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

Supplementary Material

Supplementary Contents

A. Proof of Standard Domain-Adaptation Bound

This section provides the complete proof for the standard domain-adaptation bound introduced in Sec. 3.1. We follow the same notation as in the main text.

Let $h^* \in \mathcal{H}$ be the (ideal) hypothesis defined by

$$h^* = \arg\min_{h \in \mathcal{H}} \big(\epsilon_S(h) + \epsilon_T(h)\big), \tag{7}$$

and denote $\epsilon_{\text{ideal}} = \epsilon_S(h^*) + \epsilon_T(h^*)$. By the definition of $d_{\mathcal{H}\Delta\mathcal{H}}$, we have

$$\big|\epsilon_S(h, h^*) - \epsilon_T(h, h^*)\big| \leq \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T). \tag{8}$$

Then, by the triangle inequality and Eq. (8), we have:

$$\begin{aligned} \epsilon_T(h) &\leq \epsilon_T(h^*) + \epsilon_T(h, h^*) \\ &= \epsilon_T(h^*) + \epsilon_S(h, h^*) + \big(\epsilon_T(h, h^*) - \epsilon_S(h, h^*)\big) \\ &\leq \epsilon_T(h^*) + \epsilon_S(h, h^*) + \big|\epsilon_T(h, h^*) - \epsilon_S(h, h^*)\big| \\ &\leq \epsilon_T(h^*) + [\epsilon_S(h) + \epsilon_S(h^*)] + \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) \\ &= \epsilon_S(h) + \epsilon_{ideal} + \tfrac{1}{2}\, d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T), \end{aligned} \tag{9}$$

The above proof refers to [6] and [19].

In the setting considered in this paper we examine generalisation from (i) formal geometry to broad spatial intelligence and (ii) spatial sub-skills to broader spatial intelligence. Given continued scaling of model capacity and data, along with advances in training methodologies, it is reasonable to anticipate the emergence of a sufficiently strong hypothesis h^* for which the source-target joint errors in both regimes satisfy:

$$\epsilon_S(h^*) \approx \epsilon_T(h^*) \approx 0, \tag{10}$$

i.e., the error magnitudes on geometry tasks and spatial-intelligence tasks (and likewise on spatial sub-tasks and full spatial intelligence) become simultaneously negligible. When this asymptotic condition holds for both sides of our comparisons, the corresponding ideal terms ϵ_{ideal} in each bound are effectively 0, making their omission justified. Consequently, Eq. (9) simplifies to:

$$\epsilon_T(h) \lesssim \epsilon_S(h) + \tfrac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T), \tag{11}$$

Therefore, if $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T)$ is sufficiently small, the population gap $\epsilon_T(h) - \epsilon_S(h)$ also becomes small, which allows the source distribution to serve as a reliable surrogate for the target distribution [1, 16, 17].

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 [10, 14].

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 [13]. In STEM and graphical education, higher spatial ability is associated with better problem-solving performance and more effective strategies [7]. 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 [21].

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 [22, 23]. Differences in dominance between logical-mathematical and visual-spatial intelligence yield distinct pathways for geometric reasoning, further tying geometry problem solving to spatial constructs [3]. 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 [2]. Geometrical-mechanical intelligence games, implemented in quasi-experiments with pre/post testing, significantly enhance spatial visualisation and spatial relations skills [26]. Overall, the balance of evidence indicates that well-designed geometric practice is an effective means to cultivate spatial abilities.

Moreover, neuroscience research reveals that early exposure to Euclidean geometric structures fundamentally shapes spatial representations. Studies examining hippocampal activity in rodents reared in spherical versus cuboid environments demonstrate that experience with canonical Euclidean features (edges, corners, planes) enriches the repertoire of preconfigured neuronal patterns and enhances the brain's ability to discriminate between distinct spatial layouts [9]. While these findings originate from animal models, they provide convergent biological evidence that geometric experience during development can refine spatial coding mechanisms—a principle that may extend to learning systems more broadly.

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 [25] and EasyR1 [35] to train the Qwen2.5-VL series, Qwen3-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 1024 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 lmms-eval toolkit [33] to ensure consistent decoding across models. In the test, to ensure the reproducibility of the results, we follow VSI-Bench [29] and MindCube [32] 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. For reproducibility, detailed testing scripts are provided in the supplementary materials.

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 and Qwen3VL-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 QwenVL-Instruct Prompt Template

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

During evaluation, to mitigate prompt-template bias, we run each model with both its native instruction-style prompt and the unified Euclid reasoning template, and report the better-performing variant. 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.

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 [29] contains more than 5,130 egocentric videos question-answer pairs sourced from ARKitScenes[5], ScanNet [8], and ScanNet++ [31]. 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 [15] 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 [20] 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 [29]. Specifically, we calculate mean relative accuracy (MRA) across confidence thresholds $\mathcal{C} = \{0.5, 0.55 \ldots, 0.95\}$ for the numerical question tasks and report exact-match accuracy for multiple-choice tasks.

Setup in MindCube. MindCube [32] 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 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 Experiment and Visualization

D.1. More experiment about main 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 [15], Omni3D-Bench [20], VSI-Bench [29], and MindCube [32], confirming that a compact geometry curriculum injects transferable spatial priors across both Qwen2.5VL, Qwen3VL and RoboBrain2.0 families. Additionally, we include results for Qwen2.5VL-7B and Qwen2.5VL-32B in the Fig. 4, which exhibit consistent improvements across all four benchmarks following the same geometric surrogate-task training, reaffirming that geometry serves as an effective surrogate task for spatial intelligence and further validating the robustness and generality of this approach.

Beyond the aggregate accuracy gains, Figures5–12 qualitatively illustrate how geometry tuning alters intermediate reasoning. After Euclid30K training the model produces more coherent multi view descriptions (Fig.7), applies geometric similarity relations correctly (Fig.10), uses quadrant and cardinal directional cues with fewer ambiguities (Fig.8, Fig.11), and leverages perspective driven size cues (near–far size scaling) more systematically (Fig.12). It also shows clearer distance estimation chains (Fig.5, Fig.9), improved object size estimation (Fig.6, Fig.12), more reliable counting with cross view consistency (Fig. 7), and fewer heuristic shortcuts (*e.g.*, premature guesses without spatial justification). These qualitative traces are consistent with the quantitative improvements: they suggest the model has internalized foundational Euclidean principles (similarity, proportionality, relative position, viewpoint coherence) and can deploy them across distinct downstream spatial tasks.

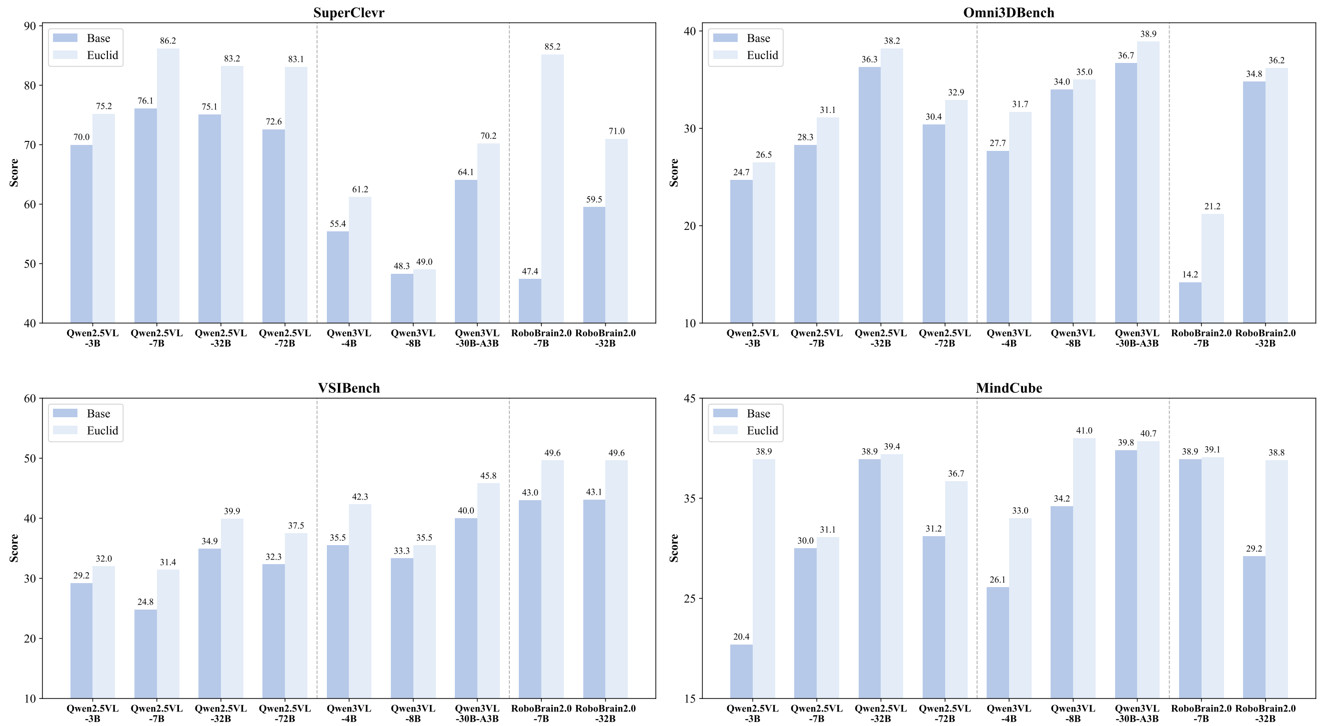


Figure 4. **Performance improvement** on SuperClevr [15], Omni3DBench [20], VSIBench [29], and MindCube [32] after the model has been trained on Eculid30K.

D.2. More experiment about ablation study

Tab. 6 presents a detailed analyses of the ablation study results on VSI-Bench [29], complementing the summary in Sec. 4.4. To isolate the contribution of Euclid30K from potential confounds such as additional training data or GRPO-induced generalization, we trained each model variant on a 30K-sample subset of the spatial-intelligence dataset Clevr-CoGenT [12] using identical hyperparameters, rollout budgets, and training epochs.

Overall Performance. Across all model families, Euclid30K training consistently yields higher overall accuracy than Clevr-CoGenT training. For instance, Qwen2.5VL-3B improves from 29.2% (base) to 32.0% with Euclid30K versus 31.3% with Clevr-CoGenT; Qwen3VL-30B-A3B rises from 40.0% to 45.8% with Euclid30K versus 44.2% with Clevr-CoGenT; and RoboBrain2.0-32B jumps from 43.1% to 49.6% with Euclid30K versus 46.7% with Clevr-CoGenT. These results indicate that Euclid30K provides a more transferable spatial foundation than an equal volume of non-geometric spatial data.

Task-Specific Patterns. Clevr-CoGenT training yields targeted improvements on tasks closely aligned with its original design, such as object counting and relative direction. For example, Qwen2.5VL-3B trained on Clevr-CoGenT achieves 40.5% on object counting, slightly outperforming Euclid30K training (38.3%), and similar patterns appear for relative direction in several variants. Conversely, Euclid30K training produces broader gains across multiple categories, particularly on size-estimation tasks (like object size and absolute distance estimation), where geometric reasoning principles directly apply. Qwen2.5VL-32B trained on Euclid30K reaches 55.8% on object size versus 44.3% with Clevr-CoGenT, and 46.0% on room size versus 40.8%.

This ablation study reveals a fundamental trade-off: task-specific spatial datasets enhance performance on related categories, whereas structured geometric datasets provide a more general-purpose spatial reasoning foundation that transfers more uniformly across diverse task types. Euclid30K's strength lies in instilling first-principles geometric knowledge—distance, proportion, similarity, and spatial relations—that applies broadly, even to tasks not explicitly represented in the training corpus. The pattern holds across Qwen2.5VL, Qwen3VL, and RoboBrain2.0 families, and across parameter scales ranging from 3B to 72B, from dense to MoE. This consistency suggests that the advantage conferred by geometric surrogate tasks is not an artifact of a particular architecture or capacity regime, but reflects a more general principle: formal Euclidean reasoning offers a robust substrate for spatial intelligence.

In summary, Tab. 6 substantiates the claim that geometry serves as an effective surrogate task for spatial intelligence,

Methods	Numerical Question				Multiple-Choice Question				Overall
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
Qwen2.5VL-series									
Qwen2.5VL-3B	35.6	23.4	34.9	16.6	34.4	40.7	26.3	**21.8**	29.2
Qwen2.5VL-Space-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-Space-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-32B	22.4	27.0	37.9	38.0	39.4	**40.1**	33.5	**41.3**	34.9
Qwen2.5VL-Space-32B	**45.6**	25.9	44.3	40.8	42.0	39.3	28.4	35.3	37.7
Qwen2.5VL-Euclid-32B	38.7	**30.9**	**55.8**	**46.0**	**43.7**	37.1	**34.0**	33.2	**39.9**
Qwen2.5VL-72B	13.6	19.6	40.9	41.1	37.7	35.3	**34.0**	36.2	32.3
Qwen2.5VL-Space-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**
Qwen3VL-series									
Qwen3VL-4B	28.5	33.0	32.6	43.5	40.3	40.0	33.0	33.2	35.5
Qwen3VL-Space-4B	32.1	37.1	44.7	**51.0**	**49.3**	43.8	**37.6**	38.5	41.8
Qwen3VL-Euclid-4B	**33.3**	**37.4**	**49.5**	48.3	46.5	**46.3**	34.0	**42.9**	**42.3**
Qwen3VL-30B-A3B	27.4	32.3	53.6	**44.0**	42.1	36.2	35.5	48.9	40.0
Qwen3VL-Space-30B-A3B	29.9	36.8	58.0	43.4	47.9	49.1	**35.9**	52.7	44.2
Qwen3VL-Euclid-30B-A3B	**33.5**	**37.5**	**64.1**	43.1	**48.7**	**49.6**	35.6	**54.4**	**45.8**
RoboBrain2.0-series									
RoboBrain2.0-7B	46.0	32.7	58.9	35.9	45.9	41.5	30.9	55.2	43.0
RoboBrain2.0-Space-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-Space-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 6. **Ablation experiment on VSI-Bench** [29]. We compare training a model on a 30K 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.

Methods	Numerical Question				Multiple-Choice Question				Overall
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
Qwen2.5VL-7B	**39.5**	17.8	16.9	5.8	33.8	36.7	24.7	22.8	24.8
Spat1-SSRL-7B [18]	37.1	**24.6**	22.9	17.0	36.3	35.1	**33.5**	**27.3**	29.2
Qwen2.5VL-Euclid-7B	38.8	22.8	**37.3**	**23.2**	**38.3**	**38.5**	25.8	26.5	**31.4**

Table 7. **Comparisons with Spat1-SSRL-7B on VSI-Bench**. **Bolding** indicates the best score within each model type.

Methods	SuperClevr	Omni3DBench	MindCube
Qwen2.5VL-7B	76.1	28.3	30.0
Spat1-SSRL-7B [18]	76.3	**33.1**	30.6
Qwen2.5VL-Euclid-7B	**86.2**	31.1	**31.1**

Table 8. **Comparisons with Spat1-SSRL-7B on SuperClevr, Omni3D Bench, and MindCube.** **Bolding** indicates the best score within each model type.

with Euclid30K training delivering superior overall performance and broader generalization compared to equal-sized non-geometric spatial datasets.

D.3. Comparison with Other Surrogate Tasks

Beyond our geometry-first curriculum, several studies explore alternative surrogate tasks to strengthen spatial intelligence. In this subsection, we compare against a representative approach, Spatial-SSRL [18]. Spatial-SSRL pro-

Methods	Numerical Question				Multiple-Choice Question				Overall
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
Qwen3VL-30B-A3B	27.4	32.3	53.6	**44.0**	42.1	36.2	**35.5**	48.9	40.0
+ GeometricSFT	**29.6**	**33.8**	**56.9**	39.1	**46.8**	**47.8**	35.1	**60.2**	**43.7**

Table 9. **Evaluation on VSI-Bench**. + GeometricSFT indicate the Qwen3VL-30B-A3B trained with SFT on the Geo170K dataset [11]. **Bolding** indicates the best score within each model type.

Methods	Numerical Question				Multiple-Choice Question				Overall
	Obj. Cnt.	Abs. Dist.	Obj. Size	Room Size	Rel. Dist.	Rel. Dir.	Route Plan	Appr. Order	
Qwen2.5VL-7B	39.5	17.8	16.9	5.8	33.8	36.7	24.7	22.8	24.8
VST-7B [30]	56.5	35.1	69.0	53.3	50.1	11.7	28.4	50.3	44.3
VST-Euclid-7B	**66.4**	**38.4**	**74.2**	**60.1**	**56.5**	**48.6**	**41.8**	**57.8**	**55.5**

Table 10. **Evaluation on VSI-Bench** . VST-Euclid indicate the VST [30] trained with GRPO [24] on the Euclid30K dataset. **Bolding** indicates the best score within each model type.

Methods	SuperClevr	Omni3DBench	MindCube
Qwen3VL-30B-A3B	64.1	36.7	**39.8**
+ GeometricSFT	**66.5**	**40.5**	38.3

Table 11. **Evaluation on SuperClevr, Omni3D Bench, and MindCube.** + GeometricSFT indicate the Qwen3VL-30B-A3B trained with SFT on the Geo170K dataset [11]. **Bolding** indicates the best score within each model type.

Methods	SuperClevr	Omni3DBench	MindCube
Qwen2.5VL-7B	76.1	28.3	30.0
VST-7B [30]	83.1	36.5	**35.5**
VST-Euclid-7B	**86.3**	**37.1**	34.8

Table 12. **Comparisons with VST-7B on SuperClevr, Omni3D Bench, and MindCube.** VST-Euclid indicate the VST [30] trained with GRPO [24] on the Euclid30K dataset. **Bolding** indicates the best score within each model type.

Category	Parameter	Value
Dataset	image max pixels	262144
Dataset	video max pixels	16384
Dataset	cutoff len	8192
Method	finetuning type	lora
Method	lora rank	8
Method	lora target	all
Train	per device train batch size	4
Train	gradient accumulation steps	4
Train	learning rate	1e-4
Train	num train epochs	1
Train	lr scheduler type	cosine
Train	weight decay	0.01
Train	warmup ratio	0.1
Train	bf16	true

Table 13. **Supervised Fine-tuning (SFT) configuration in LLaMAFactory framework [34].**

poses a self-supervised pipeline that derives verifiable signals from ordinary RGB/RGB-D images via five pretext tasks capturing 2D/3D spatial structure: shuffled patch reordering, flipped patch recognition, cropped-patch inpainting, regional depth ordering, and relative 3D position prediction. The released configuration uses 81K image-QA pairs to train a 7B model with RLVR.

Tab. 7 compares a 7B backbone trained on our 30K Euclid30K geometry corpus (RLVR, identical decoding as Sec. C) with Spatial-SSRL-7B on VSI-Bench. Euclid30K achieves a higher overall score (31.4 vs. 29.2) with substantially fewer training samples (30K vs. 81K; 63% fewer). Gains concentrate on size- and distance-estimation categories where geometric constraints are directly applicable, while Spatial-SSRL shows stronger performance on route planning and appearance order. Cross-benchmark results in Tab. 8 show that Euclid30K improves Super-CLEVR and remains competitive on Omni3D-Bench and MindCube under the same inference protocol.

E. More Discussion

E.1. Supervised Fine-tuning (SFT).

For mathematical problems, collecting step by step solution annotations is substantially more expensive than answer only labels. Accordingly, Euclid30K does not currently include process annotations, which makes it less suited to SFT routines that rely on explicit reasoning traces. Nevertheless, geometry QA can still serve as a surrogate task for spatial intelligence under SFT. In this subsection, in order to show that geometry can as a surrogate task for spatial intelligence not only under reinforcement learning but also under supervised fine tuning, we conduct SFT on Geo170K, a dataset that supplies explicit reasoning trajectories. Providing these intermediate steps helps the model internalize Euclidean principles.

As shown in Tab. 9, Qwen3VL-30B-A3B improves on VSI Bench from 40.0 to 43.7 overall after Geo170K SFT (the gain is smaller than the 40.0 to 45.8 increase achieved with Euclid30K RL). In Tab. 11, cross benchmark results also rise on SuperCLEVR (64.1 to 66.5) and Omni3D Bench (36.7 to 40.5), while MindCube shows a modest decrease (39.8 to 38.3). The decline on MindCube is plausibly due to the composition of Geo170K: most problems are plane geometry items and provide little direct supervision for viewpoint change or three dimensional mental imagery. Even with this limitation, the consistent gains on the majority of spatial benchmarks reinforce our claim that solving geometry problems is an effective surrogate task for spatial intelligence.

E.2. Model-Specific Performance Variations

As noted in Sec. 5.2, we hypothesize that the performance gains after geometry tuning are model specific and depend on capabilities inherited from earlier training data. In this subsection, we use VST-7B-RL [30] as a representative case to test this hypothesis. VST-7B first applies supervised fine-tuning on VST-P dataset with 4.1M samples that cover 19 spatial skills across single-image, multi-image, and video. It then uses reinforcement learning on VST-R dataset with 135K reasoning samples. We start from the released VST-7B-RL checkpoint, add Euclid30K training with GRPO and keep decoding and evaluation identical those to used in the rest of this appendix. Results in Tab. 10 show clear gains on VSI Bench. The overall score moves from 44.3 to 55.5. Tab. 12 shows small improvements on SuperCLEVR and Omni3DBench, and a small decrease on MindCube from 35.5 to 34.8.

A reasonable explanation is that the VST dataset has limited coverage of viewpoint changes and spatiotemporal consistency. The samples of "camera motion" and "camera-camera" combined account for only about 3% of the total dataset. However, MindCube heavily relies on these patterns. Meanwhile, the plane geometry problems in Euclid30K do not include explicit supervision signals for viewpoint transformations. This suggests that the role of Euclid30K is to formalize and reinforce the spatial concepts already present in the model, leading to significant improvements in tasks such as metric estimation and relational reasoning. However, when certain key concepts (such as robust viewpoint transformation ability) are weakly represented in the pretraining, geometric fine-tuning may offer little benefit and might even slightly interfere with the model's existing heuristic strategies. These support our hypothesis that models with richer prior spatial concepts obtain larger gains from Euclid30K, while gaps in prior coverage limit or occasionally reduce transfer.

F. More Visualization about Euclid30K

Fig. 13–15 provide additional Euclid30K examples, illustrating diversity in problem types, diagram styles, and reasoning complexity. These samples complement the main paper by showing the range of geometric configurations and textual formulations present in the Euclid30K dataset.

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. 1

[2] Laszlo Aszalos and Maria Bako. How can we improve the spatial intelligence. In *6th International Conference on Applied Informatics*, Eger, Hungary, 2004. 2

[3] 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 (IJCSE 2020)*, pages 203–207. Atlantis Press, 2020. 2

[4] 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, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 10, 11

[5] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - 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. 3

[6] 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. 1

[7] 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. 1

[8] 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. 3

[9] Usman Farooq and George Dragoi. Experience of euclidean geometry sculpts the development and dynamics of rodent hippocampal sequential cell assemblies. *Nature Communications*, 15(1):8417, 2024. 2

[10] Jacob Feldman. What is a visual object? *Trends in Cognitive Sciences*, 7(6):252–256, 2003. 1

[11] Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing HONG, Jianhua Han, Hang Xu, Zhenguo Li, et al. G-llava: Solving geometric problem with multi-modal large language model. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[12] 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. 4

[13] 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. 1

[14] 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. 1

[15] 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. 3, 4, 16, 17

[16] 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. 1

[17] 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. 1

[18] Yuhong Liu, Beichen Zhang, Yuhang Zang, Yuhang Cao, Long Xing, Xiaoyi Dong, Haodong Duan, Dahua Lin, and Jiaqi Wang. Spatial-ssrl: Enhancing spatial understanding via self-supervised reinforcement learning. *arXiv preprint arXiv:2510.27606*, 2025. 5

[19] Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009. 1

[20] Damiano Marsili, Rohun Agrawal, Yisong Yue, and Georgia Gkioxari. Visual agentic ai for spatial reasoning with a dynamic api. *arXiv preprint arXiv:2502.06787*, 2025. 3, 4, 10, 11

[21] Nora S Newcombe and Andrea Frick. Early education for spatial intelligence: Why, what, and how. *Mind, Brain, and Education*, 4(3):102–111, 2010. 2

[22] 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. 2

[23] 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. 2

[24] 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. 6

[25] 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. 2

[26] 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. 2

[27] BAAI RoboBrain Team, Mingyu Cao, Huajie Tan, Yuheng Ji, Xiansheng Chen, Minglan Lin, Zhiyu Li, Zhou Cao, Pengwei Wang, Enshen Zhou, Yi Han, Yingbo Tang, Xiangqi Xu, Wei Guo, Yaoxu Lyu, Yijie Xu, Jiayu Shi, Mengfei Du, Cheng Chi, Mengdi Zhao, Xiaoshuai Hao, Junkai Zhao, Xiaojie Zhang, Shanyu Rong, Huaihai Lyu, Zhengliang Cai, Yankai Fu, Ning Chen, Bolun Zhang, Lingfeng Zhang, Shuyi Zhang, Dong Liu, Xi Feng, Songjing Wang, Xiaodan Liu, Yance Jiao, Mengsi Lyu, Zhuo Chen, Chenrui He, Yulong Ao, Xue Sun, Zheqi He, Jingshu Zheng, Xi Yang, Donghai Shi, Kunchang Xie, Bochao Zhang, Shaokai Nie, Chunlei Men, Yonghua Lin, Zhongyuan Wang, Tiejun Huang, and Shanghang Zhang. Robobrain 2.0 technical report. *arXiv preprint arXiv:2507.02029*, 2025. 14, 15, 16, 17

[28] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 12, 13

[29] 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. 2, 3, 4, 5, 12, 13, 14, 15

[30] Rui Yang, Ziyu Zhu, Yanwei Li, Jingjia Huang, Shen Yan, Siyuan Zhou, Zhe Liu, Xiangtai Li, Shuangye Li, Wenqian Wang, et al. Visual spatial tuning. *arXiv preprint arXiv:2511.05491*, 2025. 6, 7

[31] 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. 3

[32] 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. *arXiv preprint arXiv:2506.21458*, 2025. 2, 3, 4

[33] 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. *arXiv preprint arXiv:2407.12772*, 2024. 2

[34] Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand, 2024. Association for Computational Linguistics. 6

[35] Yaowei Zheng, Junting Lu, Shenzhi Wang, Zhangchi Feng, Dongdong Kuang, and Yuwen Xiong. Easyr1: An efficient, scalable, multi-modality rl training framework. *github repository*, 2025. 2

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 -Eculid 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 5. The response and final answer for Qwen2.5VL-7B [4] and Qwen2.5VL-Eculid-7B in Omni3DBech [20].

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 m × 2 m × **0.2 m = 0.6 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 m × 0.5 m × 0.5 m = 0.125 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:

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

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)

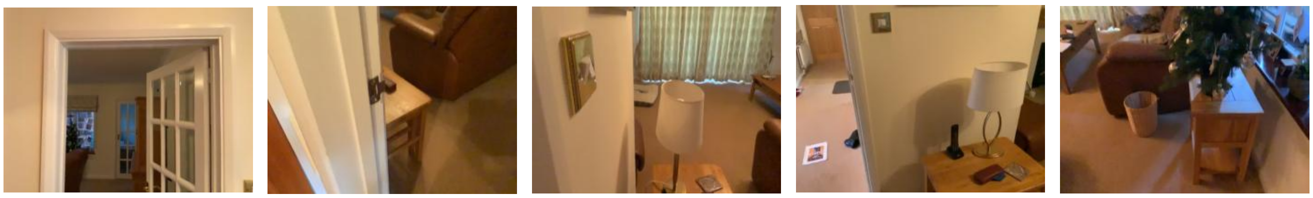
Qwen2.5VL -Eculid Response:
To determine how many objects of the volume of the bedside table can fit in an object with the volume of the bed, we need to compare the volumes of the two objects.
1. **Volume of the Bed: The bed appears to be a standard double bed, which typically has a volume of around 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:

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

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

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

Figure 6. The response and final answer for Qwen2.5VL-7B [4] and Qwen2.5VL-Eculid-7B in Omni3DBech [20].



Question: How many table(s) are in this room? **GT:** 3

Qwen3VL Response:
Looking at the frames, I can identify several tables:
1. **In the initial frames, there is a small wooden table next to a doorway, with a lamp and a phone on it.**
2. Later, in the living room, there is a larger wooden coffee table in the center of the room.
3. Also in the living room, **there is a small wooden side table next to the brown leather ottoman, which has a lamp and a phone on it.**
4. There is also a small wooden side table next to the Christmas tree, which has a phone on it.
Therefore, **there are 4 tables visible in the room.**
Qwen3VL Answer: 4

Qwen3VL-Eculid Response:
Looking at the frames, I can identify several tables:
1. In the initial frames, there is a small wooden table next to the doorway, with a lamp and a phone on it.
2. Later, in the living room, there is a larger wooden coffee table in the center of the room.
3. **Also in the living room, there is a small wooden side table next to the brown leather ottoman, which has a lamp and a phone on it (this is the same table seen earlier near the doorway).**
4. There is also a small wooden side table next to the Christmas tree, which has a small item on it.\n\nTherefore, there are at least 3 distinct tables visible in the room: the coffee table, the side table next to the ottoman, and the side table next to the Christmas tree.
Qwen3VL-Eculid Answer: 3

Figure 7. The response and final answer for Qwen3VL-4B [28] and Qwen3VL-Eculid-4B in VSIBench [29].

Question: If I am standing by the door and facing the tv, is the trash bin to my front-left, front-right, back-left, or back-right? The directions refer to the quadrants of a Cartesian plane (if I am standing at the origin and facing along the positive y-axis).",

Options: A. back-right, B. front-right, C. back-left, D. front-left

GT: D

Qwen3VL Response:
Let's analyze the scene step by step:
1. **Identify the starting position**: You are standing by the door. From the frames, the door is on the right side of the room, near the red wall and the window. So, you are on the right side of the room.
2. **Determine the direction you are facing**: You are facing the TV. The TV is mounted on the white wall, which is opposite the red wall where the door is. Therefore, you are facing the white wall.
3. **Locate the trash bin**: The trash bin is visible in the frames near the window, which is on the far side of the room, next to the door. It is positioned in the corner where the red wall and the window/wall meet.
4. **Determine the relative position**: **Since you are standing by the door (on the right side of the room) and facing the TV (which is on the opposite wall), you are facing the white wall. The trash bin is in the corner on the far side of the room, which is to your left and behind you. From your perspective, it is in the back-left quadrant.**
Therefore, the trash bin is to your back-left. The correct option is C.
Qwen3VL Answer: C

Qwen3VL-Eculid Response:
Let's analyze the scene step by step:
1. Identify the door and the TV: The door is visible in the frame, and the TV is mounted on the wall. The person is standing by the door.\n2.
Determine the facing direction: The question asks what is to the front-left, front-right, etc., when facing the TV. So, we need to determine the position of the trash bin relative to the TV from the perspective of someone standing at the door and facing the TV.
3. Locate the trash bin: The trash bin is visible in the frame, positioned near the wall, to the left of the TV from the viewer's perspective.
4. Determine the relative position: If a person is standing by the door and facing the TV, **they are facing the wall where the TV is mounted**. The trash bin is located to the left of the TV. **Therefore, from the person's perspective, the trash bin is in the front-left direction.** The final answer is D.
Qwen3VL-Eculid Answer: D

Figure 8. The response and final answer for Qwen3VL-30B-A3B [28] and Qwen3VL-Eculid-30B-A3B in VSIBench [29].

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

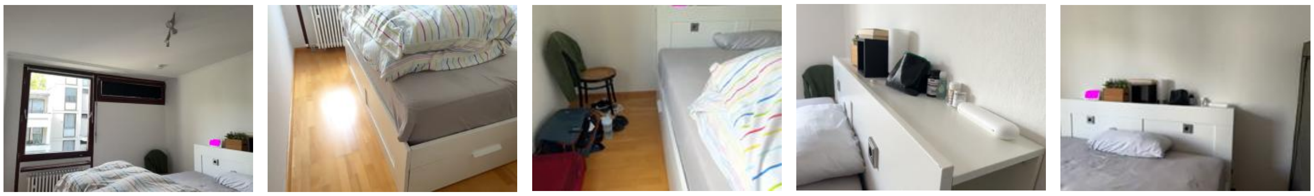
RoboBrain2.0 Response:
In the visual input, the microwave is positioned on the countertop, slightly to the right of the center. The heater is located on the left side of the countertop, near the sink. The trash can is not visible in the image, suggesting it is not in the immediate vicinity of the microwave. 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 [27] and RoboBrain2.0-Eculid-7B in VSIBench [29].



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 [27] and RoboBrain2.0-Eculid-7B in VSIBench [29].

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 [27] and RoboBrain2.0-Eculid-7B in SuperCLEVR [15].

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

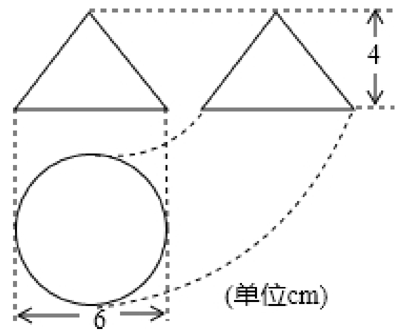
RoboBrain2.0 Answer: 2

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 [27] and RoboBrain2.0-Eculid-7B in SuperCLEVR [15].

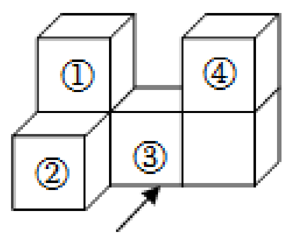
Problem: As shown in the figure, it is the three - view drawings of a part. Then the surface area of this part is?



A. $15\pi \mathrm{cm}^{2}$
B. $24\pi \mathrm{cm}^{2}$
C. $51\pi \mathrm{cm}^{2}$
D. $66\pi \mathrm{cm}^{2}$

Answer: B

Problem: As shown in the figure, six small cubes are placed as shown. If a small cube with a label is removed, and both its front view and top view change, the label of this small cube is?

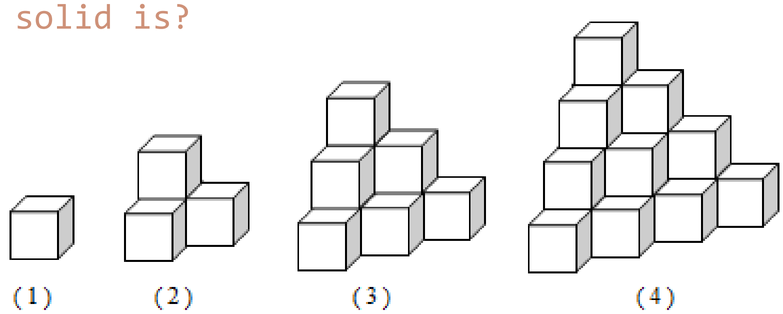


A. (1) B. (2) C. (3) D. (4)

Answer: C

Problem: The geometric solids in the figure are all stacked by cubes with a side length of 1. The surface area of the 1-th geometric solid is 6. Then the surface area of the 20-th geometric solid is?

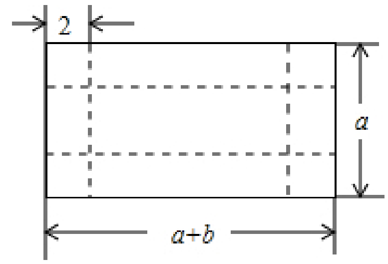
A. 1320 B. 1260
C. 1200 D. 1140



Answer: B

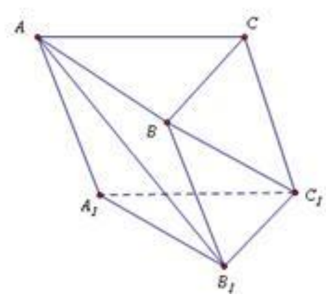
Figure 13. More examples from the Euclid30K dataset.

Problem: As shown in the figure, a rectangular iron sheet can be folded into an uncovered cuboid box after cutting off four corners. According to the data marked in the figure, find the base area of this box (unit: centimeter).



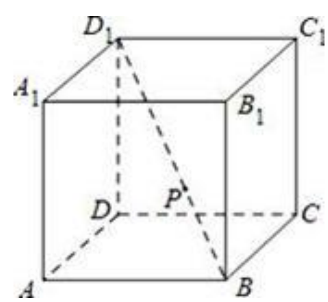
Answer: $a^2+ab - 8a - 4b + 16$

Problem: As shown in the figure, in the triangular prism $ABC - A_1B_1C_1$, both the side - lengths of the base and the length of the lateral edge are equal to 1, and angle BAA_1 = angle $ CAA_1 $ = 60°. Find the cosine value of the angle formed by the skew lines AB_1 and BC_1.",



Answer: $\sqrt{2}$

Problem: As shown in the figure, in the cube $ABCD - A_{1}B_{1}C_{1}D_{1}$, P is an equal - dividing point of the diagonal BD_{1} into three parts. The number of different values of the distances from P to each surface of the cube is?

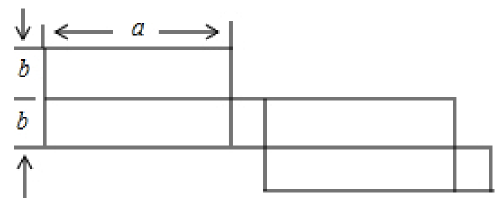


A. 2 pieces B. 3 pieces
C. 4 pieces D. 6 pieces

Answer: A

Figure 14. More examples from the Euclid30K dataset.

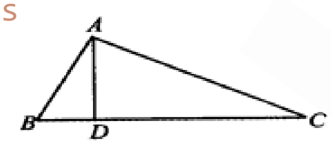
Problem: As shown in the figure, it is the lateral expansion diagram of a food packaging box. Please calculate the surface area of this packaging box according to the dimensions marked in the figure.",

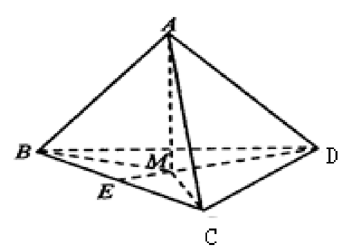


Answer: $2b^{2}+4ab$

Problem: As shown in the figure, in triangle ABC, $AB \perp AC$. If $AD \perp BC$, then $AB^{2}=BD \cdot BC$. Similarly, there is a proposition: In the triangular pyramid $A-BCD$, AD $\perp$ the plane ABC. If the projection of point A inside \triangle BCD is M, then $S_{triangle ABC}^{2}=S_{triangle BCM} \cdot S_{triangle BCD}$. The above - mentioned proposition is?

A. True proposition
B. Adding the condition “$AB \perp AC$” makes it a true proposition.
C. Adding the condition “M is the orthocenter of triangle BCD” makes it a true proposition.
D. Adding the condition "The triangular pyramid $A-BCD$ is a regular triangular pyramid" makes the proposition true."





Answer: A

Figure 15. More examples from the Euclid30K dataset.