GeoWorld-VLM: Sequential 3D Generation via Evidential Memory

Renjie Gu*, Kaichen Zhou*, Yan Luo, Mengyu Wang
Harvard AI and Robotics Lab
Kempner Institute for the Study of Natural and Artificial Intelligence
Harvard University
*Equal contribution as co-first authors

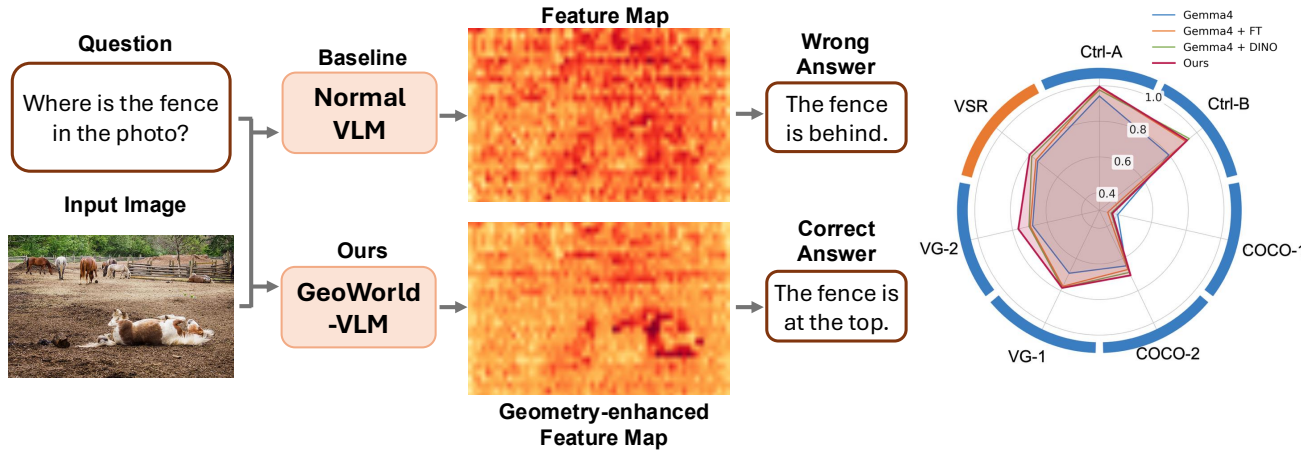


Figure 1. **Overview.** Given an input image and a spatial reasoning question, GeoWorld-VLM enhances the spatial understanding of standard vision-language models by injecting world-model priors at the feature-map level. Compared with the original VLM features, GeoWorld-VLM produces more geometry-aware representations, leading to clearer spatial grounding and improved answer accuracy. As shown on the right, our method consistently outperforms strong baselines, including the original Gemma-4, fine-tuned Gemma, and fine-tuned Gemma with DINO features, across diverse spatial reasoning benchmarks such as What'sUp and VSR.

Abstract

Modern Vision-Language Models (VLMs) achieve strong semantic recognition, yet remain brittle on elementary spatial relations such as left of, on, behind, and between. One cause of this failure arises before language reasoning begins: the visual pathway may compress or discard critical 3D structural cues during feature extraction, so the language model receives image representations that are already insufficient for reliable spatial judgment. We introduce GeoWorld-VLM, a VLM-side distillation framework that transfers geometric structure from frozen camera-conditioned video world models into VLMs. GeoWorld-VLM fine-tunes only the image encoder and multimodal projector, aligning post-projector image features with intermediate world-model representations while leaving the main backbone frozen. Given images, a prompt, and a sampled camera trajectory, the world-model teacher converts static visual input into a synthetic multi-view spatial signal. Training combines spatial answer supervision, teacher-student feature alignment, and a preservation anchor to the original VLM. Since the language model remains frozen, GeoWorld-VLM preserves the original model's linguistic capabilities while attributing spatial improvements to the enhanced visual pathway. To evaluate the effectiveness and generality of the proposed method, we apply GeoWorld-VLM to two distinct VLM architectures and observe consistent improvements across both backbones. GeoWorld-VLM improves performance by approximately 4% on both the What'sUp and VSR benchmarks, suggesting that world-model-guided visual alignment generalizes across model structures and spatial reasoning datasets. Submitted to Neurips 2026.

1. Introduction

Vision-Language Models (VLMs) have become fluent semantic recognizers, but they remain unreliable spatial reasoners [29, 48]. Even strong contemporary VLMs struggle with

elementary relations such as *left of*, *on top of*, *behind*, and *between*. These failures are surprising because the required reasoning is not linguistic: the model often knows the object names and understands the question, yet still fails to recover the geometric relation. This suggests that the bottleneck lies not in language reasoning alone, but in the visual representation delivered to the language model.

A modern VLM can be viewed as a composition of three modules [18, 30, 50]:

$$\mathrm{VLM}(x_v, x_t) = \mathrm{LLM}\left([P_\theta(V_\phi(x_v)), E(x_t)]\right), \quad (1)$$

where V_ϕ is the vision encoder, P_θ is the multimodal projector, E is the text encoder and the LLM performs language-conditioned reasoning. This decomposition exposes three possible places to repair spatial reasoning: the LLM, the text encoder, or the vision encoder & projector. In this paper, we argue that improving the spatial understanding capability of the vision encoder and multimodal projector can effectively enhance VLM performance on spatial intelligence questions.

The key challenge is how to teach the visual pathway geometric structure without relying on real multi-view supervision. Recent world models have demonstrated strong spatial understanding and prediction capabilities across diverse scenarios with sparse observation [41, 42, 52]. As shown in Figure 1, motivated by this observation, we introduce **GeoWorld-VLM**, a VLM-side distillation framework that transfers spatial structure from frozen camera-conditioned world models into VLMs. Given images, a text prompt, and a sampled egocentric camera trajectory, the world-model teacher produces an intermediate representation that captures how the scene would evolve under viewpoint motion, offering motion-aware geometric cues that are difficult to recover from the static view alone. These cues provide an implicit signal about object layout, relative position, and view-dependent structure, which are often compressed or weakened in standard VLM visual representations. GeoWorld-VLM distills this signal into the VLM by aligning its post-projector image features with the teacher representation in a shared feature space, thereby encouraging the visual pathway to preserve motion-aware spatial structure and scene geometry before these visual tokens are passed to the language model for reasoning. Crucially, GeoWorld-VLM leaves the language model frozen. Within the VLM, we update only the image encoder and multimodal projector, together with lightweight training-time alignment heads. Training combines spatial answer supervision, teacher-student feature alignment, and a preservation anchor to a frozen copy of the original VLM. This design isolates spatial improvement to the visual interface: text-only behavior is preserved by construction, while the visual tokens entering the LLM become more geometry-aware.

To evaluate the effectiveness and generality of the proposed method, we instantiate our plug-in framework with two representative VLM backbones, Gemma [38] and InternVL3.5 [46], and evaluate them on the What'sUp and VSR benchmarks. Across both architectures and datasets, our method achieves consistent improvements under world-model supervision. Ablation studies further show that supervision from camera-conditioned world models is more effective than using conventional vision foundation models such as DINO [7] or geometry foundation models such as VGGT [43].

- **World-model distillation for VLM spatial reasoning.** We introduce GeoWorld-VLM, a framework that uses frozen camera-conditioned video world models as geometry teachers for VLMs. By sampling camera trajectories from input images, the teacher provides synthetic multi-view supervision without requiring real multi-view data.
- **A vision-side repair that preserves the language model.** GeoWorld-VLM fine-tunes only the vision encoder and multimodal projector while keeping the LLM and language encoder frozen. This isolates the spatial intervention to the visual interface and preserves text-only behavior by construction.
- **Broad validation across spatial and embodied benchmarks.** We evaluate GeoWorld-VLM across multiple spatial reasoning benchmarks and VLM backbones, demonstrating consistent improvements over the base models and stronger spatial understanding.

2. Related Work

2.1. Spatial and 3D Reasoning in Vision-Language Models

Modern vision-language models such as GPT-4o [19], GPT-5 [31], Gemini [15], Claude [2], Qwen-VL [3], Qwen2.5-VL [4], Qwen3-VL [33], LLaVA [28], BLIP-2 [24], InstructBLIP [12], PaLI [9], and PaliGemma [5] have shown strong semantic visual understanding, but they remain limited in spatial and 3D reasoning. Benchmarks such as SAT [35], What's Up [20], BLINK [14], EmbSpatial-Bench [13], and PhysBench [11] reveal persistent failures on relative position, relative depth, perspective taking, ego-motion, object movement, and physical consequence prediction. These failures suggest that VLMs can often recognize objects and parse questions, yet still lack the geometric structure needed to reason about spatial relations. Several works attempt to bridge this gap by injecting 3D representations into language or vision-language models. 3D-LLM [17] introduces 3D scene features into language models for embodied reasoning. Cube-LLM [10] and Language-Image Models with 3D Understanding explore 3D-aware representations for multimodal reasoning. MiniGPT-3D [37] aligns point clouds with language models using 2D priors, while ShapeLLM focuses on 3D object understanding through point-cloud encoders and 3D instruction tuning. 3UR-LLM [49] further extends multimodal

language models toward 3D scene understanding. These approaches show that explicit 3D information can improve spatial reasoning, but typically require 3D supervision, point-cloud inputs, specialized architectures, or additional training. Training-free or test-time approaches provide a more flexible alternative. APC [23] uses abstract perspective changes to support perspective-aware reasoning. VAGEN [44] applies multi-turn reasoning and world-model interaction for spatial intelligence. SandboxVLM [29] constructs abstract 3D bounding boxes through proxy elevation, multi-view voting, and box rendering, showing that coarse structural representations [43, 56] can help VLMs reason spatially without retraining. Our work follows the same broad goal of improving VLM spatial intelligence, but differs in where the improvement is applied. Rather than providing external 3D context at test time, GeoWorld-VLM distills geometry into the VLM's own vision side, improving the visual tokens consumed by the frozen language model.

2.2. World Models and Geometric Supervision for VLMs

World models and video generative models provide a natural source of geometric supervision because they could predict how scenes evolve under time, camera motion, and object interaction [41, 42, 52]. Classical world-model approaches such as Dreamer [47], DreamerV2 [54] and DreamerV3 [16] learn predictive latent dynamics for control and planning. More recent generative world models and video models, including Genie [6], VideoPoet [21], Sora-style video models [25], CogVideoX [52], Wan [42], HunyuanVideo [22], Cosmos [1], and Stable Virtual Camera [55], scale this idea to rich visual scenes and viewpoint-conditioned generation. These models implicitly encode cues such as parallax, occlusion, support, relative depth, and object persistence. Recent VLM reasoning methods have begun to exploit such models. MindJourney [51] uses imagined visual trajectories at test time to help VLMs reason about spatial questions. These works suggest that world models contain useful spatial priors, but most use them externally: they generate images, videos, or reconstructed 3D contexts that are then fed back to a VLM during inference. GeoWorld-VLM instead uses the world model as a feature teacher. The teacher receives the initial image, a text condition, and a sampled camera trajectory, producing an intermediate representation shaped by camera-conditioned prediction. We distill this representation into the VLM's post-projector image feature, while keeping the LLM frozen. This differs from static feature distillation methods based on DINO [7], DINOv2 [32], CLIP [34], or SigLIP [53], which provide strong semantic representations but do not explicitly encode viewpoint-dependent geometry. It also differs from world-model-only image conditioning, where the teacher receives no camera trajectory and therefore behaves more like a static image encoder. By conditioning the world-model teacher on both camera motion and textual context, GeoWorld-VLM transforms a single static image into a synthetic multi-view supervision signal that encodes how the scene would change under viewpoint variation, and transfers this geometry-aware information into the VLM's visual interface through feature-level alignment.

3. Method

We introduce **GeoWorld-VLM**, a vision-side distillation method that transfers geometric structure from a frozen camera-conditioned video world model into a VLM [5, 39, 46]. As shown in Figure 2, our design follows from a simple hypothesis: spatial failures arise because the visual tokens entering the LLM lack sufficient 3D structure. We therefore keep the LLM frozen and update only the image encoder, multimodal projector, and lightweight alignment heads.

3.1. Vision-side formulation

Following most VLM architectures, we design a VLM:

$$\mathrm{VLM}(x_v, x_t) = \Psi\left([P_\theta(V_\phi(x_v)), E(x_t)]\right), \tag{2}$$

where V_ϕ is the image encoder, P_θ is the multimodal projector, E is the text encoder and Ψ is the main LLM backbone. GeoWorld-VLM updates only (ϕ, θ). This isolates the intervention to the representation consumed by the LLM: text-only behavior is unchanged by construction, and any spatial improvement must come from better visual tokens rather than language-side adaptation.

3.2. Camera-conditioned world-model teacher

Static image teachers such as DINO [7, 32, 36] provide strong semantic features, but they do not explicitly encode how a scene changes under viewpoint motion. This distinction matters for spatial reasoning: relations such as behind, between, support, and occlusion are revealed not only by what objects are present, but by how their projections change as the camera moves. For each training image x_v, we sample an egocentric camera trajectory:

$$\pi = (\pi^{(1)}, \ldots, \pi^{(F)}), \tag{3}$$

from translations and rotations, including forward, backward, left, right, forward-left, backward-left, forward-right, backward-right, and other movements, as provided in LingBot-World [41]. The frozen world-model teacher receives the initial images, the camera trajectory, and a text condition: (x_v, π, c). Here, c is "A slight camera motion with stable object layout and unchanged spatial relations.". The teacher predicts the scene under the sampled camera motion, and we extract an intermediate hidden representation $g(x_v) = \mathcal{T}_{b^\star}(x_v, \pi, c, z_{t^\star})$ where $z_{t^\star}$ is the noised video latent at timestep $t^\star$ and $\mathcal{T}_{b^\star}$ denotes the hidden state at teacher block $b^\star$. The camera trajectory is essential. If the teacher

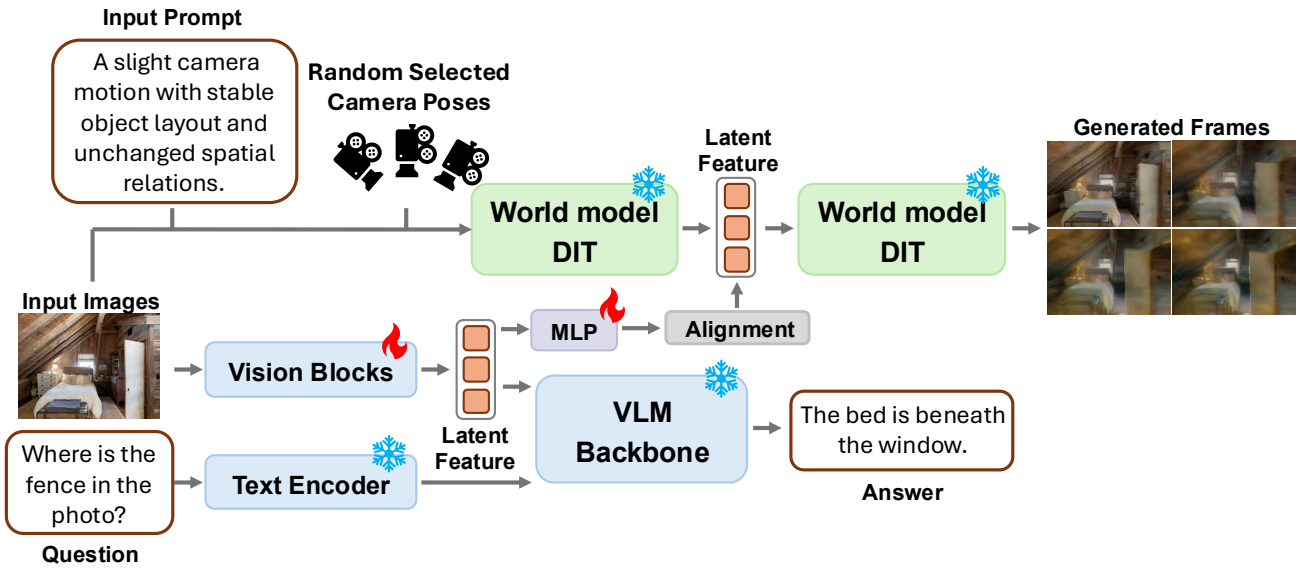


Figure 2. **GeoWorld-VLM.** During training, GeoWorld-VLM fine-tunes only the vision blocks including vision encoder and multimodal projector. It aligns the latent features produced by the VLM vision encoder with intermediate world-model representations, where the world model takes the input image, text prompt, and randomly sampled camera poses as input. At inference time, GeoWorld-VLM no longer requires the world model and can perform standard VLM inference directly.

receives only images, its representation collapses toward a static semantic embedding. Conditioning on camera motion instead forces the teacher feature to encode counterfactual views, making a single image behave like a synthetic multi-view training signal, which could be helpful for spatial understanding.

3.3. World-to-VLM alignment

The student produces post-projector image features $h_{\phi,\theta}(x_v) = P_\theta(V_\phi(x_v))$. Since the VLM and world model have different representation spaces, we use two lightweight projection heads:

$$u_s = f_\psi^s(h_{\phi,\theta}(x_v)), \qquad u_t = f_\psi^t(g(x_v)). \tag{4}$$

The alignment loss is

$$\mathcal{L}_{\text{align}} = 1 - \left\langle \frac{u_s}{\|u_s\|_2}, \frac{u_t}{\|u_t\|_2} \right\rangle. \tag{5}$$

We align at the post-projector level because this representation is directly consumed by the frozen LLM for multimodal reasoning. Aligning earlier visual features may improve the encoder but does not guarantee that geometry survives the multimodal interface.

3.4. Preserving the original interface

World-model alignment alone can move the visual tokens away from the distribution expected by the frozen LLM. To control this shift, we keep a frozen copy of the original vision side, $(V_{\phi'}, P_{\theta'})$, and add a post-projector preservation loss:

$$\mathcal{L}_{\text{preserve}} = \left\| \frac{h_{\phi,\theta}}{\|h_{\phi,\theta}\|_2} - \frac{h_{\phi',\theta'}}{\|h_{\phi',\theta'}\|_2} \right\|_2^2. \tag{6}$$

The alignment and preservation losses play complementary roles: alignment injects geometry from the world model, while preservation keeps visual tokens compatible with the frozen language stack.

3.5. Training objective

GeoWorld-VLM combines spatial supervision, world-model alignment, and interface preservation:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{preserve}}\mathcal{L}_{\text{preserve}}. \tag{7}$$

Here, $\mathcal{L}_{\text{task}}$ is the option-restricted cross-entropy loss for spatial multiple-choice supervision. Gradients update only (ϕ, θ, ψ); the LLM, the world-model teacher, and the frozen reference VLM remain fixed. This design separates the source of supervision from the locus of adaptation. The world model supplies geometry through camera-conditioned prediction; the VLM absorbs that structure only through its vision side; and the LLM remains unchanged. As a result, GeoWorld-VLM tests a focused claim: many spatial reasoning failures in VLMs can be repaired by improving the visual interface itself rather than retraining or modifying the language model.

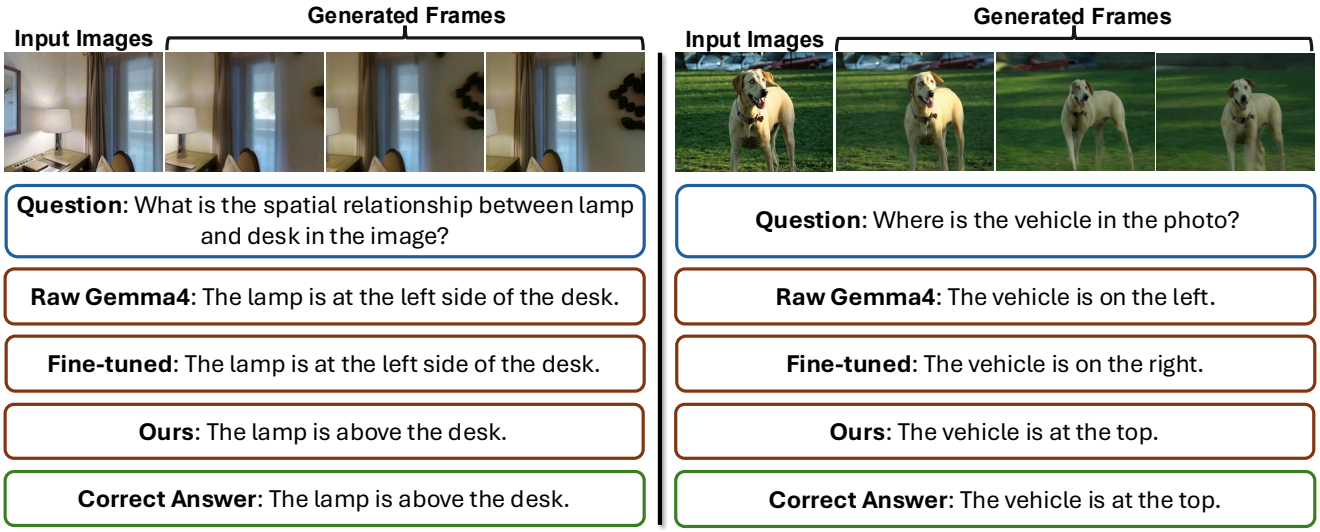


Figure 3. **Qualitative comparison.** We compare the predictions of the original Gemma4, task-only fine-tuned Gemma4, and GeoWorld-VLM on representative spatial reasoning examples. GeoWorld-VLM produces more spatially consistent answers than the baselines, illustrating the benefit of injecting camera-conditioned world-model supervision into the VLM visual pathway.

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4	94.17	79.90	**40.57**	64.55	69.14	68.49	74.36	61.70
Gemma4 + FT	99.03	92.16	34.79	66.82	77.24	69.86	75.56	62.71
Gemma4 + DINO	97.57	**94.61**	36.74	68.64	78.10	70.55	78.63	64.40
Gemma4 + Ours	**99.51**	93.14	37.46	**70.45**	**78.28**	**76.71**	**80.17**	**65.45**
InternVL3.5-2B	64.08	81.37	**40.12**	60.00	65.69	41.78	72.82	57.06
InternVL3.5-2B + FT	98.54	96.08	31.76	55.00	61.38	41.10	**83.59**	58.14
InternVL3.5-2B + DINO	97.57	**99.51**	29.54	52.73	51.38	29.45	83.25	54.81
InternVL3.5-2B + Ours	**99.51**	**99.51**	35.77	**60.91**	**70.52**	**45.89**	80.34	**61.66**

Table 1. Main results on the combined What'sUp+VSR evaluation suite. GeoWorld-VLM consistently improves the overall score over the original model, task-only fine-tuning, and DINO-based distillation. Best results are shown in bold, and second-best results are underlined.

4. Experiments

4.1. Experimental Setup

Baselines & VLMs. We evaluate GeoWorld-VLM on two representative vision-language model backbones: `google/gemma-4-E4B-it` and `OpenGVLab/InternVL3_5-2B-Instruct` [45]. For each VLM, we compare four variants. *Base* denotes the original VLM without additional spatial adaptation. *FT* denotes task-only fine-tuning, which removes the teacher-student feature alignment loss and the preservation anchor from GeoWorld-VLM and optimizes only the downstream spatial answer supervision. This baseline tests whether the gains can be explained by target-task supervision alone. *DINO* [36] denotes a DINOv3-based static feature distillation baseline, where the world-model teacher is replaced with a frozen DINOv3 encoder and the VLM visual representation is aligned with DINOv3 features. This baseline tests whether generic visual representation distillation is sufficient for spatial reasoning. *Ours* denotes GeoWorld-VLM, which aligns post-projector image features with intermediate camera-conditioned world-model representations, allowing the student VLM to absorb motion-aware geometric cues from the frozen teacher while preserving compatibility with the original visual interface and leaving the language backbone unchanged.

Datasets. We evaluate GeoWorld-VLM on three spatial reasoning benchmarks: What'sUp [20], Visual Spatial Reasoning (VSR) [27], and EmbSpatial-Bench [13]. Following the evaluation protocol of prior spatial reasoning work [8], we combine What'sUp and VSR into a unified What'sUp+VSR suite. This suite contains controlled spatial relation recognition, natural-image spatial QA, and image-text spatial verifi-

cation. Specifically, it covers object-centric relation recognition in controlled tabletop scenes, absolute and relative spatial localization in COCO and Visual Genome images, and binary verification of diverse spatial predicates in natural images. We use 3,089 examples for training and the remaining 3,091 examples for evaluation. For EmbSpatial, we use 2,000 examples for training and the remaining 1,640 examples for evaluation. For What'sUp+VSR, we report accuracy on Controlled Images A/B, COCO-based QA, VG-based QA, and VSR, together with the overall average [26, 57]. For EmbSpatial, we report relation-wise accuracy on six spatial relations: *left*, *right*, *above*, *under*, *close*, and *far*, together with the overall average.

Implementation Details. GeoWorld-VLM uses LingBot-World-Fast [40] as a frozen camera-conditioned image-to-video teacher, extracting intermediate DiT features as geometry-aware supervision. For each input image, the teacher takes a fixed lightweight prompt, "A slight camera motion with stable object layout and unchanged spatial relations.", and a randomly sampled egocentric camera trajectory. On the student side, we align post-projector image features, which are directly consumed by the frozen language model. Following Section 3, we freeze the language model and update only the image encoder, multimodal projector, and lightweight alignment heads. All main fine-tuning experiments are repeated using three different random seeds, and we report the averaged results across these independent runs. Unless otherwise specified, teacher features are extracted from the 24th teacher block using two denoising steps. All methods are trained for three epochs under the same downstream spatial supervision, with detailed hyperparameters provided in Appendix B.

4.2. Main Results on What'sUp and VSR

Table 1 reports the main results on the What'sUp+VSR suite. Across both VLM backbones, GeoWorld-VLM achieves the best overall performance. For Gemma4, GeoWorld-VLM improves the overall score from 61.70 to 65.45, outperforming both task-only fine-tuning and DINOv3-based static feature distillation. The gains are especially clear on natural-image spatial reasoning subsets such as COCO-2 and VG-2, where GeoWorld-VLM reaches 70.45 and 76.71, respectively. This suggests that world-model supervision does not merely improve controlled relation recognition, but also transfers to more complex natural-image spatial reasoning. The qualitative examples in Figure 3 further show that GeoWorld-VLM produces more spatially grounded predictions than the baselines.

For InternVL3.5-2B, GeoWorld-VLM improves the overall score from 57.06 to 61.66, yielding a 4.60-point gain over the base model and a 3.52-point gain over task-only fine-tuning. In contrast, DINOv3-based feature distillation decreases the overall score to 54.81. This comparison is important because both DINO and GeoWorld-VLM introduce auxiliary feature-level supervision, but only GeoWorld-VLM provides camera-conditioned world-model representations. The result indicates that static visual feature distillation may be insufficient, or even harmful, when the target task requires spatial judgment rather than generic visual discrimination. By contrast, world-model features provide a more geometry-aware supervision signal, while the preservation anchor helps maintain compatibility with the frozen language model. We also observe that the gains are not uniform across all subsets. For example, COCO-1 remains challenging for all adapted variants, and task-only fine-tuning can even reduce performance on this subset. This suggests that spatial adaptation may shift the visual interface toward the training distribution if it is not properly regularized. GeoWorld-VLM alleviates this issue by combining world-model supervision with the preservation anchor, improving the overall score while maintaining competitive performance across individual subsets. Figure 4 visualizes the overall comparison on the What'sUp+VSR suite. The performance pattern is consistent with Table 1: GeoWorld-VLM improves the visual pathway with geometry-aware supervision and achieves stronger spatial grounding without modifying the frozen language model.

4.3. Relation-wise Results on EmbSpatial

Table 2 further evaluates GeoWorld-VLM on EmbSpatial. This benchmark allows us to inspect whether the improvement comes from broad answer-pattern fitting or from better handling of specific spatial relations. For Gemma4, GeoWorld-VLM improves the overall score from 60.55 to 78.78, outperforming task-only fine-tuning and DINOv3-based static feature distillation. The gains are particularly clear on *above*, *under*, *close*, and *far*, which often require relative layout, support, depth, or distance estimation rather than object recognition alone.

For InternVL3.5-2B, GeoWorld-VLM improves the overall score from 52.68 to 79.70 and slightly surpasses DINOv3-based feature distillation. The strongest gains appear on *under* and *close*, suggesting that the world-model teacher is especially useful for relations that depend on scene geometry and viewpoint. Overall, the EmbSpatial results show that GeoWorld-VLM does not simply increase average accuracy; it strengthens spatial judgments that require geometry-aware visual representations. The relation-wise pattern also explains why the improvement is not identical across all categories. Relations such as *left* and *right* can often be inferred from 2D image layout, so static visual features or task-only fine-tuning may already provide strong cues. In contrast, relations such as *under*, *close*, and *far* are more sensitive to viewpoint, support, and relative depth. These are precisely the cases where camera-conditioned world-model supervision is expected to be most useful.

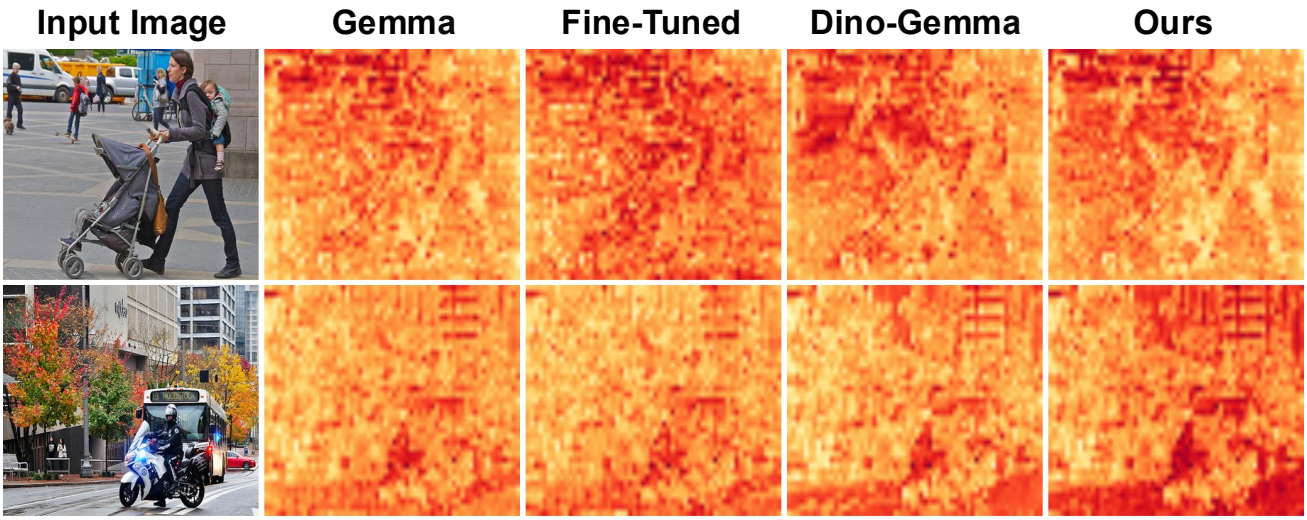


Figure 4. **Overall comparison on the What'sUp+VSR suite.** GeoWorld-VLM improves spatial reasoning performance across diverse sub-benchmarks, showing consistent gains over the original VLM, task-only fine-tuning, and DINOv3-based static feature distillation.

Model	Left	Right	Above	Under	Close	Far	Overall
Gemma4	84.53	60.57	60.45	55.35	53.99	47.76	60.55
Gemma4 + FT	**90.29**	77.78	86.57	85.24	63.04	58.96	77.01
Gemma4 + DINO	87.05	71.68	87.69	84.50	65.58	58.96	75.91
Gemma4 + Ours	89.21	**78.14**	**88.43**	**86.72**	**67.03**	**63.06**	**78.78**
InternVL3.5-2B	78.42	58.42	64.55	28.78	43.12	42.16	52.68
InternVL3.5-2B + FT	86.69	**82.80**	82.84	82.29	65.58	**68.66**	78.17
InternVL3.5-2B + DINO	**87.41**	78.85	**90.67**	84.50	67.75	65.30	79.09
InternVL3.5-2B + Ours	84.89	81.36	89.18	**87.82**	**69.93**	64.93	**79.70**

Table 2. Relation-wise results on EmbSpatial. GeoWorld-VLM improves the overall relation accuracy for both VLM backbones and shows strong gains on geometry-sensitive relations such as *under*, *close*, and *far*. The best result is shown in bold and the second-best result is underlined.

4.4. Module Selection

Table 3 studies which teacher-side conditioning module provides the most effective supervision for GeoWorld-VLM. Using image-to-video world-model features alone already improves the overall score from 61.70 to 64.01, showing that world-model representations contain spatial information beyond the original VLM visual tokens. However, adding only camera conditioning or only prompt conditioning does not further improve the overall result. The camera-only and prompt-only variants reach 63.16 and 62.61, respectively. This indicates that the two conditioning signals are complementary rather than independently sufficient. Camera trajectories provide geometric transformation cues, but without

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4	94.17	79.90	**40.57**	64.55	69.14	68.49	74.36	61.70
Gemma4 + I2V	99.03	**94.12**	35.68	68.18	77.59	67.81	79.66	64.01
Gemma4 + I2V + Camera	98.06	93.14	33.81	66.36	77.07	66.44	**81.03**	63.16
Gemma4 + I2V + Prompt	98.54	93.14	33.54	65.45	76.38	70.55	78.46	62.61
Gemma4 + I2V + Prompt + Camera	**99.51**	93.14	37.46	**70.45**	**78.28**	**76.71**	80.17	**65.45**

Table 3. Module selection results on the What'sUp+VSR suite. We compare different teacher-side conditioning modules, including image-to-video features alone and variants with prompt or camera conditioning. Based on the overall performance, we select image-to-video features with both prompt and camera conditioning as the default configuration of GeoWorld-VLM. The best result is shown in bold and the second-best result is underlined.

relation-aware text conditioning, the teacher representation may not focus on the spatial relation queried by the downstream task. Conversely, prompt conditioning alone identifies the relevant relation but lacks the viewpoint variation needed to expose geometry-aware structure.

The best performance is obtained when prompt and camera conditioning are used jointly. This full configuration improves the overall score to 65.45, outperforming the base model by 3.75 points and the I2V-only teacher by 1.44 points. The improvement is especially clear on COCO-2 and VG-2, where the full configuration reaches 70.45 and 76.71, respectively. These results suggest that the most useful teacher representation is both *question-aware* and *geometry-aware*: the prompt specifies the spatial relation of interest, while the camera trajectory encourages features that encode scene structure under viewpoint changes. Therefore, we use I2V features with both prompt and camera conditioning as the default teacher configuration of GeoWorld-VLM. Beyond module selection, we further conduct parameter ablations on world-model feature extraction layer, denoising step, and the alignment loss coefficient. As shown in Appendix C, GeoWorld-VLM achieves the best overall performance when using the 24th world-model layer, two denoising steps, and $\lambda_{\text{align}} = 0.10$. These results suggest that world-model supervision is most effective when the teacher features are sufficiently spatially structured but not overly dominated by generation-specific dynamics.

4.5. Ablation Studies

After selecting the default teacher-side conditioning module, we further ablate two key design choices in GeoWorld-VLM: the world-model feature extraction layer and the denoising step used for teacher feature extraction. Both factors directly affect the spatial signal transferred from the world-model teacher to the VLM visual interface. Unless otherwise specified, all ablation studies are conducted using Gemma4 as the VLM backbone and are evaluated on the What'sUp+VSR suite under the same experimental setting. More detailed analysis is provided in Appendix C.

World-model layer selection. Table 4 compares layers 23, 24, and 25 of the world-model teacher. Layer 24 achieves the best overall performance, improving the base Gemma4 model from 61.70 to 65.45 and outperforming layers 23 and 25 by 2.38 and 2.35 points, respectively. This suggests that the most transferable spatial supervision lies in an intermediate-to-late region of the world-model hierarchy, where the representation has incorporated camera- and prompt-conditioned spatial structure while remaining general enough to supervise the VLM visual pathway.

Denoising step selection. Table 5 evaluates the number of denoising steps used when extracting teacher features. Using zero or one denoising step improves over the base model, but both variants plateau at an overall score of 63.00. Using two denoising steps achieves the best overall performance of 65.45, while using three steps reduces the overall score to 60.23. This indicates that the teacher signal is most useful after a small amount of spatial prediction, but excessive denoising may make the representation too generation-specific and less suitable for aligning the frozen VLM visual interface.

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4	94.17	79.90	**40.57**	64.55	69.14	68.49	74.36	61.70
Gemma4 + Ours, layer 23	97.57	94.12	34.25	63.64	77.07	65.75	**80.68**	63.07
Gemma4 + Ours, layer 24	**99.51**	93.14	37.46	**70.45**	**78.28**	**76.71**	80.17	**65.45**
Gemma4 + Ours, layer 25	99.03	**95.10**	34.43	66.82	76.03	65.07	79.66	63.10

Table 4. Ablation on world-model feature extraction layers. Layer 24 achieves the best overall performance, suggesting that intermediate-to-late world-model representations provide the most transferable geometry-aware supervision.

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4	94.17	79.90	**40.57**	64.55	69.14	68.49	74.36	61.70
Gemma4 + Ours, 0 step	97.57	93.14	34.88	66.36	75.34	71.23	78.80	63.00
Gemma4 + Ours, 1 step	98.06	**95.10**	34.43	66.82	76.90	65.07	78.63	63.00
Gemma4 + Ours, 2 steps	**99.51**	93.14	37.46	**70.45**	**78.28**	**76.71**	**80.17**	**65.45**
Gemma4 + Ours, 3 steps	95.63	88.73	34.25	68.18	70.00	71.23	72.31	60.23

Table 5. Ablation on denoising step selection. Using two denoising steps gives the best overall performance, while using three steps leads to clear degradation.

Additional hyperparameter ablation. We additionally ablate the alignment loss coefficient λ_{align} in Appendix C.3. The best performance is obtained with $\lambda_{\text{align}} = 0.10$, suggesting that world-model alignment should be strong enough to inject useful spatial structure into the visual pathway, but not so strong that it disrupts the original VLM visual interface and its pretrained multimodal representation.

5. Conclusion

We presented GeoWorld-VLM, a vision-side distillation framework that transfers geometric structure from frozen camera-conditioned world models into VLMs. By fine-tuning only the vision encoder and multimodal projector, GeoWorld-VLM aligns visual tokens with world-model representations while keeping the language model frozen and requiring no world-model inference at test time. Across spatial reasoning benchmarks and two VLM backbones, GeoWorld-VLM consistently improves the performance, especially on geometry-sensitive relations such as *above*, *under*, *close*, and *far*. Ablations show that prompt- and camera-aware intermediate world-model features are most effective. Overall, our results suggest that improving the geometry encoded in visual tokens can mitigate spatial failures in VLMs without retraining the language model, highlighting world models as

reusable teachers for spatially grounded multimodal intelligence.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 3

[2] Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic, 2024. 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024. 2, 3

[6] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024. 3

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[8] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas, 2025. 5

[9] Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[10] Jang Hyun Cho, Boris Ivanovic, Yulong Cao, Edward Schmerling, Yue Wang, Xinshuo Weng, Boyi Li, Yurong You, Philipp Krähenbühl, Yan Wang, et al. Language-image models with 3d understanding. *arXiv preprint arXiv:2405.03685*, 2024. 2

[11] Wei Chow, Jiageng Mao, Boyi Li, Daniel Seita, Vitor Guizilini, and Yue Wang. PhysBench: Benchmarking and enhancing vision-language models for physical world understanding. *arXiv preprint arXiv:2501.16411*, 2025. 2

[12] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[13] Mengfei Du, Binhao Wu, Zejun Li, Xuanjing Huang, and Zhongyu Wei. EmbSpatial-Bench: Benchmarking spatial understanding for embodied tasks with large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers*, 2024. 2, 5

[14] Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024. 2

[15] Gemini Team, Google. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. 2

[16] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 3

[17] Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models. *Advances in Neural Information Processing Systems*, 36:20482–20494, 2023. 2

[18] 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. 2

[19] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, et al. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[20] Amita Kamath, Jack Hessel, and Kai-Wei Chang. What's "up" with vision-language models? Investigating their struggle with spatial reasoning. *arXiv preprint arXiv:2310.19785*, 2023. 2, 5

[21] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 3

[22] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 3

[23] Phillip Y Lee, Jihyeon Je, Chanho Park, Mikaela Angelina Uy, Leonidas Guibas, and Minhyuk Sung. Perspective-aware reasoning in vision-language models via mental imagery simulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9241–9251, 2025. 3

[24] 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 (ICML)*, 2023. 2

[25] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024. 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 6

[27] Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023. 5

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[29] Yifan Liu, Fangneng Zhan, Kaichen Zhou, Yilun Du, Paul Pu Liang, and Hanspeter Pfister. Abstract 3d perception for spatial intelligence in vision-language models. *arXiv preprint arXiv:2511.10946*, 2025. 1, 3

[30] Gen Luo, Ganlin Yang, Ziyang Gong, Guanzhou Chen, Haonan Duan, Erfei Cui, Ronglei Tong, Zhi Hou, Tianyi Zhang, Zhe Chen, et al. Visual embodied brain: Let multimodal large language models see, think, and control in spaces. *arXiv preprint arXiv:2506.00123*, 2025. 2

[31] OpenAI. Introducing GPT-5. https://openai.com/index/introducing-gpt-5/, 2025. Accessed: 2026-05-04. 2

[32] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3

[33] Qwen Team. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[35] Arijit Ray, Jiafei Duan, Ellis Brown, Reuben Tan, Dina Bashkirova, Rose Hendrix, Kiana Ehsani, Aniruddha Kembhavi, Bryan A. Plummer, Ranjay Krishna, Kuo-Hao Zeng, and Kate Saenko. SAT: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024. 2

[36] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 3, 5

[37] Yuan Tang, Xu Han, Xianzhi Li, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Minigpt-3d: Efficiently aligning 3d point clouds with large language models using 2d priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6617–6626, 2024. 2

[38] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 2

[39] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026. 3

[40] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, Yihang Chen, Jie Liu, Yansong Cheng, Yao Yao, Jiayi Zhu, Yihao Meng, Kecheng Zheng, Qingyan Bai, Jingye Chen, Zehong Shen, Yue Yu, Xing Zhu, Yujun Shen, and Hao Ouyang. Advancing open-source world models, 2026. 6

[41] Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. 2, 3, 12

[42] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 3

[43] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 2, 3

[44] Kangrui Wang, Pingyue Zhang, Zihan Wang, Yaning Gao, Linjie Li, Qineng Wang, Hanyang Chen, Chi Wan, Yiping Lu, Zhengyuan Yang, et al. Vagen: Reinforcing world model reasoning for multi-turn vlm agents. *arXiv preprint arXiv:2510.16907*, 2025. 3

[45] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. 5

[46] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 2, 3

[47] Xiaofeng Wang, Zheng Zhu, Guan Huang, Boyuan Wang, Xinze Chen, and Jiwen Lu. Worlddreamer: Towards general world models for video generation via predicting masked tokens. *arXiv preprint arXiv:2401.09985*, 2024. 3

[48] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 1

[49] Haomiao Xiong, Yunzhi Zhuge, Jiawen Zhu, Lu Zhang, and Huchuan Lu. 3ur-llm: An end-to-end multimodal large language model for 3d scene understanding. *IEEE Transactions on Multimedia*, 2025. 2

[50] Jianwei Yang, Reuben Tan, Qianhui Wu, Ruijie Zheng, Baolin Peng, Yongyuan Liang, Yu Gu, Mu Cai, Seonghyeon Ye, Joel Jang, et al. Magma: A foundation model for multimodal ai agents. In *Proceedings of the computer vision and pattern recognition conference*, pages 14203–14214, 2025. 2

[51] Yuncong Yang, Jiageng Liu, Zheyuan Zhang, Siyuan Zhou, Reuben Tan, Jianwei Yang, Yilun Du, and Chuang Gan. Mindjourney: Test-time scaling with world models for spatial reasoning. *arXiv preprint arXiv:2507.12508*, 2025. 3

[52] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 3

[53] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 3

[54] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10412–10420, 2025. 3

[55] Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12405–12414, 2025. 3

[56] Kaichen Zhou, Yuhan Wang, Grace Chen, Xinhai Chang, Gaspard Beaudouin, Fangneng Zhan, Paul Pu Liang, and Mengyu Wang. Page-4d: Disentangled pose and geometry estimation for 4d perception. *arXiv preprint arXiv:2510.17568*, 2025. 3

[57] Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. Vgbench: Evaluating large language models on vector graphics understanding and generation, 2024. 6

A. More Details about LingBot-World-Fast

GeoWorld-VLM uses LingBot-World-Fast as the default world-model teacher. LingBot-World-Fast is the efficient variant of LingBot-World, an open-source video-based world simulator designed to move beyond passive text-to-video generation toward interactive world modeling [41]. Unlike conventional video generators that mainly synthesize short visual clips, LingBot-World is trained to model temporally extended visual dynamics, action-conditioned scene evolution, and long-range spatial consistency. These properties make it particularly suitable for our setting, where the goal is not to use the generated video as an additional test-time input, but to extract intermediate representations that encode how a static scene may evolve under controlled camera motion.

LingBot-World is developed through a multi-stage training pipeline. The first stage builds on a strong image-to-video diffusion prior, which provides high-fidelity visual generation and general spatiotemporal coherence. The middle-training stage further injects world knowledge and action controllability, enabling the model to maintain scene-level consistency over extended horizons and to respond to user-specified controls. The final post-training stage adapts the model for causal and efficient interactive generation through architectural adaptation and few-step distillation. LingBot-World-Fast corresponds to this efficient post-trained variant, which is optimized for low-latency rollout while preserving the structural and physical regularities learned by the larger world model.

A key feature of LingBot-World-Fast is its action-conditioned architecture. The model takes an initial image or video, noisy video latents, textual conditions, and user-defined action signals as inputs. Its action interface supports both continuous camera motion and discrete control signals. In the original LingBot-World formulation, camera motion is represented with geometric embeddings, while discrete actions such as keyboard-style controls are encoded as action tokens and injected into the diffusion transformer through adaptive normalization. As a result, the hidden states of the model are shaped not only by image appearance and text semantics, but also by the implied camera trajectory and the corresponding counterfactual scene transformation.

This design is aligned with the motivation of GeoWorld-VLM. Spatial relations such as *behind*, *under*, *between*, *close*, and *far* are often difficult to infer from object identity alone. They depend on viewpoint, occlusion, support, relative depth, and how object projections change under camera motion. Because LingBot-World-Fast is trained to predict visually coherent future states under controllable motion, its intermediate representations are expected to contain motion-aware geometric cues that are complementary to static visual encoders such as DINO or CLIP. This is why we use LingBot-World-Fast as a feature-level teacher rather than as a video generator during inference.

In our implementation, LingBot-World-Fast is kept frozen throughout training. For each input image, we construct an image-to-video teacher input using a lightweight text condition and a sampled egocentric camera trajectory. We then run the world model for a small number of denoising steps and extract an intermediate hidden state from the diffusion transformer as the teacher representation. The extracted representation is projected into a shared alignment space and used to supervise the VLM post-projector visual tokens. Importantly, the generated frames themselves are not used as additional inputs to the student VLM at test time. The world model is only used during training to provide geometry-aware supervision, so GeoWorld-VLM retains the same inference interface and computational cost as the adapted VLM backbone.

We select LingBot-World-Fast rather than the full LingBot-World-Base model for practical reasons. The fast variant offers a better trade-off between teacher quality and

feature extraction cost, which is important because teacher representations must be extracted for thousands of training examples. At the same time, the model remains grounded in the same world-modeling pipeline as the larger base model, including long-horizon consistency, action conditioning, and causal interactive generation. Therefore, LingBot-World-Fast provides a computationally feasible source of camera-conditioned spatial supervision while preserving the core geometric properties needed by GeoWorld-VLM.

B. Training Details

This appendix provides additional implementation and training details for GeoWorld-VLM and the compared distillation baselines.

Training objective. For all GeoWorld-VLM experiments, the language model backbone is frozen, and we update only the vision encoder, the multimodal projector, and the lightweight alignment heads. The training objective consists of three terms:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{preserve}}\mathcal{L}_{\text{preserve}}. \tag{8}$$

Here, $\mathcal{L}_{\text{task}}$ is the cross-entropy loss over multiple-choice option-letter logits. $\mathcal{L}_{\text{align}}$ is a cosine alignment loss between the projected student visual representation and the projected teacher representation. $\mathcal{L}_{\text{preserve}}$ is a normalized mean-squared error loss between the current student visual representation and that of the frozen original VLM, which is used to reduce visual-interface drift. Unless otherwise specified, we set $\lambda_{\text{align}} = 0.10$ and $\lambda_{\text{preserve}} = 0.05$.

Optimization hyperparameters. All methods are trained for three epochs with AdamW. We use a learning rate of 2×10^{-5}, weight decay of 0.01, gradient accumulation of 1, maximum gradient norm of 1.0, and random seed 42. For Gemma4-based experiments, we use a batch size of 4 for both training and evaluation. For InternVL3.5-2B-based experiments, we generally use a batch size of 2 for the main-suite experiments due to the larger visual-token cost introduced by dynamic image tiling.

Alignment heads. For teacher-student feature alignment, both the student visual features and the teacher features are passed through separate two-layer MLP projection heads. The hidden dimension of the projection head is set to 1024, and the final alignment dimension is set to 512. The alignment loss is computed after both student and teacher features are projected into this shared feature space. This design avoids requiring the raw teacher and student hidden states to have the same dimensionality.

World-model teacher. For GeoWorld-VLM, we use a frozen camera-conditioned video world model as the teacher. The main experiments use the LingBot-Fast image-to-video teacher with 9 frames, 2 teacher denoising steps, and camera perturbation enabled. We extract the teacher representation from transformer block 24. The teacher prompt is:

> *A slight camera motion with stable object layout and unchanged spatial relations.*

This prompt is designed to induce mild viewpoint changes while preserving the object layout and spatial relations in the original image.

World-model visual configuration. For the LingBot/Wan-based teacher, images are processed at resolution 480×832 with 9 video frames. The target timestep is set to 300, and the shift parameter is set to 5.0. The VAE stride is $(4, 8, 8)$ and the patch size is $(1, 2, 2)$. The teacher transformer has hidden dimension 5120 and 40 transformer layers. For the LingBot-Fast denoising schedule used in our main experiments, the selected timesteps are

$$[999,\ 957,\ 899,\ 702],$$

with corresponding sigmas

$$[0.9998,\ 0.9580,\ 0.8994,\ 0.7024].$$

Unless otherwise specified, we use the two-step teacher setting in the main reported results.

Gemma4 visual processing. For Gemma4, the vision patch size is 16. In our implementation, the observed visual input representation can be organized as patch/token features, and feature alignment is performed over visual token features after projecting both the student and teacher features to the shared 512-dimensional alignment space. This matches the design choice in GeoWorld-VLM: the supervision is applied at the VLM visual interface before the frozen language model consumes the visual tokens.

InternVL3.5 visual processing. For InternVL3.5-2B, we use its dynamic image tiling strategy. Each tile has resolution 448×448. In the main reported setting, the maximum number of visual tiles is set to 4. The preprocessing procedure selects the closest aspect-ratio tiling layout under the tile budget, resizes and crops the image into 448-sized tiles, and adds a thumbnail when multiple tiles are used. The prompt is constructed by inserting repeated `<IMG_CONTEXT>` tokens according to the number of image tokens and the number of visual tiles. The student visual features are then projected to the shared 512-dimensional alignment space before computing the alignment loss.

DINOv3 distillation baseline. For the DINO baseline, we replace the camera-conditioned world-model teacher with a frozen DINOv3 ViT-S/16 encoder. Images are resized to 224×224, yielding an approximate 14×14 spatial grid with patch size 16. When the Hugging Face processor is unavailable, we use a fallback preprocessing pipeline consisting of resizing to 224×224 followed by ImageNet normalization. The DINO visual features are treated as static teacher features and are projected into the same 512-dimensional alignment space as the student features. The alignment loss is again computed using cosine distance. For both Gemma4 and InternVL3.5-2B, the DINO baseline uses the same default loss weights, namely $\lambda_{\text{align}} = 0.10$ and $\lambda_{\text{preserve}} = 0.05$, and is trained for three epochs under the same downstream supervision.

Model-specific settings. For Gemma4 with the world-model teacher, the main experiments use the LingBot-Fast image-to-video teacher with 9 frames, 2 denoising steps, transformer block 24, camera perturbation, and the layout-preserving prompt described above. For InternVL3.5-2B with the world-model teacher, we use the same teacher configuration, while adapting the student-side preprocessing to InternVL's dynamic tiling mechanism. For Gemma4 with DINOv3 and InternVL3.5-2B with DINOv3, the only change is replacing the world-model representation with DINOv3 visual features; the optimization hyperparameters and loss weights are kept the same for a controlled comparison. The time consumed to use our method on Gemma4 is about 5.5 hours if the training set is What'sup and VSR and the compute source is 2 Nvidia H200.

C. Additional Ablation Analysis

The main paper reports the ablation results on world-model layer selection and denoising step selection in Section 4.5. Here, we provide additional analysis of these two design choices and further report the ablation on the alignment loss coefficient.

C.1. Analysis of World-Model Layer Selection

The layer selection results in Table 4 show that layer 24 provides the strongest overall supervision among the tested layers. This pattern suggests that the usefulness of world-model features is not monotonic with depth. A slightly shallower layer may still retain more local appearance information and may not fully encode the camera- and prompt-conditioned spatial structure needed for downstream spatial reasoning. A later layer, on the other hand, may become more specialized toward the teacher model's generative objective and less compatible with the frozen VLM visual interface.

Layer 24 appears to offer a better balance. It has already integrated viewpoint-conditioned spatial cues, but its representation remains sufficiently general to serve as a transferable supervision signal. This is consistent with the design goal of GeoWorld-VLM: the teacher feature should provide geometry-aware information without forcing the student visual tokens to imitate generation-specific representations too strongly.

C.2. Analysis of Denoising Step Selection

The denoising step results in Table 5 show that teacher features are most effective when extracted after two denoising steps. This suggests that a small amount of world-model prediction helps reveal the geometry induced by the sampled camera trajectory. With zero or one denoising step, the teacher representation may still be close to the initial image-conditioned latent and may not fully expose viewpoint-dependent structure. This explains why both settings improve over the base model but remain limited at an overall score of 63.00.

However, more denoising is not necessarily better. Using three denoising steps substantially reduces the overall score. One possible explanation is that later denoising states become increasingly tied to video synthesis rather than representation transfer. They may contain generation-specific artifacts or drift away from the original image evidence, making them less suitable for supervising the VLM visual pathway. Therefore, GeoWorld-VLM benefits most from an intermediate predictive representation: it should be spatially enriched by the world model, but still anchored to the input image.

C.3. Alignment Loss Coefficient

Table 6 studies the effect of the alignment loss coefficient. The best performance is obtained with $\lambda_{\text{align}} = 0.10$, which reaches an overall score of 65.45. When the coefficient is too small, the teacher signal may be insufficient to inject useful spatial structure into the student visual representation. When the coefficient is too large, the adapted visual tokens may overfit to the world-model feature space and become less compatible with the frozen language model. Therefore, the alignment loss is best understood as a controlled geometric regularizer rather than a replacement objective for VLM training.

C.4. Summary

Overall, the ablation results support three design choices used in the main experiments. First, the 24th world-model layer provides the strongest supervision among the tested layers. Second, two denoising steps offer the best trade-off between exposing spatial dynamics and preserving fidelity to the input image. Third, $\lambda_{\text{align}} = 0.10$ gives the best balance between transferring geometry-aware information and maintaining the original VLM visual interface. These observations justify the default GeoWorld-VLM configuration used in the main results.

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4	94.17	79.90	**40.57**	64.55	69.14	68.49	74.36	61.70
Gemma4 + Ours, $\lambda_{\text{align}} = 0.05$	<u>99.03</u>	<u>92.65</u>	36.21	<u>66.82</u>	76.03	69.86	77.95	63.49
Gemma4 + Ours, $\lambda_{\text{align}} = 0.10$	**99.51**	**93.14**	<u>37.46</u>	**70.45**	**78.28**	**76.71**	**80.17**	**65.45**
Gemma4 + Ours, $\lambda_{\text{align}} = 0.15$	<u>99.03</u>	92.16	35.85	66.36	<u>76.90</u>	<u>72.60</u>	<u>79.49</u>	<u>63.88</u>
Gemma4 + Ours, $\lambda_{\text{align}} = 0.20$	98.54	89.22	33.99	69.55	75.69	65.75	<u>79.49</u>	62.64

Table 6. Ablation on the alignment loss coefficient. $\lambda_{\text{align}} = 0.10$ achieves the best overall result, suggesting that world-model alignment should be strong enough to inject spatial structure but not so strong that it disrupts the original VLM visual interface. The best result is shown in bold and the second-best result is underlined.

D. Qualitative Case Studies

We provide qualitative examples from EmbSpatial-Bench to further illustrate how GeoWorld-VLM changes the spatial behavior of the base VLM. As shown in Table 7, Cases 1–5 highlight examples where only our method predicts the correct answer, while the raw model, FT-only baseline, and DINO-aligned baseline all fail. These cases suggest that task supervision or static visual feature alignment alone may still be insufficient for relations that require geometric structure, such as vertical placement, left-right ordering, and viewpoint-dependent distance judgment. In contrast, GeoWorld-VLM benefits from camera-conditioned world-model supervision, which provides a more geometry-aware training signal for adapting the visual interface of the VLM. Table 8 further shows cases where our method corrects errors made by the raw model, while FT-only or DINO alignment may also succeed. Together, these examples support our main finding that world-model alignment can improve spatial reasoning not merely by memorizing task labels, but by encouraging the visual pathway to preserve spatial cues that are useful before language reasoning begins.

D.1. Cases Where Only Our Method Is Correct

D.2. Cases Where Our Method Corrects the Raw Model

E. Limitations and Broader Impact.

GeoWorld-VLM has several limitations. First, our method depends on the quality of the frozen world-model teacher: if the teacher encodes inaccurate viewpoint dynamics or weak scene geometry, the extracted features may provide noisy supervision. Second, although GeoWorld-VLM requires no world-model inference at test time, training still introduces additional cost because teacher features must be extracted from a large video world model. Finally, the method remains sensitive to feature extraction choices such as teacher layer and denoising step, and the gains are not uniform across all spatial relations or subsets. Future work could explore stronger and more efficient teachers, broader embodied evaluations, and more principled criteria for selecting geometry-aware teacher representations. GeoWorld-VLM aims to improve the spatial reasoning ability of vision-language models by enhancing the visual representations consumed by a frozen language model. This capability could have positive impacts on embodied AI, assistive agents, robotics, navigation, and multimodal systems that need more reliable understanding of object layout and spatial relations. At the same time, stronger spatial perception may also increase the capability of systems used in surveillance, autonomous decision-making, or other safety-sensitive applications where incorrect spatial judgments could cause harm. Our work is a research study on benchmarked spatial reasoning rather than a deployed system, and we do not release new high-risk datasets, pretrained generative models, or autonomous agents. Future deployments of spatially enhanced VLMs should include careful evaluation under domain-specific safety requirements, robustness testing, and human oversight in high-stakes settings.

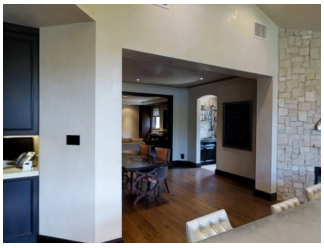
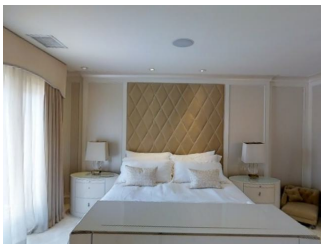
Image	Question / Ground Truth	Model Predictions
	Q: In the image, how do the positions of picture and ottoman interact with each other? **GT:** The picture is above the ottoman.	Raw: The picture is on the left side of the ottoman. ✗ FT-only: The picture is on the left side of the ottoman. ✗ DINO: The picture is on the left side of the ottoman. ✗ Ours: The picture is above the ottoman. ✓
	**Q:** In the image, how do the positions of television and bar interact with each other? **GT:** The television is left of the bar.	Raw: The television is on the right side of the bar. ✗ FT-only: The television is on the right side of the bar. ✗ DINO: The television is on the right side of the bar. ✗ Ours: The television is left of the bar. ✓
	Q: Among the listed objects, which one is farthest from your current location in the image? **GT:** plant	Raw: bathtub ✗ FT-only: bathtub ✗ DINO: bathtub ✗ Ours: plant ✓
	**Q:** From your viewpoint, which object among the options is the nearest within the image? **GT:** chest	Raw: bed ✗ FT-only: curtain ✗ DINO: bed ✗ Ours: chest ✓
	Q: What is the spatial relationship between chair and bicycle in the image? **GT:** The chair is at the right side of the bicycle.	Raw: The chair is below the bicycle. ✗ FT-only: The chair is below the bicycle. ✗ DINO: The chair is below the bicycle. ✗ Ours: The chair is at the right side of the bicycle. ✓

Table 7. Qualitative examples where only our LingBot-teacher alignment method predicts the correct spatial answer.

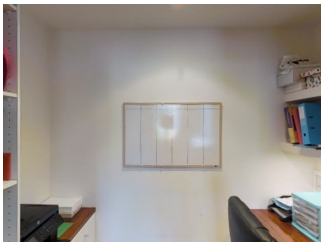
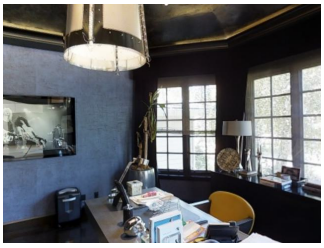
Image	Question / Ground Truth	Model Predictions
	Q: Which object, in relation to your current position, holds the nearest placement in the image? **GT:** pillow	Raw: pool table ✗ FT-only: pillow ✓ DINO: pillow ✓ Ours: pillow ✓
	Q: What is the spatial configuration between monitor and lamp in relation to each other within the image? **GT:** The monitor is beneath the lamp.	Raw: The monitor is on the left side of the lamp. ✗ FT-only: The monitor is on the left side of the lamp. ✗ DINO: The monitor is beneath the lamp. ✓ Ours: The monitor is beneath the lamp. ✓
	Q: What is the spatial arrangement of wardrobe and curtain in the image concerning each other? **GT:** The wardrobe is on the right side of the curtain.	Raw: The wardrobe is outside the curtain. ✗ FT-only: The wardrobe is on the right side of the curtain. ✓ DINO: The wardrobe is on the right side of the curtain. ✓ Ours: The wardrobe is on the right side of the curtain. ✓
	**Q:** How are printer and paper positioned in relation to each other in the image? **GT:** The printer is at the left side of the paper.	Raw: The printer is at the right side of the paper. ✗ FT-only: The printer is at the left side of the paper. ✓ DINO: The printer is at the left side of the paper. ✓ Ours: The printer is at the left side of the paper. ✓
	**Q:** What is the spatial arrangement of picture and bin in the image concerning each other? **GT:** The picture is above the bin.	Raw: The picture is left of the bin. ✗ FT-only: The picture is above the bin. ✓ DINO: The picture is above the bin. ✓ Ours: The picture is above the bin. ✓

Table 8. Qualitative examples where our method corrects errors made by the raw model. FT-only or DINO alignment may also succeed on some of these examples, but the raw VLM fails.