GEOWORLD-VLM: GEOMETRY FROM WORLD MODELS FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

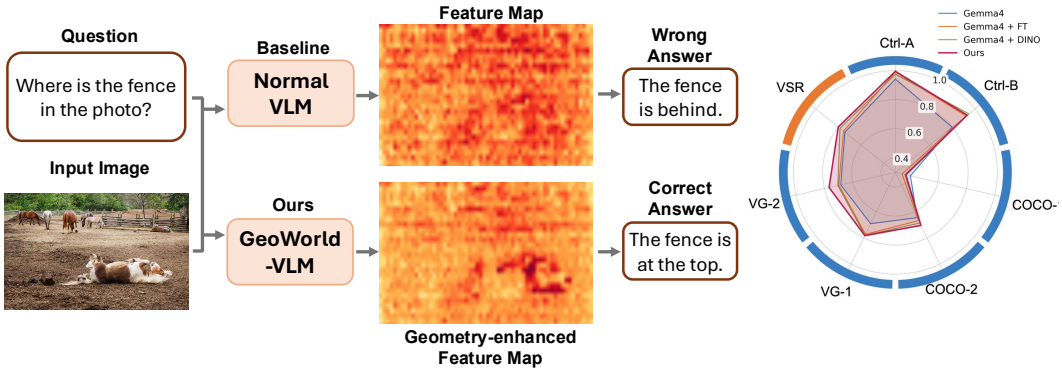


Figure 1: **Overview.** GeoWorld-VLM distills camera-conditioned world-model features into a VLM's visual pathway during training and retains standard single-image inference. The left illustrates changes in visual feature responses and the predicted spatial answer. The right compares Gemma4, task-only fine-tuning, DINO-based alignment, and GeoWorld-VLM across the What'sUp+VSR subsets; Table 1 reports the corresponding accuracies.

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. We introduce GeoWorld-VLM, which transfers predictive representations from a frozen camera-conditioned video world model into a VLM's existing visual pathway to improve single-image spatial reasoning. From each static training image, a sampled camera trajectory and fixed scene-stability prompt produce viewpoint-conditioned teacher features without real multi-view observations. Post-projector feature alignment transfers this supervision, while a preservation anchor constrains changes to the visual interface consumed by the frozen language model. Only the image encoder, multimodal projector, and student alignment head are updated; the teacher and alignment heads are removed after training. Across two VLM backbones on What'sUp+VSR and EmbSpatial, GeoWorld-VLM improves over matched task-only fine-tuning by 1.53–3.52 percentage points and achieves the highest mean accuracy among the evaluated feature teachers. These results demonstrate that predictive world-model representations can improve single-image spatial reasoning at no extra inference cost.

1 INTRODUCTION

Vision-Language Models (VLMs) have achieved strong performance in semantic recognition, but they remain unreliable spatial reasoners Liu et al. (2025); Wu et al. (2025). Even strong contemporary VLMs struggle with elementary relations such as *left of*, *on top of*, *behind*, and *between*. Recognizing the objects and parsing the question do not by themselves ensure an accurate spatial judgment. These failures motivate strengthening the spatial evidence delivered to the language model.

The image encoder and multimodal projector determine the visual evidence available to a VLM's language model Hurst et al. (2024); Luo et al. (2025); Yang et al. (2025a). We target this visual interface, improving spatial reasoning while keeping the language model fixed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Direct visual supervision offers a way to strengthen these representations beyond answer prediction. Ross (Wang et al., 2025a) introduces latent reconstruction, VIRAL (Yoon et al., 2025) aligns intermediate LLM visual tokens with vision foundation models, and 3DRS (Huang et al., 2025) transfers features from 3D foundation models. Video generative models offer another source of spatial information: VEGA-3D (Wu et al., 2026) fuses intermediate generative features into the VLM, while MindJourney (Yang et al., 2025b) uses imagined trajectories during inference. These advances motivate a complementary question: can a VLM absorb predictive world-model features during training to improve single-image spatial reasoning with a frozen LLM and its original inference architecture? As shown in Figure 1, **GeoWorld-VLM** transfers predictive world-model representations into the VLM's visual pathway for single-image spatial reasoning. Given an image, a fixed scene-stability prompt, and a sampled camera trajectory, the teacher provides intermediate features conditioned on viewpoint changes. This converts camera-conditioned prediction into visual supervision without requiring real multi-view observations. We align the VLM's post-projector image tokens with these features, directly supervising the interface consumed by the frozen LLM.

To transfer these features while maintaining the frozen LLM's visual interface, GeoWorld-VLM combines spatial answer supervision and teacher alignment with a preservation anchor to the original post-projector tokens. Adaptation updates only the image encoder, multimodal projector, and student alignment head, preserving text-only behavior by construction. After training, the world model and alignment heads are removed, leaving the original single-image inference architecture.

We evaluate this transfer on Gemma4 and InternVL3.5-2B (Wang et al., 2025b) using What'sUp, VSR, and EmbSpatial-Bench. The complete framework improves over matched task-only fine-tuning by 1.53–3.52 points and achieves the highest mean across the four backbone–suite settings against DINOv3, SigLIP2, RADIO, and VGGT teachers. Teacher comparisons and ablations examine how supervision and preservation contribute to these gains; comparison with FT+MindJourney measures both accuracy and latency under matched downstream task supervision.

- **Predictive world-model supervision from static images.** We introduce a teacher–student formulation that uses intermediate features from camera-conditioned prediction to supervise single-image VLM spatial reasoning without real multi-view training observations.
- **Transfer into the existing visual interface.** GeoWorld-VLM combines post-projector alignment and interface preservation while keeping the LLM frozen. The teacher is used only during training, leaving no auxiliary feature branch or generation step at inference.
- **Improved spatial reasoning without inference overhead.** Across two backbones and three spatial benchmarks, GeoWorld-VLM improves over matched task-only fine-tuning and achieves the highest mean accuracy among the evaluated feature teachers, while retaining the original VLM architecture and standard single-image inference procedure.

2 Related Work

2.1 Spatial and 3D Reasoning in Vision-Language Models

Spatial benchmarks expose limitations that are not captured by semantic recognition alone. What'sUp (Kamath et al., 2023), VSR (Liu et al., 2023), SAT (Ray et al., 2024), and EmbSpatial-Bench (Du et al., 2024) test relations, reference frames, and distance judgments that remain challenging for VLMs. One approach supplies explicit 3D context: 3D-LLM (Hong et al., 2023) introduces 3D scene features, while MiniGPT-3D (Tang et al., 2024) aligns point-cloud representations with language models. Another approach constructs additional spatial evidence during inference, as in the abstract perspective changes of APC (Lee et al., 2025) and the proxy geometry of SandboxVLM (Liu et al., 2025). GeoWorld-VLM addresses single-image spatial QA by adapting the VLM's existing visual pathway, retaining its original inference interface.

2.2 Visual Representation Supervision

Direct visual supervision complements the indirect signal provided by answer prediction. Ross (Wang et al., 2025a) uses latent reconstruction to encourage visual outputs to retain image detail. VIRAL (Yoon et al., 2025) aligns visual-token representations at intermediate LLM layers with frozen vision foundation models, including DINOv2, SigLIP2, and RADIO, during instruction tuning. For 3D scene understanding, 3DRS (Huang et al., 2025) distills representations from pre-trained 3D foundation models into MLLM visual features. These studies establish representation supervision as an effective tool for improving visual understanding.

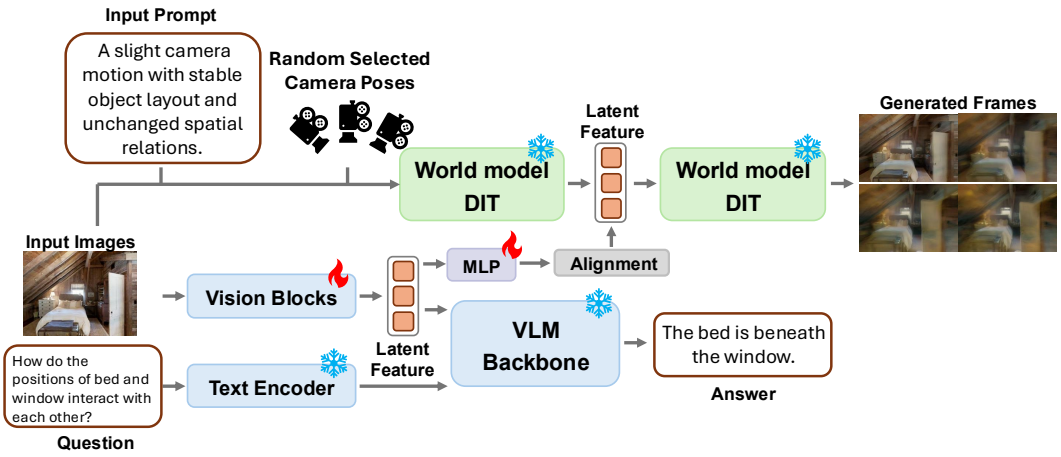


Figure 2: **GeoWorld-VLM.** A frozen world model takes an image, a scene-stability prompt, and sampled camera poses to provide predictive supervision. Post-projector alignment and interface preservation train the image encoder, multimodal projector, and student alignment head. The teacher and alignment heads are removed, retaining the original single-image VLM inference architecture.

GeoWorld-VLM builds on this principle to study the transfer of camera-conditioned predictive features into the visual interface of a frozen LLM. The teacher processes a static image together with a sampled camera trajectory and a fixed scene-stability prompt. Alignment acts on post-projector image tokens, with updates confined to the image encoder, multimodal projector, and student alignment head. A preservation objective anchors these tokens to the original interface. Our matched DINOv3, SigLIP2, RADIO, and VGGT baselines compare supervision sources within this common training recipe; they are not reproductions of the full VIRAL or 3DRS training pipelines.

2.3 Video Generative Priors for Spatial Understanding

Video generative models provide spatial priors through their learned representations of scene structure and temporal evolution (Wan et al., 2025; Yang et al., 2024; Team et al., 2026b). Two closely related methods exploit these priors at inference. MindJourney (Yang et al., 2025b) generates imagined trajectories to provide additional visual evidence for spatial reasoning. VEGA-3D (Wu et al., 2026) extracts intermediate video-diffusion features and combines them with semantic tokens through adaptive gated fusion. Its generative features remain part of the inference pathway, although caching can amortize the cost of extracting them for repeated queries on the same visual input.
GeoWorld-VLM transfers predictive features into the existing visual pathway during training, so neither a generative feature branch nor generated views are needed at inference. Explicit camera conditioning determines the predictive supervision, while alignment before a frozen LLM transfers it without retaining a separate generative feature stream at inference.

3 Method

We introduce **GeoWorld-VLM**, a vision-side distillation method that transfers predictive representations from a frozen camera-conditioned video world model into a VLM Beyer et al. (2024); Wang et al. (2025c); Team (2026). Figure 2 summarizes the teacher supervision, visual alignment, and interface preservation used to adapt the visual pathway before a frozen LLM.

3.1 Vision-side formulation

We write the VLM computation as

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

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. Within the VLM, GeoWorld-VLM updates only (ϕ, θ). This confines adaptation to the representation consumed by the LLM and leaves the text-only computation unchanged.

3.2 CAMERA-CONDITIONED WORLD-MODEL TEACHER

Static image teachers such as DINO Caron et al. (2021); Oquab et al. (2023); Siméoni et al. (2025) provide informative visual features without taking a camera trajectory as input. A camera-conditioned predictive teacher can additionally model how object projections change under specified viewpoint motion, motivating its use for spatial supervision. For each training image x_v, we sample an egocentric camera trajectory:

$$\pi = (\pi^{(1)}, \ldots, \pi^{(F)}), \tag{2}$$

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 Team et al. (2026b). The frozen world-model teacher receives the initial images, the camera trajectory, and a text condition: (x_v, π, c). Here, c is the fixed prompt "A slight camera motion with stable object layout and unchanged spatial relations." 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 trajectory specifies the viewpoint transformation, while the prompt encourages scene stability. This supplies a prediction-conditioned feature target from a single observed image. Section 4.4 compares joint, image-only, and single-condition inputs.

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:

$$u_s = f^s_{\psi_s}(h_{\phi,\theta}(x_v)), \qquad u_t = f^t_{\psi_t}(g(x_v)). \tag{3}$$

The student head $f^s_{\psi_s}$ is trainable, while the teacher head $f^t_{\psi_t}$ is randomly initialized and kept fixed, providing a fixed mapping into the alignment space. The alignment loss is

$$\mathcal{L}_{\text{align}} = 1 - \left\langle \frac{u_s}{\|u_s\|_2}, \frac{u_t}{\|u_t\|_2} \right\rangle. \tag{4}$$

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 without ensuring that the projector retains those spatial cues in the tokens delivered to the language model.

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:

$$\mathcal{L}_{\text{preserve}} = \left\| \frac{h_{\phi,\theta}}{\|h_{\phi,\theta}\|_2} - \frac{h_{\phi',\theta'}}{\|h_{\phi',\theta'}\|_2} \right\|_2^2. \tag{5}$$

The two objectives impose complementary constraints: alignment draws the student toward the predictive teacher representation, while preservation penalizes changes in the normalized tokens supplied to the frozen language model. Section 4.5 evaluates preservation with world-model alignment.

3.5 TRAINING OBJECTIVE

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

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{preserve}}\mathcal{L}_{\text{preserve}}. \tag{6}$$

Here, $\mathcal{L}_{\text{task}}$ is the option-restricted cross-entropy loss for spatial multiple-choice supervision. Gradients update only (ϕ, θ, ψ_s); the LLM, the world-model teacher, the teacher-side projection head $f^t_{\psi_t}$, and the frozen reference VLM remain fixed. At inference, the teacher, reference VLM, and alignment heads are discarded, leaving the adapted VLM's original single-image computation.

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

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 + SigLIP2	96.60	91.18	34.16	66.82	75.69	63.01	77.61	62.02
Gemma4 + RADIO	99.03	93.63	36.21	67.27	76.21	70.55	79.15	63.88
Gemma4 + VGGT	98.06	93.14	34.61	70.00	77.07	69.86	79.83	63.65
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 + SigLIP2	98.86	99.35	33.94	**62.32**	66.12	**53.07**	81.85	60.84
InternVL3.5-2B + RADIO	**100.00**	**100.00**	35.23	56.36	62.59	36.99	**84.10**	60.00
InternVL3.5-2B + VGGT	98.54	**100.00**	36.30	59.09	59.14	36.99	82.91	59.61
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 achieves the highest overall accuracy on both backbones against matched FT and feature teachers. Bold and underlining mark the best and second-best results for each backbone.

`OpenGVLab/InternVL3_5-2B-Instruct` Wang et al. (2025b). *Base* denotes the original VLM, while *FT* uses the same downstream supervision without teacher-student feature alignment or the preservation anchor. We further compare against frozen DINOv3 (Siméoni et al., 2025), SigLIP2, RADIO, and VGGT teachers under the same student architecture, downstream data, alignment location, and optimization schedule. All feature-teacher baselines retain the preservation anchor with the same default loss weights. *Ours* denotes GeoWorld-VLM: comparison with FT measures the complete adaptation objective, while teacher replacement assesses the supervision source with the alignment location and preservation objective held fixed.

Datasets. We evaluate GeoWorld-VLM on three spatial reasoning benchmarks: What'sUp (Kamath et al., 2023), Visual Spatial Reasoning (VSR) (Liu et al., 2023), and EmbSpatial-Bench (Du et al., 2024). Following the evaluation protocol of prior spatial reasoning work (Chen et al., 2025), we combine What'sUp and VSR into a unified What'sUp+VSR suite. This suite covers controlled tabletop relations, spatial QA in COCO and Visual Genome images, and natural-image spatial verification. 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 (Lin et al., 2015; Zou et al., 2024). For EmbSpatial, we report overall accuracy and results for *left*, *right*, *above*, *under*, *close*, and *far*.

Implementation Details. GeoWorld-VLM uses LingBot-World-Fast (Team et al., 2026a) as a frozen camera-conditioned image-to-video teacher, extracting intermediate DiT features as geometry-aware supervision. We use the prompt, camera sampling, and parameter-freezing scheme in Section 3. 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. This configuration was fixed before inspecting the benchmark evaluation results and reused across both backbones and benchmark suites; subsequent layer and step ablations analyze its sensitivity. All methods use three training epochs and the same spatial supervision; Appendix B lists hyperparameters.

4.2 Main Results on What'sUp and VSR

Table 1 reports the main results on the What'sUp+VSR suite. Relative to matched task-only fine-tuning, GeoWorld-VLM improves Gemma4 from 62.71 to 65.45 (+2.74) and InternVL3.5-2B from 58.14 to 61.66 (+3.52). It also exceeds the strongest teacher by 1.05 points on Gemma4 (DINOv3, 64.40) and 0.82 points on InternVL3.5-2B (SigLIP2, 60.84). These gains show that predictive supervision improves spatial QA beyond task-only adaptation and alternative teachers under the shared recipe. Improvements vary across subsets; Figure 3 illustrates representative predictions.

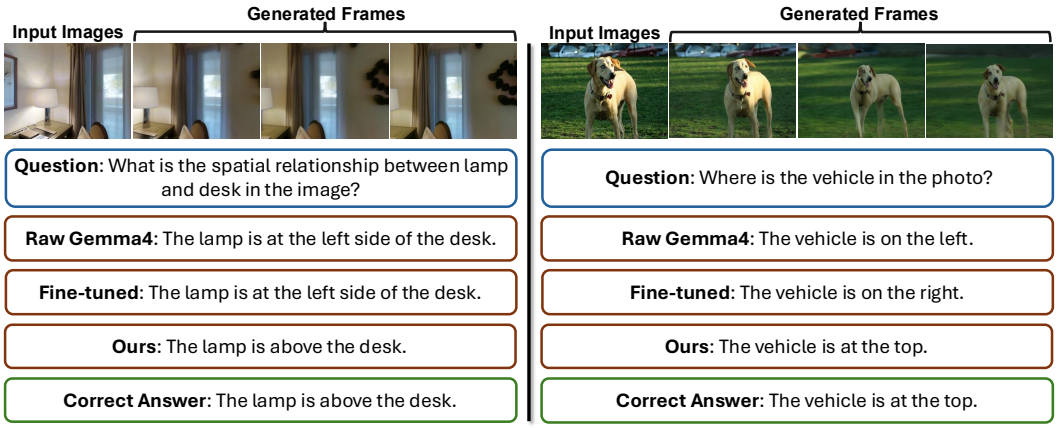


Figure 3: **Qualitative comparison.** We compare the predictions of the original Gemma4, task-only fine-tuned Gemma4, and GeoWorld-VLM on representative spatial reasoning examples. GeoWorld-VLM produces more spatially consistent answers than the baselines, illustrating the benefit of injecting camera-conditioned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

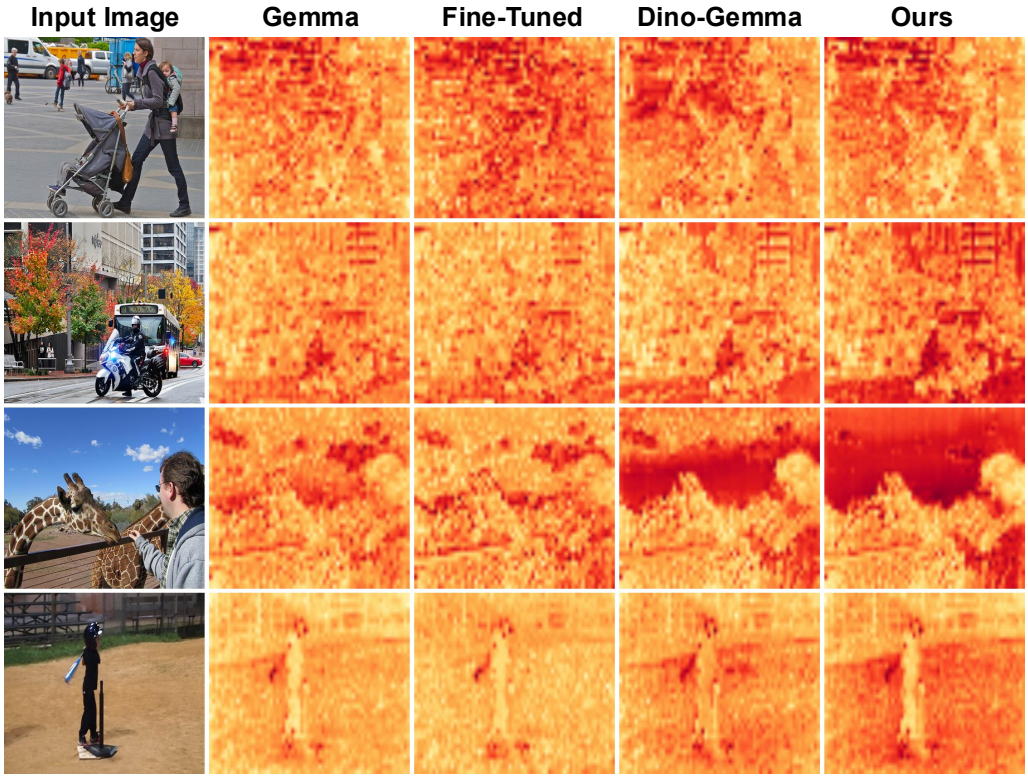


Figure 4: **Visualization of visual feature responses on the What'sUp+VSR suite.** Examples compare GeoWorld-VLM with the original VLM, task-only fine-tuning, and DINOv3-based feature distillation, showing how supervision changes visual responses in spatial QA.

4.3 Relation-wise Results on EmbSpatial

Table 2 further evaluates GeoWorld-VLM on EmbSpatial. Relative to matched FT, GeoWorld-VLM improves Gemma4 from 77.01 to 78.78 (+1.77), exceeding the strongest alternative teacher, SigLIP2, by 1.40 points. On InternVL3.5-2B, it improves FT from 78.17 to 79.70 (+1.53); the

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 + SigLIP2	**91.01**	77.06	88.06	82.29	64.13	61.57	77.38
Gemma4 + RADIO	86.69	67.03	85.07	84.87	57.61	57.84	73.17
Gemma4 + VGGT	89.57	71.33	82.46	81.92	62.68	56.34	74.09
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 + SigLIP2	88.05	79.53	90.21	83.34	68.14	65.71	79.17
InternVL3.5-2B + RADIO	86.83	78.29	89.70	86.13	70.83	64.02	79.31
InternVL3.5-2B + VGGT	**88.63**	79.73	89.70	82.45	**71.56**	65.14	79.56
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 achieves the highest overall relation accuracy across both VLM backbones, while alternative teachers remain competitive on individual relations. Bold and underlining mark the best and second-best results for each backbone.

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: Teacher conditioning analysis on the What'sUp+VSR suite. We compare image-to-video features alone with variants using prompt and camera conditioning; the native joint-conditioning configuration achieves the best overall performance. Bold and underlining identify the best and second-best scores within each evaluation subset and overall on What'sUp+VSR.

strongest alternative, VGGT, reaches 79.56, leaving a narrow 0.14-point margin. We therefore interpret the latter as a mean-level advantage rather than evidence of statistical significance. The relation-wise results are heterogeneous: for Gemma4, the clearest improvements over FT occur on *close* and *far*, where appearance alone may provide ambiguous distance cues.

4.4 Teacher Conditioning Analysis

Table 3 studies how teacher-side conditioning affects the transferred representation while holding the world-model architecture fixed. I2V-only features reach 64.01 overall; camera-only and prompt-only conditioning yield 63.16 and 62.61, respectively, while their joint use achieves the best result of 65.45. In particular, adding camera conditioning to the prompt-conditioned teacher increases the score from 62.61 to 65.45 ($+2.84$), while adding the prompt to the camera-conditioned teacher increases it from 63.16 to 65.45 ($+2.29$). Joint conditioning matches LingBot's pretraining interface and performs best, although neither condition improves over I2V-only in isolation. Appendix D reports sensitivity to layer, denoising step, and alignment strength; Section 4.8 examines how trajectory length and temporal pooling affect the transferred teacher representation.

4.5 Contributions of Alignment and Interface Preservation

Table 4 separates the available comparisons of the training objective. Adding world-model alignment to task-only FT without preservation increases accuracy from 62.71 to 63.20 ($+0.49$). Adding preservation under world-model alignment then increases accuracy to 65.45 ($+2.25$). The total $+2.74$ gain over FT therefore reflects the combined objective; it should not be attributed to alignment alone. The preservation ablation supports retaining the anchor when transferring world-model features, although it does not directly measure the mechanism of interface drift.

Table 4: Training-objective comparison on Gemma4 / What'sUp+VSR. All rows use the same task supervision. "Alignment" identifies the feature teacher; "Preservation" indicates whether the original visual interface is used as an anchor. Scores summarize Tables 1 and 13.

Configuration	Alignment	Preservation	Accuracy (%) ↑
Task-only FT	None	No	62.71
GeoWorld-VLM w/o preservation	LingBot	No	63.20
GeoWorld-VLM	LingBot	Yes	**65.45**
DINOv3 teacher	DINOv3	Yes	64.40
VGGT teacher	VGGT	Yes	63.65

Teacher replacement provides a complementary control: with preservation enabled, GeoWorld-VLM exceeds DINOv3 and VGGT by 1.05 and 1.80 points, respectively. These comparisons assess the teacher representation under the same preservation recipe. The current experiments do not include task supervision plus preservation without teacher alignment, so they do not separately identify the standalone preservation effect or its interaction with alignment. Appendix D.4 separately ablates the fixed teacher projection, which defines the alignment target in the shared feature space.

4.6 Matched Teacher Comparison and Seed Variation

Backbone / Benchmark	FT	DINOv3	SigLIP2	RADIO	VGGT	GeoWorld-VLM
Gemma4 / What'sUp+VSR	62.71±0.97	64.40±0.91	62.02±0.98	63.88±0.84	63.65±0.61	**65.45±0.41**
InternVL3.5-2B / What'sUp+VSR	58.14±0.44	54.81±1.62	60.84±0.65	60.00±0.74	59.61±0.77	**61.66±0.57**
Gemma4 / EmbSpatial	77.01±1.20	75.91±1.43	77.38±1.23	73.17±1.36	74.09±1.80	**78.78±0.85**
InternVL3.5-2B / EmbSpatial	78.17±0.52	79.09±0.12	79.17±0.58	79.31±0.41	79.56±0.38	**79.70±0.40**

Table 5: Mean ± standard deviation over three independent seeds for task-only fine-tuning, matched feature teachers, and GeoWorld-VLM. GeoWorld-VLM achieves the highest mean accuracy in all four settings, with gains varying by backbone, benchmark, and the choice of teacher.

Table 5 reports the seed-level variation underlying the main comparisons. Relative to task-only FT, GeoWorld-VLM improves by $+2.74/+3.52$ points on What'sUp+VSR and $+1.77/+1.53$ points on EmbSpatial for Gemma4/InternVL3.5-2B, respectively. The margin over the strongest alternative teacher is $+1.05$, $+0.82$, $+1.40$, and $+0.14$ points across the four settings. The mean improvement is consistent across settings, but the narrow InternVL3.5-2B/EmbSpatial margin does not establish statistical significance over the strongest alternative teacher on this benchmark.

4.7 Comparison with Test-Time World-Model Reasoning

Table 6 compares our training-time distillation with explicit test-time world-model exploration under matched task supervision. Across the four backbone–benchmark combinations, GeoWorld-VLM outperforms FT+MindJourney by 12.0–21.5 points on the matched subsets. MindJourney generates approximately five videos per question and requires 97–155 seconds per query in our evaluation, whereas GeoWorld-VLM removes the world model after training and retains the original single-image VLM inference pipeline (approximately 3–5 seconds). Thus, training-time transfer achieves higher accuracy at lower query latency in this matched-subset evaluation.

The corresponding training cost is higher: on EmbSpatial, GeoWorld-VLM uses 5.3 GPU-hours versus 0.6 for FT and 1.4 for DINO over three epochs (Appendix D.7). This cost is paid during adaptation; subsequent queries require no teacher computation. Only forward computation is needed for the frozen teacher; it receives no gradient updates and has no optimizer states.

4.8 Temporal Feature Design of the World-Model Teacher

Table 7 examines whether the transferred signal depends on a single teacher frame or on aggregation across the trajectory. Mean pooling over 9 frames reaches 65.45, compared with 61.92 for first-frame features and 63.30 for last-frame features. Using 5 or 13 frames yields 62.77 and 61.86, respectively, so longer prediction does not necessarily improve the teacher target. These results support aggregating features across a moderate trajectory for spatial QA. Mean pooling does not preserve temporal order, however, which may limit its usefulness for motion-direction questions; Appendix E examines this boundary on SAT using separate ego-motion and non-ego question splits.

Backbone	Benchmark	Base + MJ	FT + MJ	GeoWorld-VLM	Δ vs. FT+MJ
InternVL3.5-2B	What'sUp+VSR	61.50	61.50	**73.50**	**+12.00**
Gemma4	What'sUp+VSR	64.50	68.50	**84.50**	**+16.00**
InternVL3.5-2B	EmbSpatial	39.50	56.50	**78.00**	**+21.50**
Gemma4	EmbSpatial	52.00	64.00	**81.00**	**+17.00**

Table 6: Comparison with MindJourney (MJ) on matched 200-question subsets. FT+MJ applies test-time world-model exploration to the same task-only fine-tuned checkpoints used by our FT baseline. GeoWorld-VLM results are from seed 42. These scores describe the matched subsets and should be read separately from the full-benchmark averages in Tables 1 and 2.

Table 7: Temporal feature ablation on Gemma4 / What'sUp+VSR. The default uses 9 frames with mean pooling over teacher features. Bold and underlining mark the best and second-best results across the tested temporal configurations for each subset and the overall score.

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4 + Ours, 5 frames	98.54	**97.06**	34.25	67.27	73.97	<u>71.23</u>	78.12	62.77
Gemma4 + Ours, 13 frames	**99.51**	<u>96.57</u>	33.54	64.09	72.76	67.12	77.95	61.86
Gemma4 + Ours, first-frame feature	98.54	95.10	32.74	67.73	73.10	65.75	79.32	61.92
Gemma4 + Ours, last-frame feature	96.60	91.67	<u>34.88</u>	<u>68.18</u>	<u>76.55</u>	66.44	**80.51**	<u>63.30</u>
Gemma4 + Ours, 9 frames + mean pooling	**99.51**	93.14	**37.46**	**70.45**	**78.28**	**76.71**	<u>80.17</u>	**65.45**

The SAT results distinguish spatial-layout transfer from reasoning about the observer's motion. On ego-motion questions, GeoWorld-VLM scores 52.43, below FT (56.22) and DINO (63.78); on non-ego questions, it reaches 58.24, compared with 58.03 and 57.41. Thus, access to camera-conditioned predictive features does not by itself ensure improved motion reasoning after transfer. The contrast is consistent with the loss of temporal order under mean pooling, but does not isolate pooling as the cause. It motivates future teacher–student objectives that retain temporal direction while keeping the original single-image VLM architecture and avoiding test-time video generation.

Transfer beyond the training benchmark. Appendix D.5 evaluates Gemma4 in both transfer directions without target-benchmark training. Training on What'sUp+VSR yields 61.77 on EmbSpatial, exceeding FT by 5.73 points and VGGT by 1.47 points. In the reverse direction, GeoWorld-VLM reaches 62.42, only 0.10 points above DINOv3. These gains depend on the source–target pairing and provide preliminary evidence of transfer beyond the training distribution.

Retention of general multimodal capability. MMStar results in Appendix D.6 show that a frozen language model does not guarantee unchanged multimodal performance after visual adaptation. With EmbSpatial training, GeoWorld-VLM improves Gemma4 from 50.61 to 52.17 and nearly retains InternVL3.5-2B's base accuracy (55.37 versus 55.03). With What'sUp+VSR training, the corresponding scores fall to 47.66 and 48.01. These results distinguish preservation of text-only computation from retention of general visual understanding: the former follows from the frozen language pathway, whereas the latter depends on the adapted visual tokens and training distribution. The preservation anchor constrains their movement without guaranteeing that every capability is retained, making broader evaluation necessary when transferring the method to new spatial tasks.

5 Conclusion

GeoWorld-VLM demonstrates that predictive world-model representations can improve single-image VLM spatial reasoning through training-time transfer. Camera-conditioned teacher features provide the supervision, while post-projector alignment and interface preservation adapt the visual pathway before a frozen LLM. Across two backbones on What'sUp+VSR and EmbSpatial, the method improves over matched task-only fine-tuning and achieves the highest mean accuracy among the evaluated teachers, retaining the original inference architecture. The evidence remains limited to one world-model teacher, with uneven relation-wise gains and dataset-dependent retention of general multimodal capability. Broader teacher comparisons and direct geometric evaluations would clarify which predictive information is transferred and how it supports spatial decisions.

AI USE STATEMENT

We used an LLM-based assistant for manuscript drafting and editing, literature review. The authors take responsibility for the final manuscript, including all AI-assisted content and scientific claims.

ETHICS STATEMENT

This study evaluates spatial reasoning using existing benchmarks and pretrained models. Improved spatial understanding may benefit assistive and embodied systems, but incorrect spatial judgments can cause harm in navigation, robotics, and other safety-critical applications. The adapted VLM may inherit biases or errors from its training data and teacher, and freezing the language model does not guarantee preservation of general multimodal capability. Our benchmark results therefore do not establish readiness for autonomous deployment. Downstream use should respect the licenses and privacy restrictions of the source datasets and pretrained models, and include task-specific validation and human oversight where errors could affect people. Appendix G discusses limitations and potential broader impacts of the proposed method for future research and deployment.

REPRODUCIBILITY STATEMENT

Section 3 specifies the teacher inputs, alignment and preservation objectives, and trainable and frozen components. Section 4.1 describes the model backbones, benchmark splits, matched baselines, and training protocol. Appendix B provides optimization hyperparameters, loss weights, projection-head dimensions, teacher feature extraction settings, and backbone-specific image preprocessing. Table 5 reports means and standard deviations over three independent seeds. Section 4.7 and Appendix D.7 describe the matched-subset MindJourney comparison and computational costs, including the subset sampling procedure. Additional ablations and cross-benchmark evaluation protocols appear in Appendix D, supporting examination of sensitivity and transfer behavior.

REFERENCES

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.

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.

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. URL https://arxiv.org/abs/2503.01773.

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.

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.

Xiaohu Huang, Jingjing Wu, Qunyi Xie, and Kai Han. 3DRS: MLLMs need 3D-aware representation supervision for scene understanding. *arXiv preprint arXiv:2506.01946*, 2025. URL https://arxiv.org/abs/2506.01946.

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.

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.

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*, pp. 9241–9251, 2025.

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. URL https://arxiv.org/abs/1405.0312.

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

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.

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.

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.

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.

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.

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*, pp. 6617–6626, 2024.

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

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, 2026a. URL https://arxiv.org/abs/2601.20540.

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*, 2026b.

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.

Haochen Wang, Anlin Zheng, Yucheng Zhao, Tiancai Wang, Zheng Ge, Xiangyu Zhang, and Zhaoxiang Zhang. Reconstructive visual instruction tuning. In *International Conference on Learning Representations*, 2025a. URL https://arxiv.org/abs/2410.09575.

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, 2025b. URL https://arxiv.org/abs/2508.18265.

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*, 2025c.

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.

Xianjin Wu, Dingkang Liang, Tianrui Feng, Kui Xia, Yumeng Zhang, Xiaofan Li, Xiao Tan, and Xiang Bai. Generation models know space: Unleashing implicit 3D priors for scene understanding. *arXiv preprint arXiv:2603.19235*, 2026. URL https://arxiv.org/abs/2603.19235.

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*, pp. 14203–14214, 2025a.

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*, 2025b.

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.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, Chanho Eom, Sunghwan Hong, and Seungryong Kim. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025. URL https://arxiv.org/abs/2509.07979.

Bocheng Zou, Mu Cai, Jianrui Zhang, and Yong Jae Lee. Vgbench: Evaluating large language models on vector graphics understanding and generation, 2024. URL https://arxiv.org/abs/2407.10972.

Appendix

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 (Team et al., 2026b). 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 student-side alignment

head. The training objective consists of three terms:

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

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 3 random seeds. 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. Only the student-side head is optimized; the teacher-side head is randomly initialized and remains fixed throughout training. 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.

Backbone / Benchmark	FT	DINOv3	SigLIP2	RADIO	VGGT	GeoWorld-VLM
Gemma / WU+VSR	0.97	0.91	0.98	0.84	0.61	0.41
InternVL / WU+VSR	0.44	1.62	0.65	0.74	0.77	0.57
Gemma / EmbSpatial	1.20	1.43	1.23	1.36	1.80	0.85
InternVL / EmbSpatial	0.52	0.12	0.58	0.41	0.38	0.40

Table 8: Standard deviation of performance across multiple runs on What'sUp+VSR (WU+VSR) and EmbSpatial.

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 STD OF 3 SEEDS

The stds of our experiments over 3 seeds are shown in Tab 8

D ADDITIONAL ABLATION ANALYSIS

We provide additional analyses of teacher feature selection, alignment design, matched teacher baselines, transfer, general multimodal capability, and computational cost. Unless otherwise specified, the ablations use Gemma4 and the What'sUp+VSR suite.

D.1 ANALYSIS OF WORLD-MODEL LAYER SELECTION

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	<u>94.12</u>	34.25	63.64	<u>77.07</u>	<u>65.75</u>	**80.68**	63.07
Gemma4 + Ours, layer 24	**99.51**	93.14	<u>37.46</u>	**70.45**	**78.28**	**76.71**	<u>80.17</u>	**65.45**
Gemma4 + Ours, layer 25	<u>99.03</u>	**95.10**	34.43	<u>66.82</u>	76.03	65.07	79.66	<u>63.10</u>

Table 9: Ablation on world-model feature extraction layers. Layer 24 achieves the best overall performance among the tested intermediate-to-late layers.

The layer selection results in Table 9 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 retain more local appearance information, while a later layer 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 between viewpoint-conditioned spatial structure and transferable visual representation.

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	<u>71.23</u>	78.80	<u>63.00</u>
Gemma4 + Ours, 1 step	98.06	**95.10**	34.43	<u>66.82</u>	<u>76.90</u>	65.07	78.63	<u>63.00</u>
Gemma4 + Ours, 2 steps	**99.51**	<u>93.14</u>	<u>37.46</u>	**70.45**	**78.28**	**76.71**	**80.17**	**65.45**
Gemma4 + Ours, 3 steps	95.63	88.73	34.25	68.18	70.00	<u>71.23</u>	72.31	60.23

Table 10: Ablation on denoising step selection. Using two denoising steps gives the best overall performance, while the three-step state falls below the untouched base model.

D.2 Analysis of Denoising Step Selection

Table 10 shows that teacher features are most effective when extracted after two denoising steps. Zero or one step improves over the base model but reaches only 63.00, while three steps reduce the overall score to 60.23. The denoising stage is therefore the most sensitive tested design choice. One possible explanation is that later denoising states become increasingly tied to video synthesis and may drift away from the original image evidence, making them less suitable for representation transfer.

Configuration choice and subsequent sensitivity analysis. We fixed layer 24 and two denoising steps before inspecting the benchmark evaluation results. The design was informed by prior observations about the usefulness of intermediate generative representations (Wu et al., 2026) and our implementation of LingBot-World-Fast. This configuration was retained across both VLM backbones and benchmark suites. The subsequent experiments on layers 23–25 and denoising steps 0–3 assess sensitivity around this default; they were not used to select the main configuration. These settings are specific to LingBot-World-Fast: VEGA-3D's noise timestep and our number of denoising steps are different quantities. For a new teacher, we recommend calibration on held-out data when adapting the feature extraction configuration.

D.3 Alignment Loss Coefficient

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 11: 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.

Table 11 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.

D.4 Projection Design and Preservation Objective

The teacher-side projection is fixed throughout training and therefore cannot co-adapt with the student. Table 12 evaluates a stricter student-side-only design in which the fixed teacher projection is removed. Performance decreases by 1.10–2.48 points across the four settings. We interpret the fixed projection as a capacity-matching bottleneck that maps the high-dimensional generative hidden state into the shared alignment space while keeping the teacher target deterministic.

Removing the preservation loss reduces the overall score from 65.45 to 63.20. Relative to task-only FT (62.71), world-model alignment without preservation gives a 0.49-point increase, while adding preservation under alignment gives a further 2.25 points. This supports the use of the preserva-

Backbone / Benchmark	GeoWorld-VLM	Student-side-only projection	Drop
Gemma4 / What'sUp+VSR	**65.45**	63.16	−2.29
Gemma4 / EmbSpatial	**78.78**	76.83	−1.95
InternVL3.5-2B / What'sUp+VSR	**61.66**	59.18	−2.48
InternVL3.5-2B / EmbSpatial	**79.70**	78.60	−1.10

Table 12: Ablation of the fixed teacher-side projection. The implemented GeoWorld-VLM uses a trainable student-side alignment head and a randomly initialized but fixed teacher-side projection; removing the fixed teacher projection consistently reduces performance.

Model	Ctrl-A	Ctrl-B	COCO-1	COCO-2	VG-1	VG-2	VSR	Overall
Gemma4 + Ours, w/o preservation	<u>99.03</u>	**94.61**	<u>35.32</u>	68.18	74.66	67.12	78.97	63.20
Gemma4 + VGGT teacher	98.06	93.14	34.61	<u>70.00</u>	<u>77.07</u>	<u>69.86</u>	<u>79.83</u>	<u>63.65</u>
Gemma4 + Ours	**99.51**	<u>93.14</u>	**37.46**	**70.45**	**78.28**	**76.71**	**80.17**	**65.45**

Table 13: Ablation on the teacher choice and preservation objective. Removing the preservation loss degrades overall performance, while the matched VGGT teacher also remains below GeoWorld-VLM in this setting. The best result is shown in bold and the second-best result is underlined.

tion objective within the complete framework, consistent with its intended role of limiting visual-interface drift. As discussed in Section 4.5, these comparisons do not measure preservation without alignment or establish an independent decomposition of their effects. The VGGT comparison is consistent with the broader matched-teacher results in Tables 1 and 2.

D.5 Cross-Benchmark Transfer

Source training set → Target evaluation set	FT	DINOv3	VGGT	GeoWorld-VLM
What'sUp+VSR → EmbSpatial	56.04	58.35	60.30	**61.77**
EmbSpatial → What'sUp+VSR	62.19	62.32	60.16	**62.42**

Table 14: Bidirectional cross-benchmark transfer with Gemma4. Each method is trained exclusively on the source benchmark and evaluated on the target benchmark without target-benchmark training examples.

GeoWorld-VLM achieves the highest transferred accuracy in both directions. When trained on What'sUp+VSR and evaluated on EmbSpatial, it exceeds FT, DINOv3, and VGGT by 5.73, 3.42, and 1.47 points, respectively. In the reverse direction, the margin over DINOv3 is only 0.10 points. We therefore treat these results as preliminary evidence of improved cross-benchmark transfer rather than as a broad cross-dataset generalization claim.

D.6 General Multimodal Capability

Table 15 evaluates whether adapting the visual interface preserves general multimodal capability. For EmbSpatial-trained checkpoints, GeoWorld-VLM gives the strongest adapted-model MMStar result for both backbones, improving Gemma4 over Base by 1.56 points and retaining nearly all InternVL3.5-2B performance (-0.34). The What'sUp+VSR-trained checkpoints show a different pattern and can incur larger degradation, including for GeoWorld-VLM. Thus, freezing the LLM and using the preservation objective constrain, but do not universally eliminate, visual-interface drift.

D.7 Training and Inference Cost

GeoWorld-VLM requires approximately $4.4\times$ the wall-clock time of FT and $1.9\times$ that of DINO; in GPU-hours, the corresponding ratios are approximately $8.8\times$ and $3.8\times$. The additional cost comes from the online forward pass of the frozen LingBot teacher, which receives no gradient updates or optimizer states. This overhead is incurred only during training; the world model and alignment heads are removed afterward.

For the What'sUp-based subset used in the MindJourney comparison, we follow the original sampling pipeline: `random.Random(10).sample(...)` uniformly samples 200 unique exam-

Backbone	Spatial tuning data	Base	FT	DINO	VGGT	GeoWorld-VLM
Gemma4	What'sUp+VSR	50.61	50.69 (+0.08)	49.74 (−0.87)	49.05 (−1.56)	47.66 (−2.95)
Gemma4	EmbSpatial	50.61	50.87 (+0.26)	50.09 (−0.52)	50.61 (+0.00)	**52.17 (+1.56)**
InternVL3.5-2B	What'sUp+VSR	55.37	51.82 (−3.55)	47.83 (−7.54)	**52.86** (−2.51)	48.01 (−7.36)
InternVL3.5-2B	EmbSpatial	55.37	50.52 (−4.85)	52.43 (−2.94)	52.51 (−2.86)	**55.03** (−0.34)

Table 15: MMStar accuracy after spatial adaptation. Parentheses report the change relative to the corresponding Base model. Retention of general multimodal capability is dataset-dependent.

Method	Time / Step	GPUs	Wall-Clock	GPU-Hours
Task-only FT	1.45 s	1	0.6 h	0.6
DINO	3.31 s	1	1.4 h	1.4
GeoWorld-VLM	6.34 s	2	2.6 h	5.3

Table 16: Comparative training cost on EmbSpatial under the same three-epoch schedule (500 steps per epoch).

ples without replacement from the filtered 2,522-example What'sUp test set, which are then divided into eight disjoint chunks of 25 for parallel inference. The main-paper comparison uses the same 200-question subsets for Base+MindJourney, FT+MindJourney, and GeoWorld-VLM. As reported in Table 6, applying MindJourney to the same task-only fine-tuned checkpoints does not close the accuracy gap, while explicit test-time generation incurs substantially higher latency.

E SAT Results by Ego-Motion Split

Table 18 reports SAT results after separating ego-motion questions from the remaining question types. This split is motivated by an important limitation of our current world-model alignment strategy. Although the teacher is generated from a camera-conditioned image-to-video world model, our current implementation aggregates teacher features through mean pooling. This design is effective for extracting stable spatial structure, but it may suppress temporal order and motion direction, both of which are essential for ego-motion reasoning.

The EgoM-only results confirm this limitation. Our method obtains an EgoM score of 52.43, which is lower than FT-only by 3.79 points and lower than DINO by 11.35 points. This suggests that the current world-model representation is not sufficient for modeling ego-centric motion in SAT. In particular, average pooling over generated frames may convert a temporally ordered motion signal into a static scene-level representation, making it difficult for the model to infer self-motion.

For the non-Ego split, our method achieves the best overall score of 58.24, outperforming FT-only by 0.21 points and DINO by 0.83 points. Although the margin over FT-only is modest, the result suggests that world-model alignment is beneficial once ego-motion questions are excluded. The gains are mainly reflected in GoalAim and ObjectM, where our method achieves the best scores among the compared methods. This indicates that the proposed alignment is more suitable for object-centric and goal-oriented spatial reasoning than for ego-motion reasoning.

Overall, the SAT analysis reveals a clear boundary of the proposed method. World-model alignment improves spatial representation on What'sUp+VSR and achieves the best overall result on the non-Ego SAT split, but it struggles with ego-motion reasoning. This observation motivates future extensions that preserve temporal directionality, such as temporal-aware pooling, order-sensitive teacher alignment, or explicit ego-motion supervision.

F 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 19, 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 20 fur-

Backbone / Benchmark	MindJourney Videos / Q	MindJourney Time / Q	GeoWorld-VLM Time / Q
InternVL3.5-2B / What'sUp+VSR	5.17	∼148 s	∼3–5 s
Gemma4 / What'sUp+VSR	5.27	∼101 s	∼3–5 s
InternVL3.5-2B / EmbSpatial	5.19	∼155 s	∼3–5 s
Gemma4 / EmbSpatial	4.89	∼97 s	∼3–5 s

Table 17: End-to-end inference cost for MindJourney and GeoWorld-VLM on the matched subset evaluation. MindJourney invokes the world model approximately five times per question, whereas GeoWorld-VLM uses the standard single-image VLM inference pipeline.

Split	Model	ActCons	EgoM	GoalAim	ObjectM	Perspect	Val Acc.	Test Acc.	Overall
EgoM only	Gemma4 + Ours	–	52.43	–	–	–	51.23	60.87	52.43
	Gemma4 + FT-only	–	<u>56.22</u>	–	–	–	<u>54.94</u>	**65.22**	<u>56.22</u>
	Gemma4 + DINO	–	**63.78**	–	–	–	**63.58**	**65.22**	**63.78**
Non-Ego	Gemma4 + Ours	45.55	–	**69.78**	**73.37**	<u>57.46</u>	58.11	**59.06**	**58.24**
	Gemma4 + FT-only	**46.36**	–	<u>68.13</u>	<u>72.28</u>	<u>57.46</u>	**58.47**	55.12	<u>58.03</u>
	Gemma4 + DINO	<u>45.82</u>	–	63.74	71.20	**60.09**	<u>57.64</u>	<u>55.91</u>	57.41

Table 18: SAT results after separating ego-motion questions from non-ego questions. The EgoM-only split shows that our world-model alignment performs worse than both FT-only and DINO on ego-motion reasoning. On the non-Ego split, our method achieves the best overall accuracy, suggesting that world-model alignment is more effective for non-ego spatial reasoning than for ego-motion reasoning. The best result within each split is shown in bold and the second-best result is underlined.

ther 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.

F.1 Cases Where Only Our Method Is Correct

F.2 Cases Where Our Method Corrects the Raw Model

G Limitations and Broader Impact.

Our matched experiments establish the effectiveness of transfer from LingBot-World-Fast in the evaluated settings. They do not establish a universal advantage of camera-controllable teachers or separate camera conditioning from teacher scale and pretraining. Direct geometric evaluation and a preservation-only control would further clarify what information is transferred and how the objectives interact.

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.

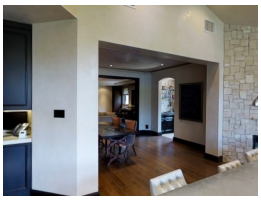
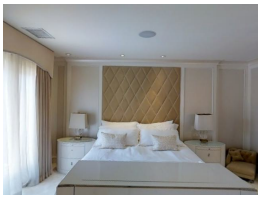
Image	Question / Ground Truth	Model Predictions
	**Q:** In the image, how do the positions of picture and ottoman interact with each other? **GT:** The picture is above the ottoman.	Raw: The picture is on the left side of the ottoman. ✗ FT-only: The picture is on the left side of the ottoman. ✗ DINO: 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 19: Qualitative examples where only our LingBot-teacher alignment method predicts the correct spatial answer.

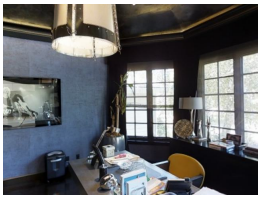
Image	Question / Ground Truth	Model Predictions
	**Q:** Which object, in relation to your current position, holds the nearest placement in the image? **GT:** pillow	Raw: pool table ✗ FT-only: pillow ✓ DINO: pillow ✓ Ours: pillow ✓
	**Q:** What is the spatial configuration between monitor and lamp in 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 20: 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.