SpaceMind: Camera-Guided Modality Fusion for Spatial Reasoning in Vision-Language Models

Supplementary Material

6. Design Philosophy

We chose an end-to-end latent fusion approach over popular explicit reconstruction pipelines because latent representations are fundamentally more robust and efficient for machine reasoning. While explicit 3D representations, such as point clouds or detailed meshes, are highly intuitive for human perception, they often introduce unnecessary complexities and rigid bottlenecks for automated systems. By prioritizing logical utility over sheer visual geometric fidelity, our latent-based design ensures that the model focuses purely on extracting the essential semantic spatial cues necessary for accurate physical understanding. The core logic behind this architectural choice is detailed below.

6.1. Avoiding the "Hard Decision" Trap

Explicit pipelines force early, *hard decisions*: errors from depth estimators or Structure-from-Motion (SfM) algorithms become permanently baked into the geometry, forcing the model to rely on flawed data. SpaceMind instead operates on *soft, probabilistic* features. By passing high-dimensional features rather than absolute coordinates, the language model treats geometric cues as evidence, leveraging semantic context to resolve low-level ambiguities.

6.2. Optimizing for Reasoning

Standard 3D reconstructions optimize for visual *geometric fidelity*, which often retains irrelevant details while missing crucial abstract spatial and physical relations. SpaceMind's latent approach instead aligns features directly with the downstream reasoning task, flexibly learning to extract critical spatial cues like relative positioning and occlusions rather than wasting model capacity on pixel-perfect reconstructions.

6.3. Handling Challenging Visuals

Explicit pipelines struggle with mirrors or transparent objects, which easily break geometric projections. SpaceMind addresses this by mapping semantics to geometry via structural grid locations rather than pure depth. This structural alignment holds firm even when depth signals are corrupted, keeping visual and spatial data securely connected and preventing misalignment.

6.4. Preserving Dense Details

Converting images to point clouds inevitably compresses rich visual textures into sparse data, fundamentally losing contextual depth. SpaceMind's latent embeddings, however, maintain a dense information manifold, retaining the subtle, high-dimensional nuances of the scene required for complex spatial logic.

7. Scalability and Future Potential

SpaceMind establishes a strong baseline with an extensible architecture that is well poised for future growth across three key avenues.

7.1. Scaling Spatial Evidence

It is a widely accepted consensus that expanding spatial context significantly enhances 3D comprehension. Aligning with this principle, increasing the number of input views yields immediate performance gains for SpaceMind. The inherently scalable architecture is primed to ingest larger context windows and denser video sampling for finer-grained understanding without requiring any redesign.

7.2. Leveraging Upstream Encoders

Because SpaceMind operates in tandem with spatial encoders like VGGT, our reasoning capabilities will naturally scale with advancements in geometric deep learning. As foundation models improve, SpaceMind's modular design allows for the seamless integration of stronger backbones, automatically lifting overall reasoning performance.

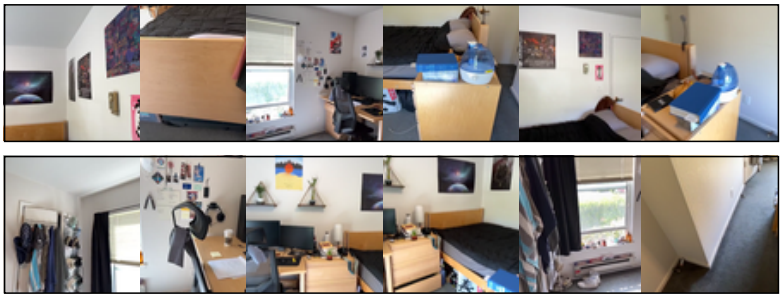
7.3. End-to-End Tuning

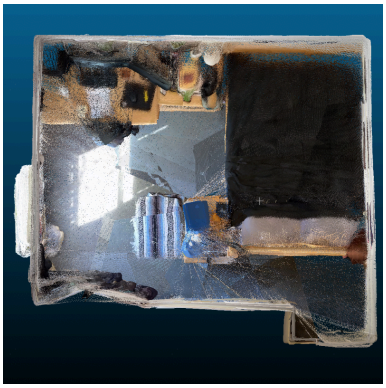
While we currently freeze the spatial encoder to retain pretrained priors, future iterations will explore end-to-end optimization. Unfreezing these layers would allow the reasoning loss to directly influence geometric extraction, transitioning the system from general-purpose reconstruction to highly specialized spatial logic extraction.

8. Qualitative Visualizations

To concretely demonstrate how SpaceMind bridges the gap between raw pixel inputs and complex spatial understanding, we provide four qualitative visual Question-Answering (QA) example visualizations across our primary evaluation benchmarks. Specifically, the selected VSI-Bench examples highlight SpaceMind's proficiency in challenging real-world tasks, such as object absolute distance and overall room-size estimation. By grounding abstract geometry in observable semantic contexts, these examples confirm that our latent fusion mechanism effectively translates visual features into accurate physical reasoning.

object_abs_distance





Question: Measuring from the closest point of each object, what is the distance between the ceiling light and the bed (in meters)?

Ground Truth: 1.4

Spatial-MLLM: 0.8
score: 0.2
VLM-3R: 1.2
score: 0.8
SpaceMind: 1.4
score: 1.0

Figure 4. Qualitative example on VSI-Bench.

object_abs_distance

Question: Measuring from the closest point of each object, what is the distance between the refrigerator and the chair (in meters)?

Ground Truth: 1.1

Spatial-MLLM: 1.9
score: 0.0
VLM-3R: 1.4
score: 0.5
SpaceMind: 1.1
score: 1.0

Figure 5. Qualitative example on VSI-Bench.

room_size_estimation

Question: What is the size of this room (in square meters)?
If multiple rooms are shown, estimate the size of the combined space.

Ground Truth: 10.1

Spatial-MLLM: 20.9
score: 0.0
VLM-3R: 14.2
score: 0.2
SpaceMind: 12.0
score: 0.7

Figure 6. Qualitative example on VSI-Bench.

room_size_estimation

Question: What is the size of this room (in square meters)?
If multiple rooms are shown, estimate the size of the combined space.

Ground Truth: 8.2

Spatial-MLLM: 15.9
score: 0.0
VLM-3R: 12.3
score: 0.0
SpaceMind: 8.8
score: 0.9

Figure 7. Qualitative example on VSI-Bench.