstrate that tackling geometry tasks can serve as an effective surrogate task for spatial intelligence. Learning to solve geometry problems assists models in acquiring the most basic perceptions of space, such as the axioms of Euclidean geometry. These low-level priors provide a principled foundation that supports the zero-shot transfer to a wide range of downstream spatial tasks.
- We collected and constructed the Euclid30K, a VQA

dataset of 29,695 geometry questions. Euclid30K provides a comprehensive range of varied and challenging geometry problems that cover a wide variety of geometric concepts to better help models learn from geometry problems formal descriptions of space.

- Extensive experiments on the Qwen2.5VL-series and RoboBrain2.0-series show significant performance gains on four spatial benchmarks (Super-CLEVR, Omni3D-Bench, VSI-Bench, and MindCube), providing empirical evidence that the geometric curriculum reliably enhances spatial reasoning across diverse evaluation settings.

2. Related Work

2.1. Spatial Intelligence

Spatial intelligence is the capacity to reason about spatial relations, including mental rotation, viewpoint switching, and object counting [27, 72]. Benchmarks such as CLEVR [35], Super-CLEVR [42], Omni3D-Bench [53], VSI-Bench [72], and MindCube [74] show that accurate visual recognition alone is not enough; models must also perform explicit spatial reasoning. Early VLMs—for example ViLT [36], METER [25], Flamingo [2], and PaLI [16]—improved perception yet still struggled with counting and orientation. Recent spatial-aware systems such as SpaceVLM [13], Spatial-MLLM [69], SpatialLLM [50], RoboBrain [34], and RoboBrain 2.0 [8] fine-tune on curated spatial datasets, but these datasets cover only part of real-world scenarios and are costly to build. Our approach instead trains on readily available Euclidean geometry problems, treating them as surrogate tasks that transfer their principles to general visual-spatial perception and reasoning without costly data collection or architectural changes.

2.2. Multimodal Math & Geometry Datasets

Research on multimodal mathematical reasoning has produced a range of datasets that progressively broaden both topic scope and problem format. GeoQA [14], Geometry3K [48], and UniGeo [15] pioneered the use of diagram-text pairs for plane-geometry question answering, laying the groundwork for visual-symbolic reasoning. Later collections such as MMMU [76], MathVista [49], We-Math [58], We-Math2 [59], MMMU-Pro [77], and GeoSense [71], expanded to a wider array of mathematical domains and introduced more challenging multimodal tasks. The recently released SolidGeo3K [67] benchmark focuses on three-dimensional geometry and offers fine-grained skill tags along with difficulty annotations, providing a dedicated testbed for solid-geometry reasoning. The GeoReasoning-10K [70] contains 10,000 pairs of carefully constructed geometric images and descriptions, ensuring perfect alignment between visual and textual information.

3. Method

This section first explores the potential mechanisms by which solving geometric problems enhances spatial intelligence through a perspective that integrates transfer learning with educational psychology. Then, we aim to enhance the geometric problem-solving capabilities of VLMs through reinforcement learning. Finally, we validate our hypothesis by verifying whether their spatial perception and reasoning skills improve through a series of benchmark tests.

3.1. Rationale

This subsection presents a domain-adaptation view explaining why training on geometric problem solving serves as an effective surrogate for spatial intelligence.

Let the source distribution \mathcal{D}_S denote geometry problem solving (e.g., Euclid30K) and the target distribution \mathcal{D}_T denote spatial intelligence. We train a VLM policy π_θ on the source. The policy induces a hypothesis:

$$h(x) = \mathbb{I}[\text{Ans}(x; \pi_\theta) \text{ is correct}] \in \{0, 1\}. \quad (1)$$

Let \mathcal{H} be the hypothesis class containing h . For $h, h' \in \mathcal{H}$, define the error (disagreement probabilities):

$$\epsilon_S(h, h') := \Pr_{x \sim \mathcal{D}_S} [h(x) \neq h'(x)],$$

Specifically, when f is an absolutely correct hypothesis, we abbreviate $\epsilon_S(h, f)$ as $\epsilon_S(h)$. The definition of $\epsilon_T(h, h')$ and $\epsilon_T(h)$ follows similarly. Furthermore, the $\mathcal{H}\Delta\mathcal{H}$ distance [11] between \mathcal{D}_S and \mathcal{D}_T is:

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h, h' \in \mathcal{H}} |\epsilon_S(h, h') - \epsilon_T(h, h')|. \quad (2)$$

Based on this definition, we can invoke the standard domain-adaptation bound [11], which connects the $\epsilon_T(h)$ to the $\epsilon_S(h)$ through the $\mathcal{H}\Delta\mathcal{H}$ distance.

Standard Domain-Adaptation Bound

For any $h \in \mathcal{H}$, we have:

$$\epsilon_T(h) \leq \epsilon_S(h) + \epsilon_{ideal} + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T), \quad (3)$$

where ϵ_{ideal} denotes the error of the ideal model and can be considered as an extremely tiny constant.

In light of the bound in Eq. (3), the target error is governed by the source error and the distribution discrepancy $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$. In practice, $\epsilon_S(h)$ can typically be reduced through optimization and data curation, which makes the discrepancy term pivotal [11, 52]. Consequently, if $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is small, we can regard the source distribution \mathcal{D}_S as a surrogate for the target distribution \mathcal{D}_T . A detailed proof of Eq. (3) can be found in Appendix A.

Statistic	Number
Total Number	29,695
Mathematical Expression	16,804
Numeric	6,321
Multiple-Choice	2,618
Type	
Plane (2D) / Solid (3D)	18,577 / 11,118
Newly collected	
Newly collected Solid Questions	3,996
Newly collected Images	3,792
Length	
Max and Avg question length	1,598 / 229.8
Max and Avg answer length	501 / 8.9
Max and Avg image numbers	8 / 1.1

Table 1. Statistics of Euclid30K, Mathematical Expression, Numeric and Multiple-Choice are the three types of answers.

We hypothesize that formal Euclidean geometry compresses a broad set of spatial regularities—congruence, similarity, perspective, parallelism, intersection, and positional reasoning—into theorem-like constraints reused across downstream spatial-intelligence tasks. Compared to training on a narrow sub-skill (e.g., object counting, depth ordering, or size estimation), this breadth plausibly yields a smaller $d_{\mathcal{H}\Delta\mathcal{H}}$ between geometry and target tasks, consistent with cognitive-science evidence on the generality of geometric knowledge in human perception and reasoning [26, 39]. We further elaborate on this connection from an educational-psychology perspective in Appendix B.

Since the spatial-intelligence target-domain distribution is unknown and there is no canonical proxy dataset $\hat{\mathcal{D}}_T$ that faithfully spans the full space of spatial-intelligence tasks, $d_{\mathcal{H}\Delta\mathcal{H}}$ is not directly observable in practice. Nevertheless, our subsequent empirical results align with this qualitative prediction: after Euclid30K fine-tuning, models exhibit consistent gains across Omni3D-Bench, VSI-Bench, Super-CLEVR, and MindCube (in Tables 2, 4, and 3), and qualitative cases in the Appendix (Figs. 7–12) further illustrate acquired, transferable skills such as perspective (near to far size), parallelism and similarity, and positional inference. These observations provide a principled rationale for geometry as a surrogate task for spatial intelligence.

3.2. Euclid30K

Plane and solid geometry give axiomatized abstractions of spatial phenomena. Training on such problems compels models to internalize Euclidean constraints such as angle and ratio preservation, similarity, and congruence, thereby providing an effective surrogate curriculum for cultivating spatial perception and reasoning skills.

Unfortunately, there are currently no large-scale, high-quality training datasets tailored for diverse geometric prob-

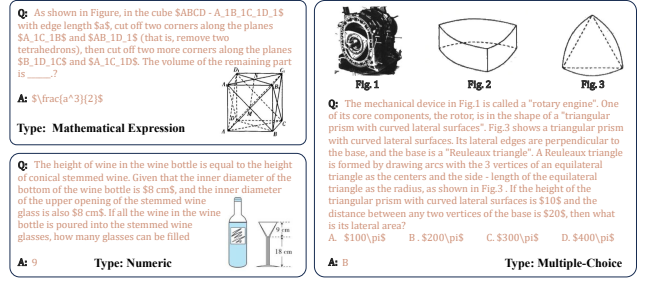


Figure 2. The examples of the newly collected questions in Euclid30K. More examples can be found in the appendix.

lems. To address this, we filtered the required corpus from open-source resources such as Geometry3K [48], MMK12 [54], SolidGeo [67], and WeMath2 [59]. We employed Qwen2.5-VL-72B as an annotator to determine whether a problem belongs to plane or solid geometry.

After filtering the existing corpus, we identified a significant imbalance: only about 7,000 solid geometry problems remained, while there were about 20,000 plane geometry problems. However, solid geometry encompasses more explicit three-dimensional spatial phenomena (e.g., perspective invariance, polyhedron truncation, volume-area relationships), which are equally crucial for VLMs in learning spatial knowledge. Furthermore, existing solid geometry problems predominantly focus on shape recognition, coordinate/angle/area calculations, and similar question types, with insufficient coverage of richer problem types. To address these gaps, we newly collected approximately 4,500 additional problems from commercially available K–12 textbooks and standard/competition practice books on junior/senior high school mathematics. These cover categories such as positional relationship determination, dynamic or moving-point problems, folding/unfolding problems, and contextualized geometry word problems.

As a result, we compiled a question pool containing approximately 32,500 candidate questions. To ensure the quality of the collected geometric data, we designed a three-stage filtering process:

- **Duplicate Filtering:** Since even similar texts can lead to vastly different meanings or solution processes when paired with different images, we uniquely identify each question through its image. Specifically, we use image-based perceptual hashing to filter duplicate questions.
- **Question Splitting:** Many materials contain multiple sub-questions under a single main question. We utilize the GPT-4o API [31] to detect and enumerate each sub-question, splitting them into independent instances.
- **Formula Formatting:** We standardize formulae in questions and answers to LaTeX format via the DeepSeek-V3.1 API [24]. This ensures answers can be correctly parsed by MATHVERIFY [32] for subsequent verification.

Methods	Numerical Question				Multiple-Choice Question				Overall
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
<i>Proprietary Models</i>									
GPT-4o [31]	46.2	5.3	43.8	38.2	37.0	41.3	31.5	28.5	34.0
Gemini-1.5 Pro [29]	49.6	28.8	58.6	49.4	46.0	48.1	42.0	68.0	48.8
Gemini-2.0 Flash [30]	52.4	30.6	66.7	31.8	56.0	46.3	24.5	55.1	45.4
Gemini-2.5-Flash-preview-04-17 [21]	-	-	-	-	-	-	-	-	48.8
Gemini-2.5-Pro-preview-05-06 [21]	-	-	-	-	-	-	-	-	47.8
<i>Open-source Models</i>									
LongVA-7B [80]	38.0	16.6	38.9	22.2	33.1	43.3	25.4	15.7	29.2
InternVL2-40B [18]	34.9	26.9	46.5	31.8	42.1	32.2	34.0	39.6	36.0
VILA-1.5-40B [43]	22.4	24.8	48.7	22.7	40.5	25.7	31.5	32.9	31.2
LLaVA-OneVision-72B [40]	43.5	23.9	57.6	37.5	42.5	39.9	32.5	44.6	40.2
LLaVA-Video-72B [81]	48.9	22.8	57.4	35.3	42.4	36.7	35.0	48.6	40.9
<i>Spatial Models</i>									
M2-Reasoning-7B [33]	41.0	34.0	60.9	55.4	40.7	47.3	29.9	28.8	42.3
Spatial-MLLM-4B [69]	65.3	34.8	63.1	45.1	41.3	46.2	33.5	46.3	48.4
<i>Qwen2.5VL-series</i>									
Qwen2.5VL-3B [9]	35.6	23.4	34.9	16.6	34.4	40.7	26.3	21.8	29.2
Qwen2.5VL-Euclid-3B	38.3 ↑	26.8 ↑	35.4 ↑	22.2 ↑	37.0 ↑	43.2 ↑	36.6 ↑	16.3	32.0 ↑
Qwen2.5VL-7B [9]	39.5	17.8	16.9	5.8	33.8	36.7	24.7	22.8	24.8
Qwen2.5VL-Euclid-7B	38.8	22.8 ↑	37.3 ↑	23.2 ↑	38.3 ↑	38.5 ↑	25.8 ↑	26.5 ↑	31.4 ↑
Qwen2.5VL-72B [9]	13.6	19.6	40.9	41.1	37.7	35.3	34.0	36.2	32.3
Qwen2.5VL-Euclid-72B	22.5 ↑	27.2 ↑	55.7 ↑	43.3 ↑	44.9 ↑	37.1 ↑	32.5	36.6 ↑	37.5 ↑
<i>RoboBrain2.0-7B-series</i>									
RoboBrain2.0-7B [8]	46.0	32.7	58.9	35.9	45.9	41.5	30.9	55.2	43.0
RoboBrain2.0-Euclid-7B	66.4 ↑	36.9 ↑	66.3 ↑	40.5 ↑	48.3 ↑	45.3 ↑	35.6 ↑	57.8 ↑	49.6 ↑
RoboBrain2.0-32B [8]	50.5	37.0	59.2	28.4	43.2	46.1	34.5	39.5	43.1
RoboBrain2.0-Euclid-32B	59.2 ↑	39.4 ↑	63.4 ↑	47.8 ↑	48.7 ↑	47.5 ↑	33.5	57.0 ↑	49.6 ↑

Table 2. **Evaluation Results on VSI-Bench [72].** The performance of Gemini-2.5 is reported from RoboBrain2.0 [8], and the performance of Spatial-MLLM-4B [69] and M2-Reasoning-7B [33] is reported from its original paper, while the results for the other Baseline, Proprietary Models, and Open-source Models are taken from the VSI-Bench benchmark [72]. Qwen2.5VL-Euclid and RoboBrain2.0-Euclid indicate the Qwen2.5VL [9] and RoboBrain2.0 [8] trained with GRPO [62] on the Euclid30K dataset.

practical relevance to the geometric conclusions, while allowing for rounding or floating-point truncation errors that are common in the generation process. In addition, for multiple choice questions, we use the exact match reward.

4. Experiment

In this section, we evaluate the zero-shot generalisation of models trained on geometry data to spatial intelligence tasks, using four benchmarks: Super CLEVR, Omni3D Bench, VSI Bench, and MindCube. We also conduct a causal ablation study that contrasts models trained on equal amounts of geometry data and spatial data. Additional experiments, along with detailed settings, prompt templates, and dataset settings, are provided in the Appendix C.

4.1. Comparisons on VSI-Bench

Tab. 2 shows a uniform rise in overall accuracy after Euclid30K fine-tuning: every checkpoint in the Qwen2.5VL and RoboBrain2.0 families increases its VSI-Bench over-

all score, with gains of 2–7% points relative to the corresponding vanilla models. The effect is most pronounced on RoboBrain2.0, where the Euclid-trained 7B and 32B versions reach 49.6% overall, outstripping the best open-source reference (Spatial-MLLM-4B at 48.4%) and also surpassing the strongest proprietary baselines reported by VSI-Bench (Gemini-1.5 Pro and Gemini-2.5-Flash-preview-04-17, both at 48.8%). We note occasional drops in performance on the subtasks like Appearance Order and Route Planning; these tasks emphasize temporal sequencing and planning rather than the static Euclidean priors that Euclid30K provides. Therefore, a slight decrease in performance is acceptable and does not offset overall performance on other tasks.

4.2. Comparisons on SuperClevr & Omni3DBench

Tab. 3 shows that Euclid30K fine-tuning consistently raises accuracy. Qwen2.5VL-3B climbs from 70.0% to 75.2% on Super-CLEVR (5.2% higher) and from 24.7% to 26.5% on Omni3D-Bench (1.8% higher). The 7B and 72B check-

Methods	SuperClevr	Omni3DBench
<i>Qwen2.5VL-series</i>		
Qwen2.5VL-3B	70.0	24.7
Qwen2.5VL-Euclid-3B	75.2 ↑	26.5 ↑
Qwen2.5VL-7B	76.1	28.3
Qwen2.5VL-Euclid-7B	86.2 ↑	31.1 ↑
Qwen2.5VL-72B	72.6	30.4
Qwen2.5VL-Euclid-72B	83.1 ↑	32.9 ↑
<i>RoboBrain2.0-series</i>		
RoboBrain2.0-7B	47.4	14.2
RoboBrain2.0-Euclid-7B	85.2 ↑	21.2 ↑
RoboBrain2.0-32B	59.5	34.8
RoboBrain2.0-Euclid-32B	71.0 ↑	36.2 ↑

Table 3. **Evaluation Results on SuperClevr [42] and Omni3DBench [53].** Qwen2.5VL-Euclid and RoboBrain2.0-Euclid indicate the Qwen2.5VL [9] and RoboBrain2.0 [8] trained with GRPO [62] on the Euclid30K dataset.

points gain about 10% absolute accuracy on Super-CLEVR and roughly 2–3% on Omni3D-Bench. Improvements are even larger for RoboBrain2.0, the 7B variant leaps from 47.4% to 85.2% on Super-CLEVR (37.8% higher) and from 14.2% to 21.2% on Omni3D-Bench (7.0% higher), while the 32B version adds 11.5% and 1.4% on the two datasets. Overall, Euclid30K delivers substantial accuracy gains, demonstrating that the geometry-centric curriculum provides broadly transferable spatial priors rather than mere data volume benefits. The unusually large boost observed for RoboBrain2.0 after Euclid30K fine-tuning likely stems from over-specialisation in its original training data: the base model had memorised patterns tailored to its own spatial QA corpus, leaving limited capacity to generalise. Exposure to Euclid30K re-anchors the network in precise Euclidean principles, and forces it to reason deductively across a much broader range of geometric configurations. In effect, the geometry curriculum supplies universal spatial priors that counteract earlier overfitting and restore the model’s ability to extrapolate to unseen spatial tasks.

4.3. Comparisons on MindCube

As shown in Tab. 4, fine-tuning on Euclid30K improves the overall MindCube accuracy of every variant in both model families. It is also worth noting that models trained using Euclidean geometry datasets outperform existing spatial models (most of which are also based on Qwen backbones, but trained on a larger spatial corpus) in terms of generalization ability. For example, Spatial-MLLM trained the Qwen 2.5-VL-3B backbone on a 120K spatial dataset and scored an overall score of 32.1% on MindCube. In contrast, Qwen2.5VL-Euclid-3B, trained with only 30k Euclidean geometric features, scored 38.9% on MindCube, repre-

Methods	Rotation	Among	Around	Overall
<i>Proprietary Models</i>				
GPT-4o [31]	32.7	40.2	29.2	38.8
Claude-4-Sonnet [4]	48.4	44.2	47.6	44.8
<i>Spatial Models</i>				
RoboBrain1.0-7B [34]	35.8	38.3	29.5	37.4
SpaceMantis [13]	37.7	21.3	29.3	22.8
Space-Qwen [13]	38.0	33.7	26.3	33.3
Spatial-MLLM [69]	38.4	20.9	32.8	32.1
<i>Qwen2.5VL-series</i>				
Qwen2.5VL-3B [9]	14.3	22.8	24.1	20.4
Qwen2.5VL-Euclid-3B	33.5 ↑	43.0 ↑	40.0 ↑	38.9 ↑
Qwen2.5VL-7B [9]	35.6	25.9	28.4	30.0
Qwen2.5VL-Euclid-7B	34.2	28.7 ↑	30.4 ↑	31.1 ↑
Qwen2.5VL-72B [9]	31.5	31.4	30.6	31.2
Qwen2.5VL-Euclid-72B	43.0 ↑	35.6 ↑	31.6 ↑	36.7 ↑
<i>RoboBrain2.0-series</i>				
RoboBrain2.0-7B [8]	39.4	38.8	38.6	38.9
RoboBrain2.0-Euclid-7B	36.0	46.5 ↑	36.2	39.4 ↑
RoboBrain2.0-32B [8]	21.2	35.4	31.0	29.2
RoboBrain2.0-Euclid-32B	39.0 ↑	42.4 ↑	34.8 ↑	38.8 ↑

Table 4. **Evaluation Results on MindCube [74].** The performance of Proprietary Models and Spatial Models are taken from the MindCube benchmark [74]. Qwen2.5VL-Euclid and RoboBrain2.0-Euclid indicate the Qwen2.5VL [9] and RoboBrain2.0 [8] trained with GRPO [62] on the Euclid30K dataset.

sending a relative improvement of 6.8 percentage points. These results suggest that learning accurate Euclidean priors from a compact geometry course provides more transferable spatial knowledge than extending generalized spatial data alone.

4.4. Ablation Study

To isolate the contribution of our Euclid30K dataset from the potential reasoning enhancements provided by the GRPO algorithm, we conducted a causal ablation study. Specifically, we randomly sampled a subset equal in size to Euclid30K on the non-geometric spatial intelligence dataset Clevr-CoGenT [35] and used the exact identical GRPO setup to train Qwen2.5VL and RoboBrain2.0. This design ensures that performance gains after training on geometric data can be directly attributed to the fact that the geometric task as surrogate task for spatial intelligence, rather than due to the effects of GRPO or data incrementation.

Tab. 5 shows that models trained on Euclid30K achieve markedly higher overall accuracy than those fine-tuned on the equal-sized Clevr-CoGenT split. Because Clevr-CoGenT mainly targets object counting and positional-relation queries, the Clevr-CoGenT variants re-

Methods	Numerical Question				Multiple-Choice Question				Overall
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
<i>Qwen2.5VL-series</i>									
Qwen2.5VL-3B	35.6	23.4	34.9	16.6	34.4	40.7	26.3	21.8	29.2
Qwen2.5VL-ClevrCoGenT-3B	40.5	24.5	30.1	29.8	33.9	43.0	29.9	18.8	31.3
Qwen2.5VL-Euclid-3B	38.3	26.8	35.4	22.2	37.0	43.2	36.6	16.3	32.0
Qwen2.5VL-7B	39.5	17.8	16.9	5.8	33.8	36.7	24.7	22.8	24.8
Qwen2.5VL-ClevrCoGenT-7B	42.4	17.8	24.5	9.2	36.7	38.5	29.4	23.8	27.8
Qwen2.5VL-Euclid-7B	38.8	22.8	37.3	23.2	38.3	38.5	25.8	26.5	31.4
Qwen2.5VL-72B	13.6	19.6	40.9	41.1	37.7	35.3	34.0	36.2	32.3
Qwen2.5VL-ClevrCoGenT-72B	15.6	24.8	40.7	41.4	43.4	37.8	29.4	33.5	33.2
Qwen2.5VL-Euclid-72B	22.5	27.2	55.7	43.3	44.9	37.1	32.5	36.6	37.5
<i>RoboBrain2.0-7B-series</i>									
RoboBrain2.0-7B	46.0	32.7	58.9	35.9	45.9	41.5	30.9	55.2	43.0
RoboBrain2.0-ClevrCoGenT-7B	66.4	34.4	65.8	41.0	46.6	46.5	36.6	52.3	48.7
RoboBrain2.0-Euclid-7B	66.4	36.9	66.3	40.5	48.3	45.3	35.6	57.8	49.6
RoboBrain2.0-32B	50.5	37.0	59.2	28.4	43.2	46.1	34.5	39.5	43.1
RoboBrain2.0-ClevrCoGenT-32B	58.0	36.9	62.2	47.8	46.9	44.5	34.0	42.1	46.7
RoboBrain2.0-Euclid-32B	59.2	39.4	63.4	47.8	48.7	47.5	33.5	57.0	49.6

Table 5. **Ablation experiment on VSI-Bench [72].** We compare training a model on a 25K subset of the spatial intelligence dataset Clevr-CoGenT v.s. the geometric dataset Euclid30K to verify that the geometric dataset serves as a surrogate task to improve the spatial intelligence capabilities of the model. **Bolding** indicates the best score within each model type.

tain a slight edge in object counting and relative-direction tasks. In contrast, the Euclid30K variant has better generalization performance, improving in almost all task categories, resulting in a one- to two-percentage-point improvement in the overall metric. This result is strong evidence that the performance improvement stems primarily from the unique, principled knowledge embedded in Euclidean geometry problems, rather than just the generic reasoning gains from the RL training process.

In addition, all Qwen2.5VL variants show a slight decline on the Appearance Order task after training on either Clevr-CoGenT or Euclid30K, which we attribute to the fact that Appearance Order primarily probes temporal reasoning and memory skills, and not directly reinforced by either corpus. By contrast, the RoboBrain2.0 variants were exposed to similar temporal tasks during their original pre-training; once Euclid30K fine-tuning alleviates their overfitting to the original data distribution, their temporal reasoning rebounds, and RoboBrain2.0-Euclid-32B, in particular, registers a marked improvement in Appearance Order.

5. Limitations

Although Euclidean geometry tasks can serve as a surrogate task for spatial intelligence and supply useful Euclidean priors that boost spatial reasoning in many situations, these priors do not always translate into higher scores on every task. For example, the Appearance Order task in VSI-Bench de-

pends on temporal memory, whereas static geometry data does not encode temporal information, so accuracy on this task can remain flat or decline. Rotation tasks in MindCube may show limited benefit because plane-geometry problems rarely require reasoning about object rotation, which offsets the expected gains. More detailed analysis about the limitations can be found in Appendix E.

6. Conclusion

This study shows that using Euclidean geometry as a surrogate task provides an alternative way to achieve transferable Spatial Perception and Reasoning. MLLMs are trained on planar and solid geometry problems with basic Euclidean priors that can be transferred to various spatial benchmarks without additional fine-tuning. After training on our proposed Euclid30K alone, our model achieves consistent and significant performance gains on four different, unseen benchmarks. This increase in generalization validates our core hypothesis that learning basic Euclidean geometric principles is a more effective strategy for developing transferable spatial skills. More importantly, by identifying the boundaries of this transferability, our work provides a direction for future research: namely, to provide the ability to combine geometry lessons with data (e.g., video) containing rich temporal information to achieve more comprehensive and balanced spatial intelligence.

Future Work

We are closely monitoring the updates of Qwen3VL [65]. After the release of Qwen3-VL-30B-A3B-Instruct and Qwen3-VL-4B-Instruct, we will further update our results.

References

- [1] David Acuna, Guojun Zhang, Marc T. Law, and Sanja Fidler. f-domain adversarial learning: Theory and algorithms. In *Proceedings of the 38th International Conference on Machine Learning*, pages 66–75. PMLR, 2021. 13
- [2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 1, 3
- [3] Anthropic. Claude 3.5 sonnet. Blog, 2024. Accessed: November 22, 2024. 1
- [4] Anthropic. Claude 4 sonnet system card, 2025. Version 20250514, accessed 2025-06-23. 1, 7
- [5] Laszlo Aszalos and Maria Bako. How can we improve the spatial intelligence. In *6th International Conference on Applied Informatics*, Eger, Hungary, 2004. 13
- [6] Jafar A Aziz, Dwi Juniati, and Pradnyo Wijayanti. Students’ reasoning with logical mathematical and visual spatial intelligence in geometry problem solving. In *International Joint Conference on Science and Engineering (IJCE 2020)*, pages 203–207. Atlantis Press, 2020. 13
- [7] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2
- [8] BAAI-RoboBrain-Team. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 1, 3, 6, 7, 16, 19, 20, 21, 22
- [9] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025. 1, 6, 7, 16, 17, 18
- [10] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARK-itscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 14
- [11] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79(1-2):151–175, 2010. 3, 13
- [12] Jeffrey Buckley, Niall Seery, and Donal Canty. Investigating the use of spatial reasoning strategies in geometric problem solving. *International Journal of Technology and Design Education*, 29(2):341–362, 2019. 13
- [13] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brian Ichter, Danny Driess, Pete Florence, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. *arXiv preprint arXiv:2401.12168*, 2024. 1, 3, 7
- [14] Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, 2021. 3
- [15] Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, 2022. 3
- [16] Xi Chen, Xiao Wang, Soravit Changpinyo, Anthony J Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 3
- [17] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei 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*, pages 24185–24198, 2024. 1
- [18] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2025. 6
- [19] Sijie Cheng, Zhicheng Guo, Jingwen Wu, Kechen Fang, Peng Li, Huaping Liu, and Yang Liu. Egothink: Evaluating first-person perspective thinking capability of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14291–14302, 2024. 2
- [20] Douglas H Clements and Michael T Battista. Geometry and spatial reasoning. *Handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics*, pages 420–464, 1992. 2
- [21] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Naveen Sachdeva, Inderjit Dhillon, Marcel Blstein, 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. 6, 15

- [22] Yasser Dahou, Ngoc Dung Huynh, Phuc H Le-Khac, Wamiq Reyaz Para, Ankit Singh, and Sanath Narayan. Vision-language models can't see the obvious. *arXiv preprint arXiv:2507.04741*, 2025. 1
- [23] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2432–2443, 2017. 14
- [24] DeepSeek-AI. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 4
- [25] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18166–18176, 2022. 3
- [26] Jacob Feldman. What is a visual object? *Trends in Cognitive Sciences*, 7(6):252–256, 2003. 4, 13
- [27] Howard Gardner. *Frames of mind: The theory of multiple intelligences*. Basic books, 2011. 1, 3, 15
- [28] GemmaTeam, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 1
- [29] Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 1, 6
- [30] Google. Gemini 2.0 flash: Model card. <https://storage.googleapis.com/model-cards/documents/gemini-2-flash.pdf>, 2025. Model card published April 15, 2025. 6
- [31] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 4, 6, 7
- [32] Křtylčik Hynek and Ganderberger Greg. Math-verify. <https://github.com/huggingface/Math-Verify>, 2025. 4, 5
- [33] Inclusion-AI. M2-reasoning: Empowering mllms with unified general and spatial reasoning, 2025. 6, 16
- [34] 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. *arXiv preprint arXiv:2502.21257*, 2025. 1, 3, 7
- [35] 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. 3, 7
- [36] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 3
- [37] KimiTeam, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chen-zhuang Du, Chu Wei, Congcong Wang, Dehao Zhang, Dikang Du, Dongliang Wang, Enming Yuan, Enzhe Lu, Fang Li, Flood Sung, Guangda Wei, Guokun Lai, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haoning Wu, Hao-tian Yao, Haoyu Lu, Heng Wang, Hongcheng Gao, Huabin Zheng, Jiaming Li, Jianlin Su, Jianzhou Wang, Jiaqi Deng, Jiezhong Qiu, Jin Xie, Jinhong Wang, Jingyuan Liu, Junjie Yan, Kun Ouyang, Liang Chen, Lin Sui, Longhui Yu, Mengfan Dong, Mengnan Dong, Nuo Xu, Pengyu Cheng, Qizheng Gu, Runjie Zhou, Shaowei Liu, Sihao Cao, Tao Yu, Tianhui Song, Tongtong Bai, Wei Song, Weiran He, Weixiao Huang, Weixin Xu, Xiaokun Yuan, Xingcheng Yao, Xingzhe Wu, Xinxing Zu, Xinyu Zhou, Xinyuan Wang, Y. Charles, Yan Zhong, Yang Li, Yangyang Hu, Yanru Chen, Yejie Wang, Yibo Liu, Yibo Miao, Yidao Qin, Yimin Chen, Yiping Bao, Yiqin Wang, Yongsheng Kang, Yuanxin Liu, Yulun Du, Yuxin Wu, Yuzhi Wang, Yuzi Yan, Zaida Zhou, Zhaowei Li, Zhejun Jiang, Zheng Zhang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Zijia Zhao, and Ziwei Chen. Kimi-VL technical report, 2025. 1
- [38] K. M. Kyaw and T. Vidákovich. The relationship between spatial reasoning and geometric reasoning in teachers. *Eurasia Journal of Mathematics, Science and Technology Education*, 21(8):em2684, 2025. 13
- [39] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. 4, 13
- [40] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6
- [41] 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
- [42] Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023. 1, 2, 3, 7, 14, 15, 21, 22
- [43] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26679–26689, 2023. 6
- [44] Dingkun Liu, Zhu Chen, Jingwei Luo, Shijie Lian, and Dongrui Wu. Mirepnet: A pipeline and foundation model for eeg-based motor imagery classification. *arXiv preprint arXiv:2507.20254*, 2025. 13

- [45] Dingkun Liu, Siyang Li, Ziwei Wang, Wei Li, and Dongrui Wu. Spatial distillation based distribution alignment (sdda) for cross-headset eeg classification. *arXiv preprint arXiv:2503.05349*, 2025. 13
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 1
- [47] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 1
- [48] Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)*, 2021. 2, 3, 4
- [49] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, 2024. 1, 3
- [50] Wufei Ma, Luoxin Ye, Celso M de Melo, Alan Yuille, and Jieneng Chen. Spatialllm: A compound 3d-informed design towards spatially-intelligent large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17249–17260, 2025. 3
- [51] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations*, 2023. 2
- [52] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009. 3, 13
- [53] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api, 2025. 1, 2, 3, 7, 14, 15, 17, 18
- [54] Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, et al. Mm-eureka: Exploring visual aha moment with rule-based large-scale reinforcement learning. *arXiv preprint arXiv:2503.07365*, 2025. 2, 4
- [55] Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988. 1
- [56] Nora S Newcombe and Andrea Frick. Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, 4(3):102–111, 2010. 13
- [57] Marios Pittalis and Constantinos Christou. Types of reasoning in 3d geometry thinking and their relation with spatial ability. *Educational Studies in mathematics*, 75(2):191–212, 2010. 2
- [58] Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *arXiv preprint arXiv:2407.01284*, 2024. 3
- [59] Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. *arXiv preprint arXiv:2508.10433*, 2025. 3, 4
- [60] Nova Riastuti, Mardiyana, and Ikrar Pramudya. Analysis of students geometry skills viewed from spatial intelligence. In *AIP Conference Proceedings*, page 020024. AIP Publishing LLC, 2017. 13
- [61] N Riastuti, M Mardiyana, and I Pramudya. Students’ errors in geometry viewed from spatial intelligence. In *Journal of Physics: Conference Series*, page 012029. IOP Publishing, 2017. 13
- [62] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 5, 6, 7, 16
- [63] Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297, 2025. 13
- [64] Neşe Dokumacı Sütçü and Behçet Oral. The effects of geometrical-mechanical intelligence games on the spatial abilities. *International Online Journal of Primary Education*, 9(2):171–196, 2020. 13
- [65] Qwen Team. Easyrl: An efficient, scalable, multi-modality rl training framework. <https://github.com/QwenLM/Qwen3-VL>, 2025. 9
- [66] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggg: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 15
- [67] Peijie Wang, Chao Yang, Zhong-Zhi Li, Fei Yin, Dekang Ran, Mi Tian, Zhilong Ji, Jinfeng Bai, and Cheng-Lin Liu. Solidgeo: Measuring multimodal spatial math reasoning in solid geometry, 2025. 2, 3, 4, 5
- [68] Lilian Weng. Reward hacking in reinforcement learning. "<https://lilianweng.github.io/posts/2024-11-28-reward-hacking/>", 2024. 5
- [69] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, and Yueqi Duan. Spatial-mlm: Boosting mllm capabilities in visual-based spatial intelligence. *arXiv preprint arXiv:2505.23747*, 2025. 1, 3, 5, 6, 7, 15, 16
- [70] Yue Xin, Wenyuan Wang, Rui Pan, Ruida Wang, Howard Meng, Shizhe Diao, Renjie Pi, and Tong Zhang. Generalizable geometric image caption synthesis. *arXiv preprint arXiv:2509.15217*, 2025. 3
- [71] Liangyu Xu, Yingxiu Zhao, Jingyun Wang, Yingyao Wang, Bu Pi, Chen Wang, Mingliang Zhang, Jihao Gu, Xiang Li, Xiaoyong Zhu, et al. Geosense: Evaluating identification and application of geometric principles in multimodal reasoning. *arXiv preprint arXiv:2504.12597*, 2025. 3

- [72] Jihan Yang, Shusheng Yang, Anjali W. Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10632–10643, 2025. [1](#), [2](#), [3](#), [5](#), [6](#), [8](#), [14](#), [15](#), [16](#), [19](#), [20](#)
- [73] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12–22, 2023. [14](#)
- [74] Baiqiao Yin, Qineng Wang, Pingyue Zhang, Jianshu Zhang, Kangrui Wang, Zihan Wang, Jieyu Zhang, Keshigeyan Chandrasegaran, Han Liu, Ranjay Krishna, Saining Xie, Manling Li, Jiajun Wu, and Li Fei-Fei. Spatial mental modeling from limited views, 2025. [2](#), [3](#), [7](#), [14](#), [15](#), [16](#)
- [75] Qiyang Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gao-hong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. [5](#)
- [76] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruofei Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhui Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. [3](#)
- [77] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhui Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of The Association for Computational Linguistics*, 2025. [3](#)
- [78] Huanyu Zhang, Chengzu Li, Wenshan Wu, Shaoguang Mao, Ivan Vulić, Zhang Zhang, Liang Wang, Tieniu Tan, Furu Wei, et al. A call for new recipes to enhance spatial reasoning in mllms. *arXiv e-prints*, pages arXiv–2504, 2025. [1](#)
- [79] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024. [14](#)
- [80] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *ArXiv*, abs/2406.16852, 2024. [6](#)
- [81] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *ArXiv*, abs/2410.02713, 2024. [6](#)
- [82] Yaowei Zheng, Juntong Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyrl: An efficient, scalable, multi-modality rl training framework. <https://github.com/hiyouga/EasyRl>, 2025. [13](#)

Euclid’s Gift: Enhancing Spatial Perception and Reasoning in Vision-Language Models via Geometric Surrogate Tasks

Supplementary Material

A. Proof of Standard Domain-Adaptation Bound

This section provides the complete proof for the standard domain-adaptation bound introduced in Eq. (3). We follow the same notation as in the main text.

Let $h^* \in \mathcal{H}$ be a reference (ideal) hypothesis induced by a near-optimal policy. There is clearly that:

$$|\epsilon_S(h, h^*) - \epsilon_T(h, h^*)| \leq \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T). \quad (7)$$

By the triangle inequality and Eq. (7), we have:

$$\begin{aligned} \epsilon_T(h) &\leq \epsilon_T(h^*) + \epsilon_T(h, h^*) \\ &= \epsilon_T(h^*) + \epsilon_S(h, h^*) + (\epsilon_T(h, h^*) - \epsilon_S(h, h^*)) \\ &\leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + |\epsilon_T(h, h^*) - \epsilon_S(h, h^*)| \\ &\leq \epsilon_T(h^*) + [\epsilon_S(h) + \epsilon_S(h^*)] + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\ &= \epsilon_S(h) + \epsilon_{ideal} + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T), \end{aligned} \quad (8)$$

where $\epsilon_{ideal} = \epsilon_S(h^*) + \epsilon_T(h^*)$ can be considered as an extremely tiny constant (e.g., the error rate of an AGI model on these two datasets). The above proof refers to [11] and [52]. Therefore, if $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is sufficiently small, the population gap $\epsilon_T(h) - \hat{\epsilon}_S(h)$ also becomes small, which allows the source distribution to serve as a reliable surrogate for the target distribution [1, 44, 45].

B. Evidence from Educational Psychology

Complementing the domain-adaptation view in the previous subsection, we now present evidence from educational psychology that echoes the cognitive-science perspective on the generality of geometric knowledge in perception and reasoning [26, 39].

There is extensive evidence in educational psychology that geometry problem solving is closely related to spatial intelligence, can serve as an informative indicator of spatial ability, and can be used to improve it through targeted practice.

First, numerous correlational studies document a substantive link between geometric and spatial reasoning. Kyaw and Vidákovich report a moderate positive correlation between teachers’ geometric and spatial reasoning ($r = 0.47$), with 3D matching and measurement tasks predicting spatial scores [38]. In STEM and graphical education, higher spatial ability is associated with better problem-solving performance and more effective strategies

[12]. Newcombe and Frick emphasize that spatial representations and transformations are central cognitive resources that support reasoning in domains that are not obviously spatial—for example, through the use of graphs and diagrams [56].

Second, several studies show that performance on geometry tasks is a sensitive proxy for spatial ability. Analyses of middle-school students reveal that geometry skills and error patterns systematically vary with spatial-intelligence levels [60, 61]. Differences in dominance between logical-mathematical and visual-spatial intelligence yield distinct pathways for geometric reasoning, further tying geometry problem solving to spatial constructs [6]. These results support the use of geometry assessments as indicators of students’ spatial proficiency.

Third, intervention studies demonstrate that providing structured geometric activities can improve spatial intelligence. Programmatic practice with polyhedra and computer-generated spatial problems yields measurable gains [5]. Geometrical-mechanical intelligence games, implemented in quasi-experiments with pre/post testing, significantly enhance spatial visualisation and spatial relations skills [64]. Overall, the balance of evidence indicates that well-designed geometric practice is an effective means to cultivate spatial abilities.

Taken together, these findings motivate our surrogate-task choice. Our results suggest that the same relationship generalises beyond human learners to large multimodal models: training on formal geometry induces domain-invariant structure that transfers to diverse spatial-intelligence benchmarks. This observation is consistent with the domain-adaptation analysis in the previous subsection and provides an educational-psychology rationale for our geometry-first curriculum.

C. Detailed Experimental Setup

This section summarises the key hyperparameters, evaluation settings, prompt templates, and datasets settings used throughout the paper.

C.1. Training setup

In this paper, we follow the default settings of VeRL [63] and EasyR1 [82] to train the Qwen2.5-VL series and the RoboBrain2.0 series. Specifically, we train for 10 epochs in 64 NVIDIA H100 GPUs using Adam optimizer with a learning rate of 1×10^{-6} and a weight decay of 1×10^{-2} . In GRPO, we perform 8 rollouts per question and set the

default sampling temperature to 1. The KL divergence coefficient β in Eq. 5 is set to 1×10^{-2} .

Unless stated otherwise, we fix the random seed at 1 to guarantee determinism. We adopt a context window of 2048 tokens for both the prompt and the response, and use a rollout batch of 512 samples. The actor network updates with a global batch size of 128 and a maximum gradient norm of 1.0. Images are resized so that the total pixel count lies between 512×512 and 2048×2048 . All remaining hyper-parameters, including PPO clip ratio, learning-rate schedule, and parallelism settings, follow the default EasyR1 recipe and can be found in the supplied supplementary material.

C.2. Test setup

Inference is conducted with the Imms-eval toolkit [79] to ensure consistent decoding across models. In the test, to ensure the reproducibility of the results, we follow VSIBench [72] and MindCube [74] to set the temperature to 0. Finally, to ensure that the model performs sufficient spatial inference, we set the maximum generation length of model responses at 1024 tokens.

C.3. Prompt templates

Euclid-tuned models During both training and evaluation, we use the following template:

Euclid-tuned Models Prompt Template

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think>` `</think>` tags. The final answer MUST BE put in `\boxed{\}`.

Baseline variants RoboBrain2.0 expects the answer inside `<answer>` `</answer>` tags; we therefore replace the last line with, like:

Vanilla RoboBrain2.0 Prompt Template

You FIRST think about the reasoning process as an internal monologue and then provide the final answer. The reasoning process MUST BE enclosed within `<think>` `</think>` tags. The final answer MUST BE put in `<answer>` `</answer>`.

Because Qwen2.5VL-Instruct was tuned with supervised instruction data that often begins with phrases like “think step by step,” keeping the same cue in your evaluation prompt aligns the test-time input with the style encountered during training. This consistency helps the model interpret

the prompt as intended and reduces the risk of unexpected formatting effects.

Vanilla Qwen2.5VL-Instruct Prompt Template

You FIRST think step by step and then provide the final answer. The final answer MUST BE put in `\boxed{\}`.

When processing VSI-Bench tasks, we make minor adjustments to the above templates to follow the benchmark’s original settings as closely as possible and ensure consistent results—for example, we prepend “These are frames of a video.” to every prompt. In addition, all exact prompt templates are provided in the `lmms_eval/task` folder of the supplemental code.

C.4. Dataset Setup

In this subsection, we provide an introduction and configuration details for the dataset used in the main page.

Setup in VSI-Bench. VSI-Bench [72] contains more than 5,130 egocentric videos question-answer pairs sourced from ARKitScenes[10], ScanNet [23], and ScanNet++ [73]. The task types are divided into numerical question tasks (e.g. object counting, absolute distance estimation, object size estimation, and room size estimation) and multiple choice tasks (e.g. relative distance estimation, relative direction reasoning, route planning, and spatiotemporal appearance-order). For the evaluation metrics, we align with the VSIBench setting. In addition, for the Qwen2.5VL-series and RoboBrain2.0-series, we use 32 frames uniformly sampled from the scene video as input frames in the inference process.

Setup in Super-CLEVR and Omni3D-Bench. Super-CLEVR [42] contains a 5,000-image test split that probes how well a model handles changes in visual complexity, concept distribution, and composition, making it a strong measure of two-dimensional spatial reasoning. Omni3D-Bench [53] adds 500 questions to the Omni3D dataset, each requiring a model to locate objects in three-dimensional space and estimate their relative distances and sizes. Together, these benchmarks test both planar and volumetric aspects of spatial understanding, providing complementary evidence of a model’s geometric competence. For the evaluation metrics, we follow the settings of VSIBench [72]. Specifically, we calculate mean relative accuracy (MRA) across confidence thresholds $\mathcal{C} = \{0.5, 0.55 \dots, 0.95\}$ for the numerical question tasks and report exact-match accuracy for multiple-choice tasks.

Setup in MindCube. MindCube [74] is a recent benchmark crafted to scrutinize the spatial-reasoning capabilities of VLMs under partial observability and dynamic viewpoints, challenging the VLM to maintain object consistency

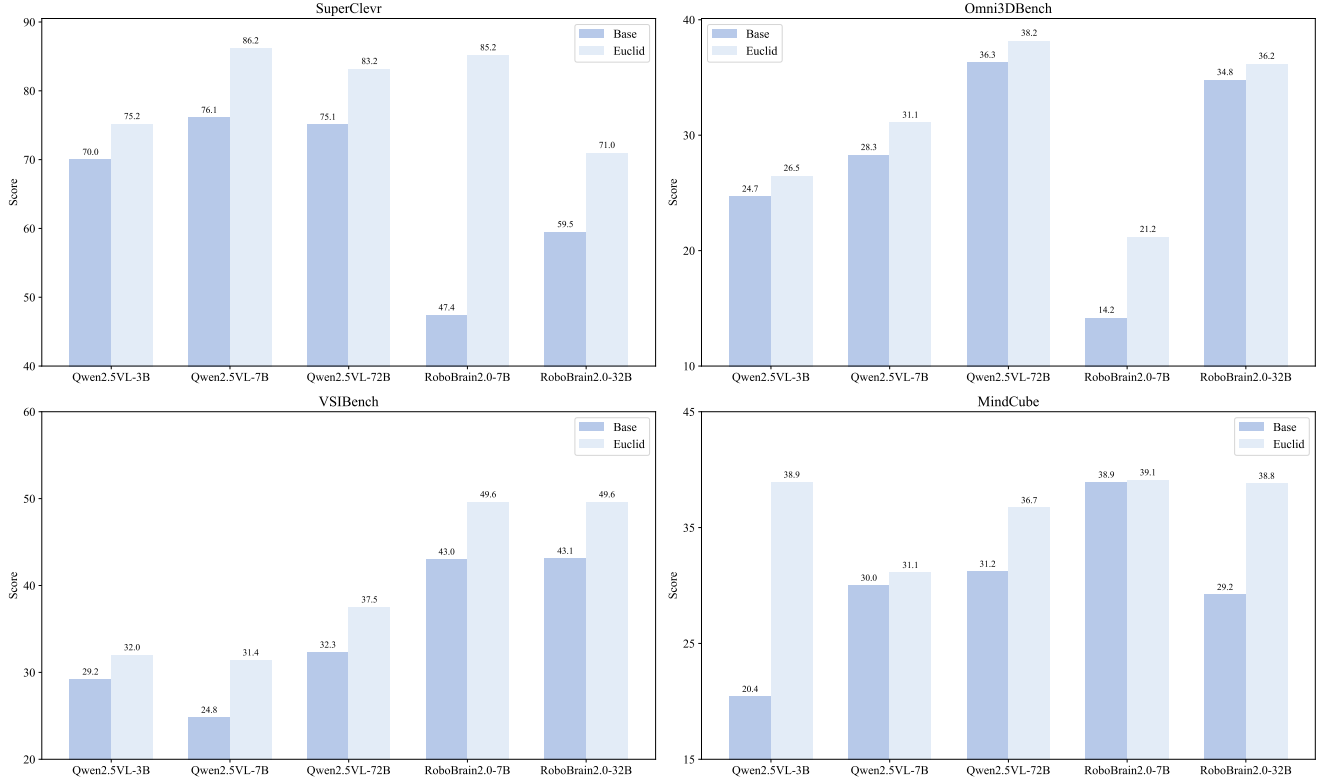


Figure 4. **Performance improvement** on SuperClevr [42], Omni3DBench [53], VSIBench [72], and MindCube [74] after the model has been trained on Euclid25K.

across viewpoints and to reason about occluded or invisible elements. MindCube defines three canonical camera trajectories: Rotation (camera stays in place but rotates to look around; 1,081 samples), Around (camera moves around objects in a circular path; 1,869 samples), and Among (camera moves among objects in a circular path; 18,204 samples). Since all questions follow a multiple-choice format, we evaluate models by exact-match accuracy between the predicted option and the ground-truth answer.

D. More visualization and analysis of results

To present the quantitative gains more intuitively, Fig. 4 plots the base models and their Euclid30K-tuned counterparts side by side. The light bars show consistent accuracy improvements on Super-CLEVR [42], Omni3D-Bench [53], VSI-Bench [72], and MindCube [74], confirming that a compact geometry curriculum injects transferable spatial priors across both Qwen2.5VL and RoboBrain2.0 families.

Fig. 5 and Fig. 6 compare our Euclid30K models with the strongest proprietary, open-source, and purpose-built spatial systems. On VSI-Bench, RoboBrain2.0-Euclid-7B and RoboBrain2.0-Euclid-32B achieve the highest overall scores, surpassing Gemini-2.5 [21] and the Spatial-MLLM [69]. We attribute this result to two factors: (1) the Robo-

Brain2.0 backbone has already learned rich spatial and temporal patterns from large-scale video and spatial datasets; (2) Euclid30K fine-tuning re-introduces general Euclidean constraints, which enhance the model’s ability to generalise without erasing its original strengths.

A similar trend appears on MindCube: Qwen2.5-VL-Euclid-3B variant outperforms Spatial-MLLM, even though Spatial-MLLM augments Qwen 2.5-VL-3B with a dedicated spatial encoder (VGGT) [66] and is trained on a much larger 120 K corpus. These observations highlight the efficiency and broad generalisation afforded by training on a focused set of formal geometry problems.

E. Further Analysis of Limitations

E.1. Temporal skills beyond basic spatial abilities

Educational psychology views spatial intelligence as one branch of a hierarchical model of human abilities. According to Gardner’s theory of multiple intelligences [27] decomposes spatial intelligence into five *Basic Spatial Abilities* (BSAs)—spatial perception, spatial relations, spatial orientation, mental rotation, and spatial visualisation—that serve as building-blocks for higher reasoning. The Euclid30K curriculum targets precisely these BSAs, which

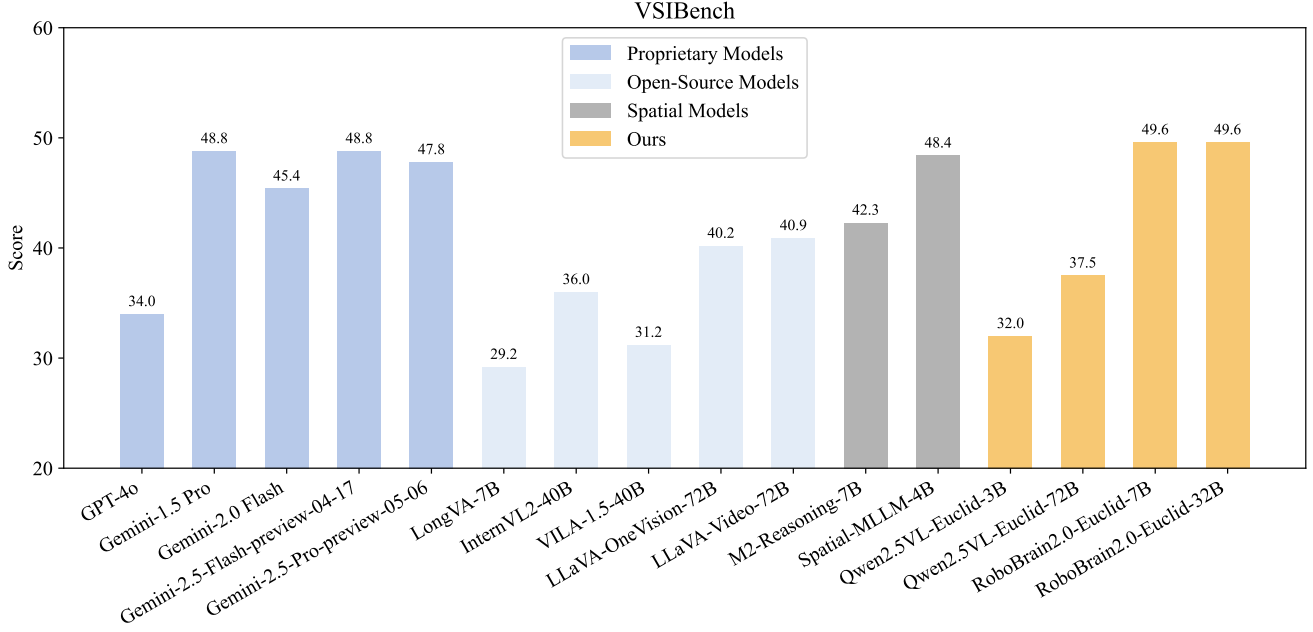


Figure 5. **Visualization Evaluation Results on VSI-Bench** [72]. The performance of Gemini-2.5 is reported from RoboBrain2.0 [8], and the performance of Spatial-MLLM-4B [69] and M2-Reasoning-7B [33] is reported from its original paper, while the results for the other Baseline, Proprietary Models, and Open-source Models are taken from the VSI-Bench benchmark [72]. Qwen2.5VL-Euclid and RoboBrain2.0-Euclid indicate the Qwen2.5VL [9] and RoboBrain2.0 [8] trained with GRPO [62] on the Euclid30K dataset.

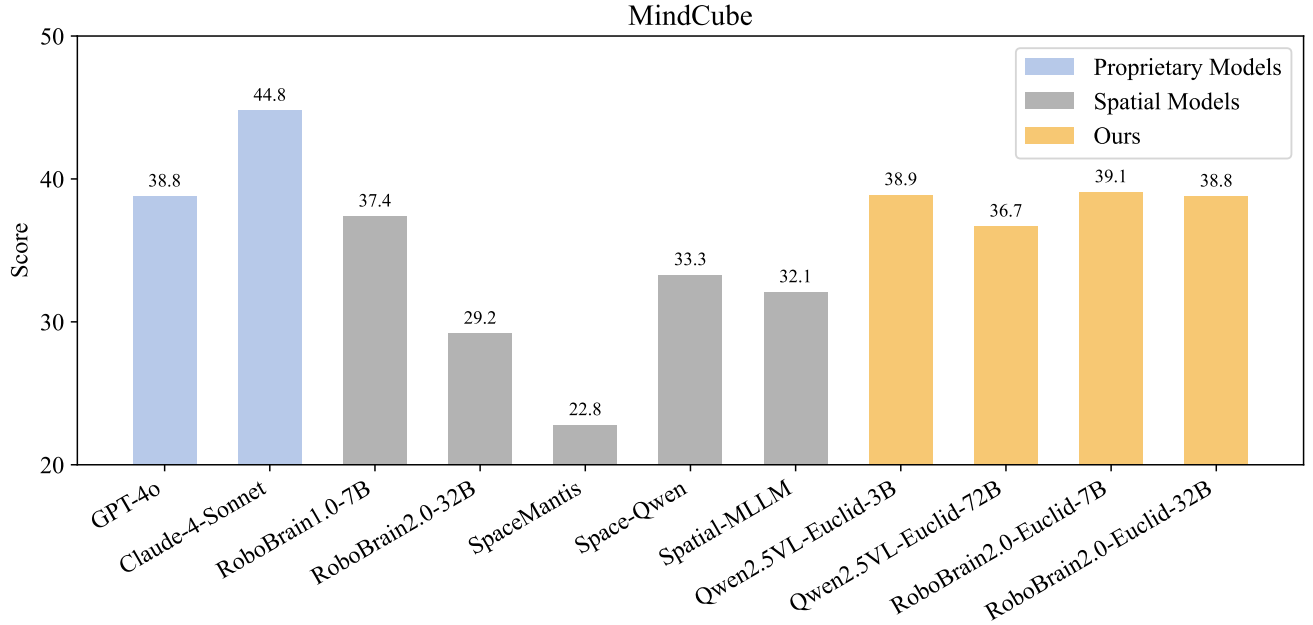


Figure 6. **Visualization Evaluation Results on MindCube** [74]. The performance of Proprietary Models and Spatial Models are taken from the MindCube benchmark [74]. Qwen2.5VL-Euclid and RoboBrain2.0-Euclid indicate the Qwen2.5VL [9] and RoboBrain2.0 [8] trained with GRPO [62] on the Euclid30K dataset.

explains the broad transfer we observe on Super-CLEVR, Omni3D-Bench, and VSI-Bench.

However, tasks such as *Appearance Order* in VSI-Bench require a form of *spatio-temporal memory* that goes beyond



Question: Which object is closer to the camera: the washing machine or the kettle?

Options: Washing Machine, Kettle

GT: Washing Machine

Qwen2.5VL Response:

To determine which object is closer to the camera, let's analyze the image step by step:

1. Identify the washing machine: It is located in the lower part of the image, partially under the counter.
2. Identify the kettle: It is located on the counter, to the right of the washing machine.
3. Compare the positions: **The kettle is positioned on the counter, which is closer to the camera than the washing machine, which is partially hidden under the counter.**

Therefore, the object closer to the camera is the kettle.

Qwen2.5VL Answer: Kettle

Qwen2.5VL -Euclid Response:

To determine which object is closer to the camera, let's analyze the image:

1. The washing machine is positioned in the foreground, directly in front of the camera.
2. The kettle is placed on the countertop, behind the washing machine.

Since the washing machine is in front of the kettle and closer to the camera, it is the object that is closer to the camera.

Final answer: washing machine

Qwen2.5VL Answer: Washing Machine

Figure 7. The response and final answer for Qwen2.5VL-7B [9] and Qwen2,5VL-7B-Euclid in Omni3DBech [53].

the static BSAs. Euclid30K supplies no temporal cues, so models fine-tuned only on Euclidean geometry show limited or negative gains in this category. This boundary clarifies the scope of our surrogate task: mastering Euclidean principles is an effective route to transferable *spatial* skills, but additional temporal supervision is needed to reach full spatio-temporal intelligence.

In addition, on the Rotation tasks in MindCube, the model sometimes shows comparable or slightly decreased performance. This may be because the mental rotation ability required by this task is primarily stimulated by solid-geometry problems, whereas Euclid30K is dominated by plane-geometry questions that rarely involve imagining object rotations. As a result, transfer in this category is less effective. This observation suggests that future work should incorporate more solid-geometry data to achieve a better balance across categories and further strengthen performance on rotation-related reasoning tasks.

F. Discussion of Future Work

Three avenues appear promising.

- **Temporal augmentation.** Combining Euclid30K with video-centered dataset may fill gap in the timing capabilities of models trained with Euclid30K.
- **Higher-level spatial knowledge.** Mining CAD blueprints, architectural plans, and even non-Euclidean geometries might expose the model to more abstract spatial rules.
- **Layered RLVR schedules.** The gains on RoboBrain 2.0 suggest that a second RL-with-verifiable-reward (RLVR) stage on Euclid30K can refine a model that was already trained on generic spatial data. Exploring multi-stage RLVR schedules across diverse spatial domains is a fruitful next step.

By expanding both the temporal and conceptual breadth of our surrogate tasks, future work can push multimodal models toward a more comprehensive and balanced notion of spatial intelligence.



Question: How many objects of the volume of the bedside table can you fit in an object with the volume of the bed?

GT: 15.407

Qwen2.5VL Response:

To determine how many objects of the volume of the bedside table can fit in an object with the volume of the bed, we need to compare the volumes of the two objects.

1. Volume of the Bed: The bed appears to be a standard double bed, which typically has dimensions of approximately 1.5 meters by 2 meters. **Assuming the height of the bed is about 0.2 meters, the volume of the bed can be calculated as:** Volume of the bed = $1.5 \text{ m} \times 2 \text{ m} \times 0.2 \text{ m} = 0.6 \text{ m}^3$
2. Volume of the Bedside Table: The bedside table appears to be a standard nightstand, which typically has dimensions of approximately 0.5 meters by 0.5 meters by 0.5 meters. The volume of the bedside table can be calculated as: Volume of the bedside table = $0.5 \text{ m} \times 0.5 \text{ m} \times 0.5 \text{ m} = 0.125 \text{ m}^3$
3. Comparison: To find out how many bedside tables can fit into the bed, we divide the volume of the bed by the volume of the bedside table:

$$\frac{\text{Volume of the bedside table}}{\text{Volume of the bed}} = \frac{0.125 \text{ m}^3}{0.6 \text{ m}^3} = 4.8$$

Since we are looking for a whole number of bedside tables that can fit, we round down to the nearest whole number.

Qwen2.5VL Answer: 4 (MRA:0)

Qwen2.5VL -Eculid Response:

To determine how many objects of the volume of the bedside table can fit in an object with the volume of the bed, we need to compare the volumes of the two objects.

1. **Volume of the Bed: The bed appears to be a standard double bed, which typically has a volume of around 1.5 cubic meters (assuming standard dimensions and a thickness of the mattress).**
2. Volume of the Bedside Table: The bedside table appears to be a small, rectangular table. Estimating its dimensions, it might be about 0.5 meters in length, 0.4 meters in width, and 0.5 meters in height, giving it a volume of approximately 0.1 cubic meters.
3. Comparison: To find out how many bedside tables can fit into the bed, we divide the volume of the bed by the volume of the bedside table:

$$\frac{0.1 \text{ cubic meters}}{1.5 \text{ cubic meters}} = 15.$$

Therefore, the number of bedside tables that can fit into the bed is 15.

Qwen2.5VL Answer: 15 (MRA:1)

Figure 8. The response and final answer for Qwen2.5VL-7B [9] and Qwen2,5VL-7B-Eculid in Omni3DBech [53].



Question: Measuring from the closest point of each object, which of these objects (heater, trash can, door, cup) is the closest to the microwave?

Options: A. heater, B. trash can, C. door, D. Cup

GT: D. Cup

RoboBrain2.0 Response:

In the visual input, the microwave is positioned on the countertop, slightly to the right of the center. The heater is located on the left side of the countertop, near the sink. The trash can is not visible in the image, suggesting it is not in the immediate vicinity of the microwave. The door is visible on the far left side of the frame, indicating it is not near the microwave. The cup is on the countertop, closer to the microwave than the heater, but not as close as the heater itself.

To determine which object is closest to the microwave, I focus on the spatial arrangement of the heater and the cup. The heater is positioned directly adjacent to the microwave, with no visible gap between them. The cup, while also on the countertop, is further away from the microwave compared to the heater.

Therefore, based on the visual analysis, the heater is the closest object to the microwave. The heater is directly adjacent to the microwave, making it the closest object in terms of proximity.

Thus, the answer is A. heater

RoboBrain2.0 Answer: A. heater

RoboBrain2.0-Eculid Response:

In the visual scene, the microwave is positioned on the countertop, slightly to the right of the center. The heater is located on the left side of the countertop, near the sink. The trash can is not visible in the image, suggesting it is not in the immediate vicinity of the microwave. The door is visible in the background, slightly to the right of the microwave. The cup is on the countertop, near the microwave, and appears to be the closest object to the microwave.

To determine which object is closest to the microwave, I focus on the spatial arrangement of these objects. The cup is directly on the countertop, adjacent to the microwave, indicating it is the closest object to the microwave. The heater is on the left side of the countertop, further away from the microwave. The door is in the background, suggesting it is not the closest. The trash can, if present, would be further away from the microwave than the cup.

Therefore, based on the visual analysis of the scene and the spatial proximity of the objects to the microwave, the object closest to the microwave is the cup. This aligns with option D.

Thus, the answer is: D.

RoboBrain2.0 Answer: D. Cup

Figure 9. The response and final answer for RoboBrain2.0-7B [8] and RoboBrain2.0-7B-Eculid in VSIBench [72].



Question: What is the length of the longest dimension (length, width, or height) of the pillow, measured in centimeters?

GT: 73

RoboBrain2.0 Response:

In the visual input, the pillow is prominently visible on the bed, positioned towards the center of the frame. The pillow is covered with a striped duvet, which provides a clear visual cue for its approximate size. The pillow appears to be elongated, with a noticeable length extending horizontally across the bed. To determine the length of the pillow, I focus on its visible dimensions relative to the bed and other objects in the scene. The pillow's length is visually estimated by comparing it to the bed's width and the visible portion of the bed's length. **The pillow's length seems to be approximately 100 cm, as it spans a significant portion of the bed's width. This estimation is based on the pillow's apparent size relative to the bed's dimensions and the visible length of the bed itself.** Therefore, the length of the pillow, measured in centimeters, is approximately 100 cm.

RoboBrain2.0 Answer: 100 (MRA: 0.3)

RoboBrain2.0-Eculid Response:

From the visual input, the pillow is prominently placed on the bed, with its distinct rectangular shape and soft texture. The pillow's dimensions are visually apparent, with its length extending horizontally across the bed. The pillow's height is relatively low, typical of a standard pillow, and its width spans a significant portion of the bed's surface.

To determine the pillow's length, I focus on its horizontal span across the bed. The pillow's length is visually estimated by comparing it to the bed's width and the visible portion of the bed's surface. **The pillow appears to extend approximately 70 centimeters along the bed's length, which is consistent with the typical dimensions of a standard pillow.**

The pillow's length is the primary dimension of interest, as the task specifies measuring the longest dimension. The pillow's width and height are not the focus here, as the task specifies the length.

Therefore, based on the visual analysis of the pillow's dimensions in the scene, the pillow's length is approximately 70 centimeters. This conclusion is supported by the pillow's visible horizontal span across the bed, which aligns with the typical dimensions of a standard pillow.

RoboBrain2.0 Answer: 70 (MRA: 1.0)

Figure 10. The response and final answer for RoboBrain2.0-7B [8] and RoboBrain2.0-7B-Eculid in VSIBench [72].



Question: Is the number of big yellow objects to the left of the rubber car the same as the number of motorbikes that are right of the regular bus? True or False.

GT: False

RoboBrain2.0 Response:

In the visual input, I focus on identifying the big yellow objects to the left of the rubber car. The rubber car is a prominent yellow vehicle located centrally in the scene. To its left, there is a large yellow bicycle, which is a big yellow object. This bicycle is positioned prominently to the left of the rubber car.

Next, I identify the motorbikes in the scene. There are two motorbikes visible: one is a red motorbike located to the left of the yellow bicycle, and the other is a gold motorbike positioned to the right of the yellow bicycle. **The gold motorbike is the one that is right of the regular bus, which is a blue bus located on the far right side of the scene.**

Now, I compare the number of big yellow objects to the left of the rubber car with the number of motorbikes to the right of the regular bus. There is one big yellow object (the bicycle) to the left of the rubber car, and there is one motorbike (the gold one) to the right of the regular bus.

Therefore, the number of big yellow objects to the left of the rubber car is equal to the number of motorbikes to the right of the regular bus. This leads to the conclusion that the statement is True.

RoboBrain2.0 Answer: True

RoboBrain2.0-Eculid Response:

To determine the number of big yellow objects to the left of the rubber car, I first identify the rubber car, which is a yellow minivan located centrally in the image. To its left, I observe a large yellow bicycle. This bicycle is a prominent, big yellow object.

Next, I focus on the motorbikes in the scene. There are two motorbikes visible: one is a large yellow motorcycle positioned to the right of the minivan, and the other is a smaller red motorbike located to the left of the minivan.

Now, I compare the number of big yellow objects to the left of the rubber car with the number of motorbikes to the right of the regular bus. The big yellow object to the left of the minivan is the bicycle.

The motorbikes are positioned to the left of the minivan, not to the right of the bus.

Therefore, the number of big yellow objects to the left of the rubber car (the bicycle) is not the same as the number of motorbikes to the right of the regular bus. **The bicycle is not on the right side of the bus, and the motorbikes are not on the right side of the bus either.** Thus, the answer is: False.

RoboBrain2.0 Answer: False

Figure 11. The response and final answer for RoboBrain2.0-7B [8] and RoboBrain2.0-7B-Eculid in SuperClver [42].



Question: What number of objects are tiny cyan rubber things or metallic objects that are on the left side of the bus?

GT: 3

RoboBrain2.0 Response:

In the visual input, I observe several objects, including motorcycles, a bus, and airplanes. The task requires identifying objects that are either tiny cyan rubber things or metallic objects located on the left side of the bus.

First, I focus on the metallic objects on the left side of the bus. The bus is positioned towards the upper center of the scene. To its left, there are two prominent metallic objects: a gold-colored motorcycle and a gold-colored scooter. These objects are clearly metallic and positioned to the left of the bus.

Next, I consider the tiny cyan rubber things. Upon examining the scene, there are no objects that appear to be tiny and distinctly cyan rubber. **The bus itself is not a tiny object, and the other objects, such as the motorcycles and airplanes, are not tiny either.**

Based on this analysis, the metallic objects on the left side of the bus are the gold-colored motorcycle and scooter. These are the only objects that meet the criteria of being metallic and positioned to the left of the bus.

Therefore, the answer is that there are two objects that are metallic and on the left side of the bus

RoboBrain2.0 Answer: 2

RoboBrain2.0-Eculid Response:

In the visual scene before me, I observe several metallic objects, including a blue airplane, a yellow scooter, and a yellow motorcycle. The blue airplane is positioned prominently in the foreground, while the yellow scooter and motorcycle are situated towards the left side of the scene. The bus, which is a metallic object, is located further back in the scene, behind the motorcycle.

To determine the number of metallic objects on the left side of the bus, I focus on the objects that are positioned to the left of the bus. The yellow scooter and the yellow motorcycle are both on the left side of the bus. **The blue airplane, although not directly to the left of the bus, is also considered as part of the left side due to its position relative to the bus.**

The blue airplane, being a metallic object, is included in the count. The yellow scooter and motorcycle are also metallic and are positioned to the left of the bus. **Therefore, the metallic objects on the left side of the bus are the blue airplane, the yellow scooter, and the yellow motorcycle.**

Thus, based on the visual analysis of the scene and the spatial arrangement of the objects, the number of metallic objects on the left side of the bus is three. Therefore, the answer is: 3.

RoboBrain2.0 Answer: 3

Figure 12. The response and final answer for RoboBrain2.0-7B [8] and RoboBrain2.0-7B-Eculid in SuperClver [42].