PanoEnv: Supplementary Material

1. More Related Works

RL for Multimodal and Spatial Reasoning. Supervised fine-tuning (SFT) is the standard training method for MLLMs, but its reasoning processes can be rigid [1]. RL offers a promising alternative, as demonstrated by pioneering works [2, 7], such as Vision-R1 [4], which converts visual inputs into structured reasoning chains.

Recently, this RL-style reinforcement learning paradigm has been rapidly extended beyond 2D vision to 3D spatial understanding and embodied agent domains. Recent breakthroughs, including 3D-R1 [3], Robot-R1 [5], VLN-R1 [8], and Nav-R1 [6], have collectively demonstrated that RL fine-tuning significantly enhances complex geometric reasoning, physical interaction, and spatial navigation in embodied scenes.

In the panoramic VQA domain specifically, 360-R1 [10] also applies an RL framework with a reward function based on semantic similarity. However, previous methods rely heavily on other LLMs or human annotations to obtain answers or CoT, which inevitably introduces biases or hallucinations. The core innovation in our PanoEnv research is the use of geometric ground truth from TartanAir [9] to **programmatically generate correct textual answers**, which then serve as the basis for reward calculation. During RL training, the reward is based on the correspondence between the model's output and this physically-grounded ground truth text. This ensures the learning signal originates entirely from the objective physical world, providing a more direct and rigorous incentive for the model to learn true 3D spatial structures.

2. Dataset Construction Details

2.1. Object Filtering and Sampling Heuristics

Objective. Before question generation, we implement a rigorous object filtering and sampling pipeline to ensure the high quality and physical validity of the generated QA pairs, mitigating the geometric distortions inherent in Equirectangular Projection (ERP) images.

Ground Truth Formulation. The filtering mechanism operates on three hierarchical levels:

1. Size and Aspect Ratio Constraints. Objects are retained only if their 2D bounding box area in the ERP image satisfies Area ≥ 900 pixels, with both width ≥ 25 and height ≥ 25 pixels. Highly distorted or degenerate masks are removed by thresholding the aspect ratio to strictly < 5.

2. Semantic Exclusion. We systematically exclude background surfaces with low QA discrimination (e.g., *sky*, *ground*, *wall*), extremely thin objects prone to detection errors (e.g., *wire*), and unstable semantic debris (e.g., *rubble*).

3. Containment Filtering. For relational and comparative questions, we exclude object pairs with substantial bounding box overlap to prevent visual ambiguity. Two objects are deemed unsuitable for comparison if their bounding box overlap ratio exceeds 0.90.

2.2. Robust Depth Profiling

As mentioned in the main text, relying solely on a single median depth value can lead to significant errors for "thick" objects or objects exhibiting severe partial occlusion. Therefore, we utilize the extracted point cloud to compute a robust depth profile based on percentiles (p_n) and the Inter-Quartile Range ($IQR = p_{75} - p_{25}$).

Effective Depth for Comparison. An object is defined as having a significant physical thickness along the camera ray if its depth variation satisfies:

$$IQR > \max(0.6, 0.15 \cdot p_{50}) \tag{1}$$

For objects meeting this condition, using the median depth (p_{50}) may misrepresent their closest visible surface. Instead, we dynamically route the depth metric to the near-end quantile (p_{20}) to serve as the effective depth for distance-based QA comparisons.

Distance Similarity Thresholds. When generating "True/False" questions regarding whether two objects are at a similar distance, we evaluate both absolute median differences and interval overlaps. Two objects, A and B, are considered to be at a similar depth if the Jaccard similarity of their inter-quartile depth intervals exceeds 0.30:

$$\frac{|[p_{25}^A, p_{75}^A] \cap [p_{25}^B, p_{75}^B]|}{|[p_{25}^A, p_{75}^A] \cup [p_{25}^B, p_{75}^B]|} > 0.30 \tag{2}$$

Alternatively, they are considered similar if the absolute difference between their median depths is exceptionally small:

$$|p_{50}^A - p_{50}^B| < \max(0.5, 0.10 \cdot \min(p_{50}^A, p_{50}^B)) \tag{3}$$

2.3. Data Quality Assurance via Human Verification

To rigorously validate the semantic soundness of our programmatically generated ground truth, we conducted a human evaluation study on a randomly sampled subset of the PanoEnv-QA dataset. Independent human annotators verified the correctness of the generated answers against the visual evidence in the ERP images. The human verification yielded a **96% accuracy rate**, confirming that our physics-derived geometric ground truth is highly reliable, semantically sound, and strongly aligns with human perception of 3D spaces.

3. Prompt Designs for Generation and Evaluation

The design of the prompt is critical for both the GRPO exploration phase and the final model evaluation. In this section, we detail the unified templates used for eliciting model responses, as well as the strict scoring prompts utilized by the LLM-as-a-judge.

3.1. Generation and Training Prompts

To encourage the model to output transparent Chain-of-Thought (CoT) reasoning before providing the final answer, we strictly enforce a unified prompt structure across all baseline evaluations and our RL training.

The complete prompt is constructed dynamically by concatenating a Base Prompt with a Task-Specific Suffix depending on the question type.

Base Prompt (Shared across all tasks):

```
You are an expert in analyzing 360°
panoramic (ERP) images (2560x1280).
Analyze the image carefully and
focus on the specific objects
mentioned in bounding boxes.

{Question}

Provide your reasoning based on the
panoramic scene within <Reasoning>
tags, then give your final answer
within <Answer> tags.
```

Task-Specific Format Constraints:
To ensure robust parsing for our routed reward system, one of the following rule suffixes is appended to the base prompt based on the question category:

1. True/False Questions:

```
This is a YES/NO question.  Your
<Answer> must be EXACTLY "Yes"
or "No" (case-sensitive, no extra
words).
Example:
<Reasoning>...</Reasoning>
<Answer>Yes</Answer>
```

2. Multiple-Choice Questions:

```
This is a MULTIPLE CHOICE question.
Your <Answer> must be EXACTLY one of
the provided options.
Example:
<Reasoning>...</Reasoning>
<Answer>AbandonedCable</Answer>
```

3. Open-Ended Questions:

```
This is an OPEN-ENDED question.
Your <Answer> should be concise and
direct (under 20 words).
Example:
<Reasoning>...</Reasoning>
<Answer>The building is behind
and to the right of and below the
truck</Answer>
```

3.2. LLM-as-a-Judge Evaluation Prompts

In addition to the strict rule-based accuracy metrics, we employ an LLM-as-a-judge approach (e.g., QwenScore and PrometheusScore) to evaluate the semantic correctness and reasoning quality of the models' outputs. To ensure consistency and transparency, we use a highly structured prompt that explicitly defines the scoring rules (0-10) for each specific question type.

1. System Prompt (Evaluator Rules):

```
You are an evaluator.  Compare
answers and give a score 0-10.

Rules:

- YES/NO: yes/true/1 = YES,
no/false/0 = NO.
Same meaning -> 10, different -> 0.

- Multiple choice:  output a
single option only.  Ignore
articles/case/punctuation.
Match -> 10, else -> 0.

- Numeric (distance, e.g.  "About X
meters"):  extract numbers.
If no valid number -> 0.
* <=10% relative error -> 10
* <=20% -> 5-9
* >20% -> 0

- Numeric (counting, e.g.  "3
different views"):  use only the
```

```
integer.
* Exact match -> 10
* Otherwise -> 0

- Spatial: compare direction
words (left/right, front/behind,
above/below).
If any axis is opposite (e.g. left
vs right), score -> 0.
Otherwise, let N = axes used in
reference, C = correctly matched
axes; score ≈ 10 * C / N.

- Open-ended: judge semantic match.
* All key info correct -> 10
* Most info correct -> 8-9
* Partially correct -> 6-7
* Mostly wrong -> 0-5

IMPORTANT: End response with "Score:
X" where X is 0-10.
```

2. User Prompt (Evaluation Instance):
During evaluation, the test samples are dynamically formatted into the following template and passed to the LLM judge:

```
Q: {question}
Type: {question_type}
Reference: {expected_answer}
Answer: {model_answer}

Give score 0-10. Must end with
"Score: X".
```

4. Training Details and Hyperparameters

Hardware & Optimization. Models are trained on 4 $\times$ NVIDIA B200 GPUs with bfloat16, DeepSpeed ZeRO-2, and Flash Attention 2. For reproducibility, training is also feasible on 80GB A100 GPUs by halving the group size ($K = 2$) and per-device batch size. Hyperparameters are detailed in Table 1.

5. Comparison with Existing Benchmarks

To further clarify the unique positioning of PanoEnv-QA, we provide a detailed comparison with recent panoramic reasoning benchmarks, such as OSR-Bench and OmniVQA, in Table 2. Specifically, PanoEnv uniquely focuses on recovering 3D metric and physical relationships from 2D pixels using precise simulation engine ground truth.

Unlike OSR-Bench, which focuses on 2D topological relations and achieves its scale through extensive negative sampling, PanoEnv prioritizes reasoning depth over massive dataset scale. Our benchmark evaluates the implicit recovery of 3D physical reality (e.g., true volume and metric distance) from monocular panoramas. Furthermore, in contrast to OmniVQA which relies on LLM-generated answers, PanoEnv derives its rewards strictly from pixel-perfect geometric annotations provided by the simulation engine, thereby preventing the reinforcement of hallucinated reasoning.

Table 1. Hyperparameter configurations for GRPO fine-tuning.

Hyperparameter	**Value**
Base Model	Qwen2.5-VL-7B-Instruct
Tuning Method	LoRA (Language Decoder only)
LoRA Rank (r) / Alpha (α)	16 / 32
LoRA Dropout	0.05
Group Size (K)	4
Training Epochs	2
Peak Learning Rate	5e-6
Warmup Ratio	0.05
Global Batch Size	32
KL Coefficient (β)	0.01
Max Completion Length	256
Max Gradient Norm	10.0
Reward Weights (Acc, Fmt)	(0.90, 0.10)

Table 2. Comprehensive comparison between PanoEnv and existing panoramic benchmarks.

Feature	**PanoEnv (Ours)**	**OSR-Bench**	**OmniVQA**
Dataset Base	TartanAir	Replica / DeepPano	Stan. 2D-3D-S
Scenes	60 (Indoor & Outdoor)	1,526 (Indoor)	6 (Indoor)
Images	595	4,100	1,213
QA Pairs	∼14.8K	∼153K	∼4.9K
Core Focus	**2D-to-3D Inference**	Topology	Polar Distortion
GT Source	**3D Annotations**	Cognitive Map	MLLM + Human

6. Additional Experiments: Sim-to-Real Generalization

To validate the real-world transferability of PanoEnv-RL, we conducted a zero-shot evaluation on the **OSR-Bench**, a benchmark comprised of photorealistic panoramas derived from high-fidelity scene scans.

As shown in Table 3, despite being trained *solely* on simulation data (TartanAir), our PanoEnv-RL (7B) consistently outperforms the Base 7B model and even surpasses the significantly larger **Qwen2.5-VL-72B** on tasks such as *Object Counting* and *Relative Distance*.

Table 3. Zero-Shot Performance on OSR-Bench (Real-World).

Model	**Obj. Count**	**Rel. Dist.**	**Rel. Dir.**
Qwen2.5-VL-7B (Base)	0.477	0.321	0.089
Qwen2.5-VL-72B	0.498	0.325	**0.181**
PanoEnv-RL (Ours)	**0.507**	**0.371**	0.105

Transferability. The results exhibit the robust transferability of our method to real-world domains. The performance gap in *Relative Direction* compared to the 72B model is primarily attributed to a **2D-vs-3D mismatch**: OSR-Bench lacks verticality and evaluates flat 2D topological relations, which inherently penalizes our model's fine-grained 3D spherical predictions.

Logic over Scale. The ability to generalize directly to the real-world OSR-Bench without any domain-specific fine-tuning confirms that PanoEnv-RL acquires transferable **3D spatial reasoning** (i.e., geometric logic) rather than merely memorizing simulation data or textures. PanoEnv provides the exact physical ground truth essential to encode this reasoning core, successfully enabling robust sim-to-real transfer and proving that geometric logic can bridge the reality gap more effectively than simply scaling model parameters.

References

[1] Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. How abilities in large language models are affected by supervised fine-tuning data composition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, 2024.

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[3] Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*, 2025.

[4] Wenxuan Huang, Bohan Jia, Zijie Zhai, Shaosheng Cao, Zheyu Ye, Fei Zhao, Zhe Xu, Xu Tang, Yao Hu, and Shaohui Lin. Vision-r1: Incentivizing reasoning capability in multimodal large language models. *arXiv preprint arXiv:2503.06749*, 2025.

[5] Dongyoung Kim, Sumin Park, Huiwon Jang, Jinwoo Shin, Jaehyung Kim, and Younggyo Seo. Robot-r1: Reinforcement learning for enhanced embodied reasoning in robotics. *arXiv preprint arXiv:2506.00070*, 2025.

[6] Qingxiang Liu, Ting Huang, Zeyu Zhang, and Hao Tang. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*, 2025.

[7] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[8] Zhangyang Qi, Zhixiong Zhang, Yizhou Yu, Jiaqi Wang, and Hengshuang Zhao. Vln-r1: Vision-language navigation via reinforcement fine-tuning. *arXiv preprint arXiv:2506.17221*, 2025.

[9] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916. IEEE, 2020.

[10] Xinshen Zhang, Zhen Ye, and Xu Zheng. Towards omnidirectional reasoning with 360-r1: A dataset, benchmark, and grpo-based method. *arXiv preprint arXiv:2505.14197*, 2025.