GEOFIRST: LEARNING WORLD-CONSISTENT GEOMETRY FOR ROBUST ROBOT MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

World-action models (WAMs) have shown promise for embodied intelligence, but their core value lies in the latent representations learned through prediction rather than explicit future imagination. Existing WAMs rely on RGB video prediction, which lacks 3D geometric structure, while depth-based approaches treat depth as an auxiliary task from limited viewpoints—yielding representations that lack cross-view consistency and degrade under observation perturbations. We present **GC-WAM** (Geometry Consistent World-Action Model), which replaces RGB video prediction with metric depth prediction as the primary future-state supervision signal, encouraging the video foundation model to learn geometry-grounded latent representations rather than appearance-centric ones. GC-WAM introduces per-view depth normalization to preserve geometric detail across camera views, cross-view world consistency via an InfoNCE-based loss for viewpoint-invariant representations, and a geometry-first curriculum that first learns geometry then jointly learns geometry and action. Experiments on LIBERO Plus show that GC-WAM achieves **79.7%** average success rate (**+29.8 pp** over RGB baseline), and achieves **85.4%** on RobotTwin Clean2Clean. **Different Views, Same World.**

1 INTRODUCTION

World-action models (WAMs) (Yuan et al., 2026; Yang et al., 2026a; Li et al., 2026b) jointly learn future observation prediction and action generation, bringing video-model priors to robot control. Recent work (Yuan et al., 2026; DYNA, 2026) show that predictive training can benefit policies that generate actions without synthesizing future video at deployment. For manipulation, where object geometry, contact surfaces, and spatial relationships matter, we ask: *do existing predictive objectives learn sufficiently accessible 3D information, and does that information remain useful when the viewpoint changes?*

We investigate this question through a controlled study of frozen features, comparing RGB, RGB-D, and depth prediction under matched backbones, data, and camera augmentation, and examining the effect of cross-view consistency. We examine two aspects of representation quality. First, RGB-predictive Fast-WAM features yield weaker linear geometry readouts than geometry-supervised variants, indicating that RGB prediction alone leaves relevant 3D information less accessible. Recent methods introduce geometric information through future-depth supervision in WAM4D (Li et al., 2026b), trajectory-field distillation in 4D-WAM (Yang et al., 2026a), or a geometric foundation backbone in GAM (Han et al., 2026). Second, RGB-D supervision improves in-view geometric performance, but substantial deficits remain under unseen camera poses. Even at comparable in-view depth-prediction quality, GC-WAM yields stronger geometric readouts and more reliable cross-view correspondences. These observations motivate a distinction between predicting per-view depth and learning geometry that generalizes across views. Among variants that predict future depth, we additionally relate depth accuracy, frozen-feature geometry, and policy performance. Across these controlled variants, lower depth-prediction error and stronger geometric readouts and correspondence scores are associated with higher policy success, particularly under camera perturbations (see figure 1).

These diagnostics motivate two complementary properties of representations for robust control. First, *geometric accessibility*: task-relevant 3D structure should be readily recoverable from the features available to the action head, addressing the limited geometric readout of RGB-predictive

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.

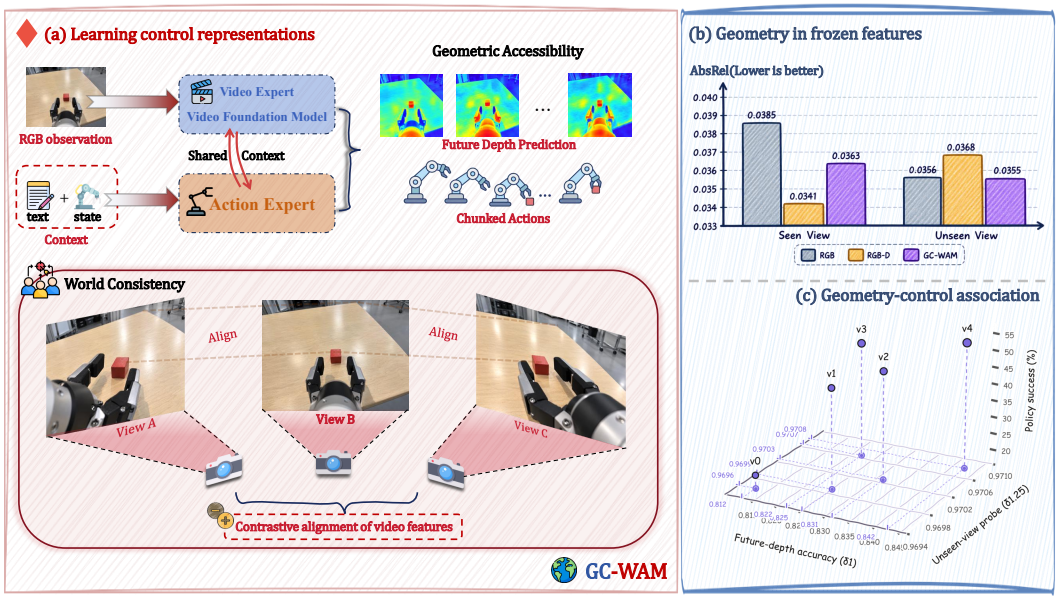


Figure 1: **(a)** Overview of GC-WAM: given a single RGB observation and a language instruction, the model jointly predicts future depth video and action trajectories, with cross-view world consistency enforcing viewpoint-invariant representations. **(b)** Depth probe results comparing frozen features across prediction objectives. **(c)** Correlation between future depth prediction quality, unseen-view depth probe accuracy, and policy success rate—variants with better depth prediction and stronger cross-view geometric readouts achieve higher task success.

features. Second, *world consistency*: local descriptors of the same physical surface point should remain matchable across viewpoints, while retaining each view's spatial layout. As in binocular correspondence, this property connects different projections of the same surface and addresses the cross-view gap left by per-view depth supervision. Together, these properties guide both the geometric content of control representations and their correspondence across observations.

To learn representations with these properties, we present **GC-WAM** (Geometry Consistent World-Action Model), a world-action model that embodies this principle through native 3D world modeling. GC-WAM replaces RGB video prediction with **metric depth prediction** as the primary future-state supervision signal, leveraging the video foundation model's generative capacity to **lift 2D RGB observations into 3D depth representations**. By training the video diffusion transformer to generate future depth frames, the prediction objective encourages its latent representations to encode 3D spatial structure rather than surface appearance, yielding a richer, **geometry-grounded** latent space that the action expert can exploit. Furthermore, we introduce a **cross-view world consistency** objective using an InfoNCE-based geometric consistency loss between different viewpoints, and a **geometry-first curriculum** with two-stage training—first learning geometry alone, then jointly learning geometry and action—to stabilize the emergence of robust 3D representations.

Our contributions are as follows:

- We present a controlled study of predictive supervision in WAMs, examining the geometric accessibility of frozen features, their consistency across viewpoints, and their association with policy performance.
- We introduce GC-WAM, which combines future-depth prediction with bidirectional, calibration-guided alignment of corresponding local features. These objectives shape the video representations used for control while retaining RGB-to-action deployment without future-depth generation.

- We propose a representation-level constraint using an InfoNCE-based geometric consistency loss between different viewpoints of the same scene, encouraging the world model to learn viewpoint-invariant geometric representations.
- We develop a geometry-first, two-stage training scheme that first optimizes depth prediction and cross-view consistency with the action expert frozen, then jointly optimizes geometry and action while retaining both geometric objectives.

We evaluate GC-WAM on the LIBERO Plus benchmark and RobotTwin, demonstrating that it achieves a **79.7%** average success rate on LIBERO Plus, representing a **+29.8 pp** improvement over the RGB baseline across all seven challenge dimensions, and **85.4%** under Clean2Clean evaluation on RobotTwin.

2 Related Works

2.1 World Models for Robot Manipulation

World models have emerged as a promising paradigm for robot learning, enabling agents to predict future states and plan actions in learned dynamics models. Early work in this direction includes Dreamer (Hafner et al., 2019), which learns a latent dynamics model for model-based reinforcement learning, and TD-MPC (Hansen et al., 2022), which combines temporal difference learning with model predictive control. More recently, world-action models (WAMs) have been proposed to jointly learn world dynamics and action prediction in a unified framework. Fast-WAM (Yuan et al., 2026) demonstrates that the core value of WAMs lies not in explicit future imagination at inference time, but in the robust latent representations learned during training through the prediction task itself. Subsequent works such as 4D-WAM (Yang et al., 2026a), GAM (Han et al., 2026), and VGA (Song et al., 2026) have explored various architectural improvements and training strategies to enhance the quality of learned representations. Our work builds upon this line of research by introducing **Geometry Consistent World-Action Model (GC-WAM)**, which emphasizes learning viewpoint-invariant representations that capture the intrinsic dynamics of the physical world.

2.2 Depth-based Representations for Manipulation

Depth information has been widely used in robotic manipulation to provide geometric understanding of the scene. Prior works have incorporated depth or RGB-D observations into robotic perception systems, using them for open-set 3D spatial reasoning (Jatavallabhula et al., 2023), object pose estimation and tracking (Wen et al., 2024), and grasp perception (Fang et al., 2023). However, these approaches use depth as an input to downstream perception systems, rather than as the primary future-state prediction target of a world model. In contrast, our work proposes to use metric depth prediction as the **future-state supervision signal** for the world model, replacing RGB video prediction entirely. This design choice encourages the video expert to learn 3D-aware latent representations that explicitly encode scene geometry, providing a more robust foundation for action prediction. Our experiments demonstrate that this depth-first approach yields significant improvements over RGB-based baselines across multiple benchmark dimensions.

2.3 Viewpoint-Invariant Representation Learning

Learning representations that are robust to viewpoint changes is a fundamental challenge in computer vision and robotics. Prior works have explored various approaches to achieve viewpoint invariance, including contrastive learning (Chen et al., 2020), view-conditioned representation learning (Wang et al., 2026), and zero-shot novel view synthesis for policy learning (Tian et al., 2024). In the context of robotic manipulation, viewpoint robustness is particularly important as robots must operate from diverse camera perspectives in real-world environments. Recent works such as AnyCamVLA (Heo et al., 2026) have specifically addressed zero-shot camera adaptation for vision-language-action models. The LIBERO Plus Benchmark (Fei et al., 2025) evaluates this capability through its Camera dimension, which tests robustness to unseen viewpoint variations. Our work contributes to this direction by encouraging viewpoint-invariant representations through both data augmentation and representation-level consistency objectives, as detailed in section 3.2.

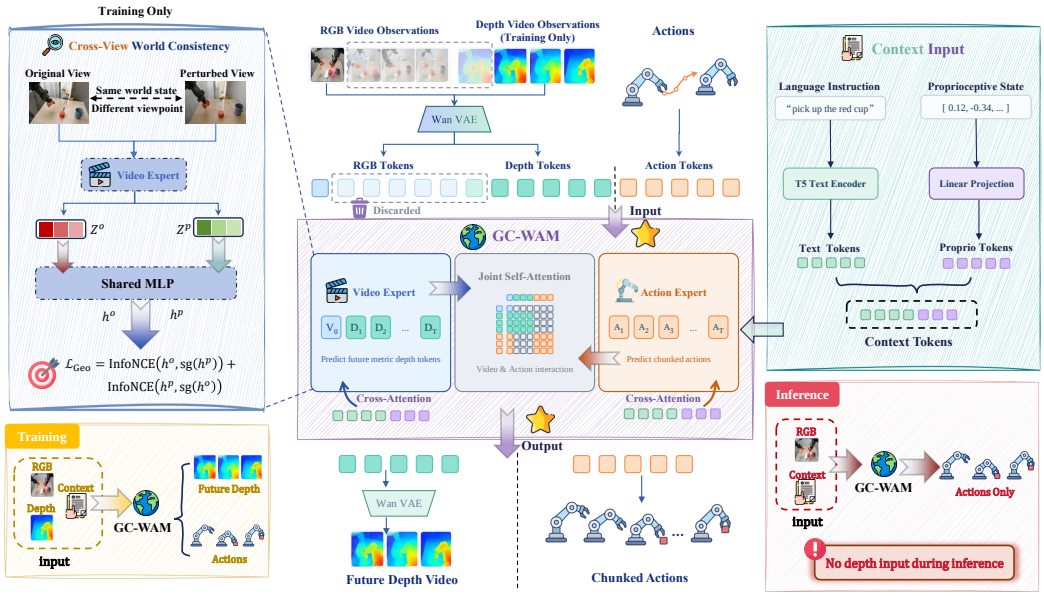


Figure 2: Overview of GC-WAM architecture. The model takes a single RGB first-frame observation and a language instruction as input, and jointly predicts future depth video frames and action trajectories through a Mixture-of-Transformers backbone.

3 METHODS

We present **GC-WAM** (Geometry Consistent World-Action Model), a world-action model that learns **world-consistent geometric representations** for robust robot manipulation. We begin with a formal problem definition in section 3.1, analyze the limitations of existing formulations, and introduce our improved design. We then detail the GC-WAM architecture in section 3.2 and describe training and inference in section 3.3.

3.1 PRELIMINARY

World-Action Models (WAMs). A world-action model (WAM) (Yuan et al., 2026; Yang et al., 2026a; Li et al., 2026b) jointly learns visual dynamics and robot actions in a unified framework. Formally, given an initial RGB observation $o_0 \in \mathbb{R}^{H\times W\times 3}$ and a language instruction ℓ, a WAM parameterized by θ learns a joint mapping:

$$\mathcal{M}_\theta : (o_0, \ell) \longrightarrow (\hat{V}_{1:T},\ \hat{A}_{1:T}), \tag{1}$$

where $\hat{V}_{1:T} = \{\hat{v}_1, \ldots, \hat{v}_T\}$ is a predicted future video rollout of length T representing the visual evolution of the scene, and $\hat{A}_{1:T} = \{\hat{a}_1, \ldots, \hat{a}_T\}$ is a chunked action trajectory for the robot to execute. The two prediction targets are handled by separate expert modules that interact through joint self-attention, enabling the visual dynamics prediction to inform action prediction via shared latent representations. As argued by Yuan et al. (2026), the core value of WAMs lies not in the explicit video rollout at inference time, but in the **latent representations learned through the prediction task itself**.

Limitations and our formulation. Existing WAMs (Yuan et al., 2026; Yang et al., 2026a; Li et al., 2026b) predict future frames in RGB space ($\hat{v}_t \in \mathbb{R}^{H\times W\times 3}$), which has two key limitations: (1) RGB supervision encourages *appearance-centric* representations (textures, lighting) rather than the *geometric* structure (3D shape, spatial relationships) critical for manipulation; and (2) there is no mechanism to enforce **cross-view consistency**.the principle that different observations x^{v_1}, x^{v_2} of the same world state should yield similar representations $\boldsymbol{z}^{v_1} \approx \boldsymbol{z}^{v_2}$.

We address both issues by replacing the RGB rollout with a **metric depth** rollout and adding a cross-view consistency objective. Our GC-WAM learns:

$$\mathcal{M}_\theta : (o_0, \ell) \longrightarrow (\hat{D}_{1:T},\ \hat{A}_{1:T}), \tag{2}$$

where $\hat{D}_{1:T} = \{\hat{d}_1, \ldots, \hat{d}_T\}$ with $\hat{d}_t \in \mathbb{R}^{H\times W}$ is a predicted future metric depth map. Depth supervision naturally encodes **future dynamics**—how 3D spatial structure evolves—in a texture-agnostic manner, yielding geometry-grounded latent representations. The overall objective is:

$$\min_\theta \underbrace{\mathcal{L}_{\text{depth}}(\hat{D}_{1:T}, D_{1:T})}_{\text{depth prediction}} + \lambda_{\text{action}} \underbrace{\mathcal{L}_{\text{action}}(\hat{A}_{1:T}, A_{1:T})}_{\text{action prediction}} + \lambda_{\text{Geo}} \underbrace{\mathcal{L}_{\text{Geo}}(\boldsymbol{z}^o, \boldsymbol{z}^p)}_{\text{cross-view consistency}}, \tag{3}$$

where $D_{1:T}$ denotes ground-truth depth, $\boldsymbol{z}^o$ and $\boldsymbol{z}^p$ are latent representations from different viewpoints of the same scene, and λ_{action}, λ_{Geo} are loss weights. We detail each component below.

3.2 GC-WAM: Geometry Consistent World-Action Model

We now describe our approach to achieve the above objective. As illustrated in figure 2, GC-WAM is a single unified transformer that, conditioned on a single RGB first-frame observation and a language instruction, jointly denoises a future depth-video rollout and the corresponding chunked action trajectory.

3.2.1 Future-Depth Predictive Learning

We replace the video branch's future RGB targets with metric-depth-derived targets. Predicting their evolution supplies an appearance-reduced learning signal for visible scene geometry. The RGB observation remains the conditioning input, so depth serves as a training target rather than a required sensor modality at deployment.

Per-view depth encoding. To accommodate different depth distributions across cameras, we normalize metric depth d with fixed camera-specific parameters:

$$u_k(d) = \text{clip}\left(\alpha_k \log\left(1 + \frac{d}{m_k}\right),\ 0,\ 1\right), \tag{4}$$

where k indexes the camera stream and m_k is a reference depth in meters. We use $(\alpha_k, m_k) = (1.0, 1.0\,\text{m})$ for the agent camera and $(0.9, 0.2\,\text{m})$ for the wrist camera.

Channel replication. We replicate u_k across three channels and rescale it to $[-1, 1]$ to match the video VAE input format. The depth sequence is VAE-encoded, and its first latent slot is replaced by the independently encoded current RGB observation. This conditioning slot remains clean and is excluded from the depth prediction loss. The scalar normalization has an analytical inverse within its unsaturated range, but clipping and VAE reconstruction preclude exact metric-depth recovery.

Predictive objectives. Both experts are trained with flow matching. Let x_0^D denote the clean future-depth latents and x_0^A the normalized action chunk. For modality $m \in \{D, A\}$, Gaussian noise ϵ^m and noise level σ_m define

$$x^m_{\sigma_m} = (1 - \sigma_m)x_0^m + \sigma_m \epsilon^m. \tag{5}$$

The experts predict the corresponding velocity targets:

$$\mathcal{L}_m = \mathbb{E}\Big[w_m(\sigma_m)\, \|v_\theta^m(\mathcal{X}_\sigma, c_0) - (\epsilon^m - x_0^m)\|^2_{\text{valid}}\Big], \tag{6}$$

where $\mathcal{X}_\sigma$ contains the noisy depth and action sequences and their noise levels, w_m is the scheduler's timestep weighting, and the squared error is averaged over valid prediction elements. We write $\mathcal{L}_{\text{depth}} = \mathcal{L}_D$ and $\mathcal{L}_{\text{action}} = \mathcal{L}_A$. The depth objective is evaluated in VAE latent space, rather than as a direct regression loss on decoded metric-depth images.

3.2.2 Cross-View World Consistency

Per-view depth prediction does not directly specify how features should relate across cameras. We define *world consistency* through local geometric correspondences: descriptors of the same scene point should align across viewpoints, while each view retains its spatial layout.

Paired observations and features. We construct paired training sequences by rendering the same demonstration states at the original and perturbed agent-camera poses. Each sample uses one perturbed view from the left/right camera perturbations. Reprojection establishes feature correspondences between these observations; it is not used to synthesize the target RGB image. We run the paired sequences through the shared MoT with matched noise realizations and diffusion timesteps, and extract final-block video features. Denote the agent-camera token maps by $Z^v \in \mathbb{R}^{T_{\text{lat}} \times S \times C}$ for $v \in \{o, p\}$, where S indexes spatial tokens and the batch dimension is omitted.

Geometric correspondence. For a source view v and target view w, let r_i be the image coordinate of source token i. Source metric depth and camera calibration give

$$q_{t,i}^{v \to w} = \pi_w\big(T_{w \leftarrow v}\, \pi_v^{-1}\big(r_i, D_t^v(r_i)\big)\big)\,, \tag{7}$$

where π_v^{-1} back-projects a pixel and its depth using the source intrinsics, $T_{w \leftarrow v}$ transforms the resulting point between camera coordinates, and π_w projects it into the target view. Here, D_t^v is the metric-depth frame associated with latent time index t. We map $q_{t,i}^{v \to w}$ to the target token grid and bilinearly sample its feature. The valid set $\mathcal{V}_t^{v \to w}$ excludes invalid or out-of-range source depths, points behind the target camera, and projections outside the target image or token grid. These geometric validity checks do not include an explicit target-depth occlusion test.

Consistency objectives. A shared linear projection $g_\phi : \mathbb{R}^C \to \mathbb{R}^{64}$ produces contrastive descriptors. With norm denoting L2 normalization, we define

$$\begin{aligned} a_{t,i}^v &= \mathrm{norm}\,\big(g_\phi(Z_{t,i}^v)\big), \\ b_{t,i}^{v \to w} &= \mathrm{norm}\big(g_\phi\big(\mathrm{sg}\big[\mathrm{Bilinear}(Z_t^w, q_{t,i}^{v \to w})\big]\big)\big)\,. \end{aligned} \tag{8}$$

The stop-gradient operator acts on the sampled target feature before projection; the shared projection remains trainable from both descriptor branches. For each valid source token,

$$\ell_{t,i}^{v \to w} = -\log \frac{\exp\big((a_{t,i}^v)^\top b_{t,i}^{v \to w}/\tau\big)}{\sum_{j \in \mathcal{V}_t^{v \to w}} \exp\big((a_{t,i}^v)^\top b_{t,j}^{v \to w}/\tau\big)}, \tag{9}$$

where $\tau = 0.1$. The positive is the geometrically matched target descriptor; negatives are the descriptors sampled at the other valid reprojected locations in the same sequence and timestep. Averaging over valid anchors across time and batch gives $\mathcal{L}_{v \to w}$. We apply the constraint in both directions:

$$\mathcal{L}_{\text{Geo}} = \mathcal{L}_{o \to p} + \mathcal{L}_{p \to o}. \tag{10}$$

Each direction uses its own source-view depth and validity set. Both backbone passes retain gradients for their anchor descriptors; stopping gradients through the target features is specific to each directional term. Thus, both views contribute backbone gradients through their respective anchor roles.

3.3 Geometry-First Two-Stage Training

Our curriculum separates geometric representation learning from subsequent joint adaptation to control. Stage I optimizes future-depth prediction and bidirectional correspondence with the action expert frozen; Stage II introduces action supervision while retaining both geometric objectives.

Stage I: geometric representation learning. We freeze the action expert and train the video expert and shared projection head with

$$\mathcal{L}_{\text{I}} = \mathcal{L}_{\text{depth}} + \lambda_{\text{Geo}} \mathcal{L}_{\text{Geo}}. \tag{11}$$

Both geometric objectives are active throughout this stage, and no action prediction loss is applied.

Stage II: joint geometry–action learning. Starting from Stage I, we unfreeze the action expert and jointly optimize both experts and the projection head:

$$\mathcal{L}_{\text{II}} = \mathcal{L}_{\text{I}} + \lambda_{\text{action}} \mathcal{L}_{\text{action}}, \tag{12}$$

We use $\lambda_{\text{action}} = 1.0$ and keep $\lambda_{\text{Geo}} = 0.1$ in both stages. Depth and action prediction losses are computed on the original-view sequence, while the paired perturbed sequence supplies the additional features and depth required for bidirectional geometric alignment.

Table 1: **LIBERO Origin** (Liu et al., 2023) results (% success rate). **Bold**: best; underline: second best.

Model	Spatial	Object	Goal	Long	Avg.
VLA-based					
WorldVLA (Cen et al., 2025)	87.6	96.2	83.4	60.0	81.8
SpatialVLA (Qu et al., 2025)	88.2	89.9	78.6	55.5	78.1
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	98.8	98.2	98.0	92.4	96.9
GeoVLA (Sun et al., 2025)	98.4	99.0	96.6	96.6	97.7
GeoAwareVLA (Abouzeid et al., 2025)	95.0	**100**	**99.0**	93.0	96.8
X-VLA (Zheng et al., 2026)	98.2	98.6	97.8	**97.6**	98.1
WAM-based					
Fast-WAM (Yuan et al., 2026)	98.2	**100**	97.0	95.2	97.6
FastWAM-Joint (Yuan et al., 2026)	**99.6**	99.0	98.4	95.8	98.2
4D-WAM (Yang et al., 2026a)	98.8	99.5	98.6	97.4	**98.6**
GAM (Han et al., 2026)	98.6	99.6	98.6	94.6	97.6
VGA (Song et al., 2026)	99.0	99.6	98.6	95.0	98.1
Motus (Bi et al., 2026)	96.8	99.8	96.6	**97.6**	97.7
GC-WAM (Ours)	98.0	**100**	97.6	95.6	97.8

4 EXPERIMENTS

We evaluate GC-WAM on the **LIBERO Plus Benchmark** (Liu et al., 2023), a comprehensive robot manipulation benchmark that spans seven distinct challenge dimensions: layout variations, camera viewpoint changes, robot configuration differences, language instruction diversity, lighting condition changes, background variations, and sensor noise. This benchmark provides a rigorous testbed for assessing the robustness of world-action models across diverse visual and structural perturbations.

4.1 BASELINES

We compare our method against **Fast-WAM** (Yuan et al., 2026), the original RGB-based world-action model that predicts future RGB video frames and action trajectories jointly. Our variant replaces the RGB video prediction target with metric depth video prediction, using log-median depth normalization with per-view parameterization (see section 3.2.1 for details). We refer to this configuration as **GC-WAM** throughout.

4.2 MAIN RESULTS

4.2.1 LIBERO RESULTS

We evaluate GC-WAM on the LIBERO Origin and LIBERO Plus benchmarks. On LIBERO Origin (Table 1), GC-WAM achieves **97.8%** average success rate, competitive with state-of-the-art methods. On the more challenging LIBERO Plus benchmark (Table 2), which spans seven distinct challenge dimensions, GC-WAM achieves **79.7%** weighted average—a **+29.8 pp** improvement over the RGB baseline (49.9%). GC-WAM outperforms all baselines on Layout, Language, and Sensor dimensions. On the Camera dimension, GC-WAM (67.7%) trails GAM (83.1%), likely because GAM inherits multi-view geometric priors from its 3D reconstruction backbone, whereas GC-WAM learns viewpoint invariance purely from the cross-view consistency objective.

4.2.2 ROBOTWIN RESULTS

We further evaluate on the RobotTwin Benchmark (Chen et al., 2025) under Clean2Clean and Clean2Random protocols (Table 3). GC-WAM achieves **85.4%** Clean2Clean, surpassing all baselines. However, Clean2Random drops to **4.7%**, which we attribute to training from scratch: 4D-WAM leverages a pretrained video foundation model with diverse prior knowledge, while GC-WAM learns entirely on the target domain. This suggests combining our depth objective with pretrained video models as a promising direction.

Table 2: **LIBERO Plus** (Fei et al., 2025) Benchmark results (% success rate). Avg: weighted average. **Bold**: best; underline: second best.

Method	Layout	Camera	Robot	Language	Light	Background	Sensor	Avg
Fast-WAM (Yuan et al., 2026)	61.4	16.5	44.5	67.0	79.9	52.6	38.2	49.9
Fast-WAM-Joint (Yuan et al., 2026)	74.8	27.9	60.9	89.3	86.7	49.3	51.1	62.2
4D-WAM (Yang et al., 2026a)	79.2	45.2	64.3	90.6	94.3	57.7	69.1	71.0
GAM (Han et al., 2026)	79.1	**83.1**	**70.0**	84.8	**97.2**	**94.3**	**95.3**	**85.5**
GC-WAM (Ours)	**83.0**	67.7	67.2	**93.6**	94.0	68.8	84.5	79.7

Table 3: **RobotTwin 2.0** (Chen et al., 2025) results (% success rate). Models are trained on clean data only. **Bold**: best; underline: second best.

Method	Clean2Clean	Clean2Random	Average
VLA-based			
GR00T-N1.7 (NVIDIA et al., 2025)	43.6	20.7	32.2
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	68.5	**46.0**	57.3
Abot-M0 (Yang et al., 2026b)	70.7	36.0	53.4
StarVLA (Community, 2026)	58.1	10.6	34.3
WAM-based			
Lingbot-VA (Li et al., 2026a)	80.7	34.6	57.7
Fast-WAM (Yuan et al., 2026)	70.2	1.2	35.7
4D-WAM (Yang et al., 2026a)	81.5	41.8	**61.7**
GC-WAM (Ours)	**85.4**	4.7	45.1

4.3 Ablation Studies

We conduct controlled ablation studies to investigate the contribution of each component in our approach.

4.3.1 Cross-View World Consistency

We first investigate the contribution of cross-view world consistency. As shown in Table 4, we progressively introduce bidirectional geometric consistency and direct supervision on the perturbed view.

Adding bidirectional geometric consistency improves from **75.0%** to **78.7%** (+3.7 pp), and further adding perturbed-view supervision reaches **79.7%**. These results confirm that constraining representations across viewpoints provides strong regularization, and that direct supervision on diverse views is complementary to representation-level consistency. We also compare against camera augmentation (Base + Cam. Aug), which mixes data from three viewpoints and thus trains the model on all three views with $3\times$ the data volume and correspondingly longer training steps. Despite using only half the data exposure on the perturbed viewpoint and no additional training steps, our approach (**+ Perturb VA**) achieves comparable overall performance (**79.7%** vs. **79.8%**) with a more balanced per-dimension profile (see Table 6). Specifically, while camera augmentation excels on Camera (+6.8 pp) and Light (+2.3 pp) due to direct viewpoint exposure, it degrades on Robot (-6.0 pp) and Background (-3.6 pp), likely because the augmented viewpoints introduce visual artifacts that interfere with morphology-sensitive and scene-level reasoning. In contrast, our geometric consistency objective maintains stable performance across all dimensions, suggesting that structured geometric constraints provide a more robust mechanism for learning viewpoint-invariant representations than brute-force data scaling.

Table 4: **LIBERO Plus** (Fei et al., 2025) ablation study of cross-view world consistency (% success rate). The origin view is always supervised by the video and action objectives. "Geo" denotes bidirectional geometric consistency. "Perturb VA" denotes video and action supervision on the perturbed view.

Method	Origin VA	Geo	Perturb VA	Cam. Aug	Avg. ↑
Base	✓				75.0
+ Bi-directional Geo	✓	✓			78.7
+ Perturb VA	✓	✓	✓		79.7
Base + Cam. Aug	✓			✓	79.8

Table 5: **LIBERO Plus** (Fei et al., 2025) ablation of the two-stage training strategy (% success rate). Stage 0: video-only pre-training (Action DiT frozen). Stage 1: joint video–action training. **Bold**: best.

Training Paradigm	Avg.
Stage 1 only (Joint from scratch)	58.6
Stage 0 +1 (Ours)	**79.7**

4.3.2 Two-Stage Training Strategy

We evaluate the contribution of our two-stage geometry-first curriculum (section 3.3). As shown in Table 5, we compare two training paradigms on the LIBERO Plus benchmark: (1) **Stage 1 only**, which skips video pre-training and directly trains both the Video DiT and Action DiT jointly from scratch; and (2) **Stage 0 +1**, which first pre-trains the Video DiT alone with the depth prediction objective (Stage 0), and then jointly fine-tunes both experts end-to-end (Stage 1).

Training both objectives jointly from scratch (Stage 1 only) yields a lower success rate, as the action gradients interfere with the video expert's early optimization landscape, preventing it from settling into a well-structured geometric latent space. In contrast, our full two-stage curriculum (Stage 0 +1) first establishes geometry-grounded representations through depth prediction alone, then introduces action supervision on this stable foundation. This **geometry-first curriculum** achieves **79.7%** average success rate, substantially outperforming Stage 1 only. The result confirms that *decoupling* geometric representation learning from action learning is critical: pre-training the video expert without action gradients allows it to develop robust 3D scene understanding, which in turn provides a stronger basis for the action expert to learn precise spatial correspondences.

5 Conclusions

We presented **GC-WAM**, a world-action model that replaces RGB video prediction with **metric depth prediction** to learn **world-consistent geometric representations** for robot manipulation. Our key insight is that a robust representation should capture viewpoint-invariant world dynamics, since the required action remains unchanged regardless of camera viewpoint. To this end, GC-WAM introduces **per-view depth normalization** to preserve geometric detail across cameras with different depth distributions, a **cross-view world consistency** objective that aligns local features across viewpoints via calibrated geometric correspondences, and a **geometry-first curriculum** that first learns geometry before jointly optimizing geometry and action. Experiments on LIBERO Plus demonstrate a **+29.8 pp** improvement over the RGB baseline across all challenge dimensions, and GC-WAM achieves **85.4%** on RobotTwin Clean2Clean, confirming that depth-supervised, cross-view-consistent representations provide a stronger foundation for manipulation. **Different Views, Same World.**

AI USE STATEMENT

In this work, we used generative AI tools for paper writing refinement and code analysis. Specifically, we used AI-assisted tools for language polishing and improving the clarity of the manuscript. We have not used generative AI tools for research idea generation, experiment design, or data analysis. All AI-assisted text has been reviewed and verified by the authors. We take full responsibility for the final content of this work, including text, claims, and artifacts.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions to facilitate reproduction of our results. The model architecture and all key design choices, including per-view depth normalization, the cross-view world consistency objective, and the geometry-first training curriculum, are specified in section 3. Training hyperparameters and implementation details are described in section 4. Additional ablation studies and depth normalization alternatives are provided in the appendix. Our experiments are conducted on the publicly available LIBERO Plus (Fei et al., 2025) and RobotTwin (Chen et al., 2025) benchmarks. We plan to release the code and trained checkpoints upon acceptance.

REFERENCES

Ali Abouzeid, Malak Mansour, Qinbo Sun, Zezhou Sun, and Dezhen Song. Geoaware-vla: Implicit geometry aware vision-language-action model. *arXiv preprint arXiv:2509.14117*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

Jun Cen, Chaohui Yu, Hangjie Yuan, Yuming Jiang, Siteng Huang, Jiayan Guo, Xin Li, Yibing Song, Hao Luo, Fan Wang, et al. Worldvla: Towards autoregressive action world model. *arXiv preprint arXiv:2506.21539*, 2025.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

StarVLA Community. Starvla: A lego-like codebase for vision-language-action model developing. *arXiv preprint arXiv:2604.05014*, 2026.

DYNA. Dyna-2: A 1-million-hour scaling law for world-action models, 2026. URL `https://www.dyna.co/dyna-2`.

Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains. *IEEE Transactions on Robotics*, 39(5):3929–3945, 2023.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, et al. Libero-plus: In-depth robustness analysis of vision-language-action models. *arXiv preprint arXiv:2510.13626*, 2025.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Jisang Han, Seonghu Jeon, Jaewoo Jung, René Zurbrügg, Honggyu An, Tifanny Portela, Marco Hutter, Marc Pollefeys, Seungryong Kim, and Sunghwan Hong. Geometric action model for robot policy learning. *arXiv preprint arXiv:2606.17046*, 2026.

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. *arXiv preprint arXiv:2203.04955*, 2022.

Hyeongjun Heo, Seungyeon Woo, Sang Min Kim, Junho Kim, Junho Lee, Yonghyeon Lee, and Young Min Kim. Anycamvla: Zero-shot camera adaptation for viewpoint robust vision-language-action models. *arXiv preprint arXiv:2603.05868*, 2026.

Krishna Murthy Jatavallabhula, Alihusein Kuwajerwala, Qiao Gu, Mohd Omama, Tao Chen, Alaa Maalouf, Shuang Li, Ganesh Iyer, Soroush Saryazdi, Nikhil Keetha, et al. Conceptfusion: Open-set multimodal 3d mapping. *arXiv preprint arXiv:2302.07241*, 2023.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026a.

Ying Li, Xiaobao Wei, Jiajun Cao, Hao Wang, Xiaowei Chi, Chengyu Bai, Qianpu Sun, Jiajun Li, Xiaojie Zhang, Peidong Jia, et al. Wam4d: Fast 4d world action model via spatial register tokens. *arXiv preprint arXiv:2606.14048*, 2026b.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems*, 2023.

NVIDIA, Guanzhi Wang, Fanqi Lin, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_0.5$: a vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Delin Qu, Haoming Song, Qizhi Chen, Yuanqi Yao, Xinyi Ye, Yan Ding, Zhigang Wang, JiaYuan Gu, Bin Zhao, Dong Wang, et al. Spatialvla: Exploring spatial representations for visual-language-action model. *arXiv preprint arXiv:2501.15830*, 2025.

Zijian Song, Qichang Li, Jiawei Zhou, Zhenlong Yuan, Tianshui Chen, Liang Lin, and Guangrun Wang. Robotic manipulation is vision-to-geometry mapping ($f(v) \rightarrow g$): Vision-geometry backbones over language and video models. *arXiv preprint arXiv:2604.12908*, 2026.

Lin Sun, Bin Xie, Yingfei Liu, Hao Shi, Tiancai Wang, and Jiale Cao. Geovla: Empowering 3d representations in vision-language-action models. *arXiv preprint arXiv:2508.09071*, 2025.

Stephen Tian, Blake Wulfe, Kyle Sargent, Katherine Liu, Sergey Zakharov, Vitor Guizilini, and Jiajun Wu. View-invariant policy learning via zero-shot novel view synthesis. *arXiv preprint arXiv:2409.03685*, 2024.

Qian Wang, Longrui Chen, Peiran Sun, Aleksandar Taranovic, Niklas Freymuth, Ge Li, Weiran Liao, CF Nagy, Yucheng Tan, Tao Chen, et al. Rayvit: Ray-conditioned visual representations for viewpoint-robust imitation learning. *arXiv preprint arXiv:2607.29622*, 2026.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Lishan Yang, Wenxuan Song, Xi Wang, Pingyue Sheng, Zheng Fang, Ziyang Zhou, Junjie He, Haodong Yan, Jiayi Chen, Nan Sun, et al. 4d-wam: Infusing spatiotemporal awareness into world action models through trajectory fields. *arXiv preprint arXiv:2608.08023*, 2026a.

Yandan Yang, Shuang Zeng, Tong Lin, Xinyuan Chang, Dekang Qi, Junjin Xiao, Haoyun Liu, Ronghan Chen, Yuzhi Chen, Dongjie Huo, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026b.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-wam: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, Dongxiu Liu, Xirui Kang, Yuchun Feng, Yinan Zheng, Jiayin Zou, Yilun Chen, Jia Zeng, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. In *International Conference on Learning Representations*, volume 2026, pp. 60580–60606, 2026.

A LIMITATIONS

We discuss two main limitations of our work.

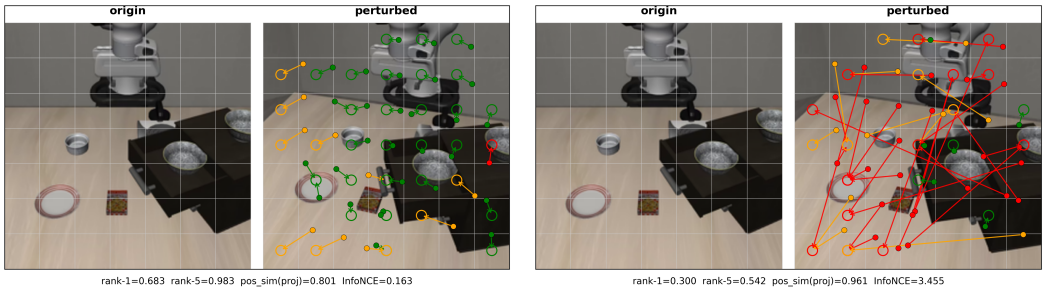
Dependence on high-quality depth data. GC-WAM requires synchronized RGB and metric depth data for training, where depth serves as the primary future-state supervision signal. Acquiring such data at scale is more challenging than collecting RGB-only video: it either requires depth sensors (e.g., RealSense cameras) tightly synchronized with RGB streams, or accurate off-the-shelf depth estimators applied to existing datasets. This limits the applicability of our method to domains where high-quality depth annotations are readily available.

Limited scale of training. Our current implementation builds on Fast-WAM (Yuan et al., 2026) and is trained on the target domain datasets without leveraging large-scale pretrained video foundation models. As shown in our RobotTwin experiments (Table 3), this results in poor generalization under distribution shifts (Clean2Clean 85.4% vs. Clean2Random 4.7%). Scaling the depth-supervised training to diverse, large-scale video datasets—similar to how general-purpose video foundation models are pretrained—is likely necessary to learn broadly transferable geometric representations.

B CAMERA VIEWPOINT AUGMENTATION

Perturbed viewpoint construction. We describe the construction of perturbed viewpoints used in the cross-view world consistency objective (section 3.2.2). Given the original agent-view camera, we generate two perturbed viewpoints by orbiting the camera 15° leftward and 15° rightward around a reference point in the scene. This orbital rotation simultaneously changes both the camera position and orientation, producing viewpoints that observe the same scene from slightly different angles while preserving the scene content. For each training sample, we randomly select one of the two perturbed views and project the original observation to this viewpoint via the 2D-3D-2D mapping in Eq. (1), producing a perturbed observation x^p that shares the same world state and action but differs in visual appearance. The wrist-view camera is not perturbed, as its viewpoint is determined by the robot arm's configuration rather than the global camera setup.

Cross-view correspondence visualization. Figure 3 visualizes the token-level cross-view retrieval used by our cross-view world consistency objective. For each token in the **origin** view, we back-project its representative 3D point using ground-truth depth and camera pose, then re-project it into the **perturbed** view to obtain the geometric correspondence (solid dot). We retrieve the most similar perturbed token by cosine similarity (ring at the retrieved cell center), and draw an arrow from the geometric point to the retrieved token to show the discrepancy between geometry-based and feature-based correspondences. Arrow colors indicate retrieval quality: **green** = rank-1 hit, **orange** = rank-5 hit (GT in top-5 but not top-1), **red** = miss (GT not in top-5). Tokens with invalid depth projection are omitted. The numbers below each image summarize rank-1, rank-5, pos_sim(proj) (mean cosine similarity at the geometric-GT position after the projection head; higher is better), and InfoNCE (dense contrastive loss; lower is better).



(a) w/ geometric consistency (ours) (b) w/o geometric consistency (no-gc)

Figure 3: **Cross-view feature consistency visualization.**

Per-perturbation comparison. Table 6 provides a per-dimension breakdown comparing our geometric consistency approach with camera augmentation. While camera augmentation achieves higher scores on Camera (+6.8 pp) and Light (+2.3 pp) through direct viewpoint exposure, it significantly degrades on Robot (−6.0 pp) and Background (−3.6 pp). In contrast, our method maintains stable performance across all dimensions without severe degradation on any individual perturbation type.

Table 6: **Per-dimension comparison** on LIBERO Plus (Fei et al., 2025) (% success rate). "Camera Augmentation" trains on 3× data from three viewpoints; "Geometric Consistency" uses cross-view consistency with half the data exposure on the perturbed viewpoint. **Bold**: best.

Method	Camera	Robot	Language	Light	Background	Noise	Layout	Avg.
Camera Augmentation	**74.5**	61.2	93.4	**96.3**	65.2	**85.1**	83.0	79.8
Geometric Consistency (Ours)	67.7	**67.2**	**93.6**	94.0	**68.8**	84.5	83.0	79.7

C LIBERO Future Depth Prediction Visualizations

This appendix presents qualitative results of our future depth prediction model across the four LIBERO benchmark suites (Liu et al., 2023): LIBERO-Spatial, LIBERO-Object, LIBERO-Goal, and LIBERO-10. For each task suite, we visualize the predicted depth rollouts over 8 future timesteps under different camera viewpoint perturbations (original, left, right, up, down). The visualizations demonstrate that our model generates geometrically consistent and temporally coherent depth predictions across diverse manipulation scenarios.

C.1 LIBERO-Spatial

Figures 4–8 show the future depth predictions for the LIBERO-Spatial task suite under different viewpoint perturbations.

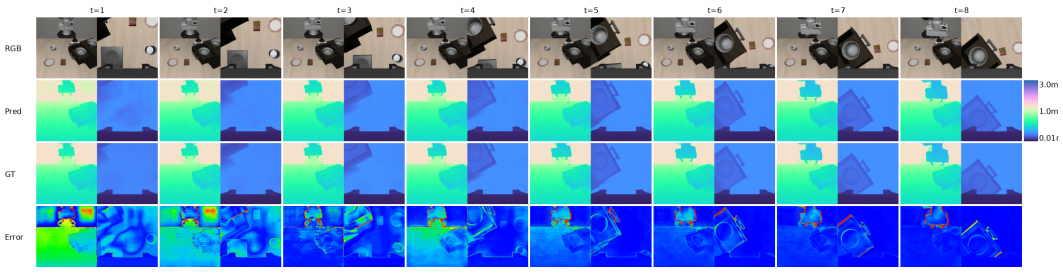


Figure 4: **LIBERO-Spatial depth rollout (original viewpoint).** Predicted future depth maps over 8 timesteps.

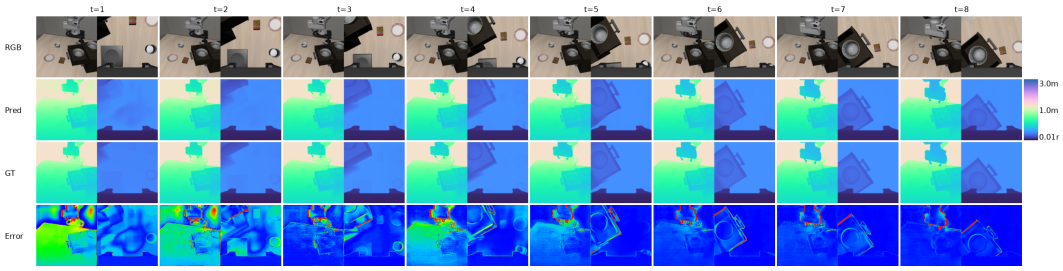


Figure 5: **LIBERO-Spatial depth rollout (left perturbation).** Predicted future depth maps over 8 timesteps.

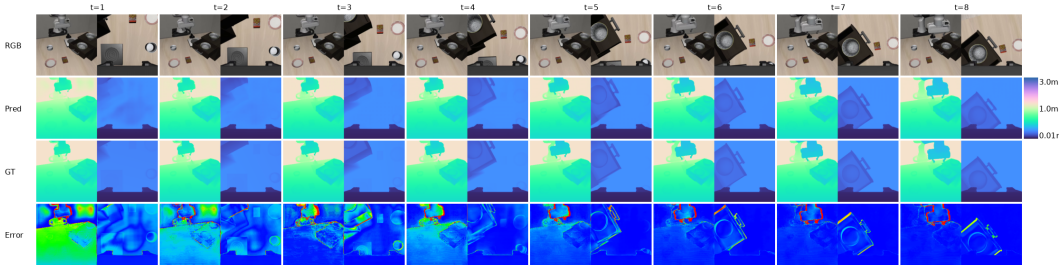


Figure 6: **LIBERO-Spatial depth rollout (right perturbation).** Predicted future depth maps over 8 timesteps.

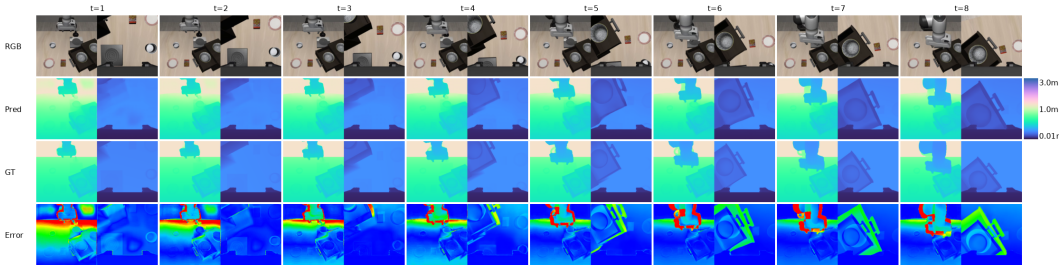


Figure 7: **LIBERO-Spatial depth rollout (up perturbation).** Predicted future depth maps over 8 timesteps.

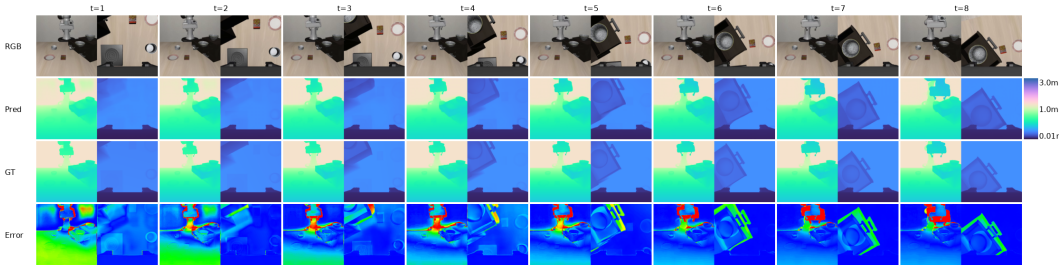


Figure 8: **LIBERO-Spatial depth rollout (down perturbation).** Predicted future depth maps over 8 timesteps.

C.2 LIBERO-OBJECT

Figures 9–13 present the depth rollout predictions for the LIBERO-Object task suite, which focuses on object-centric manipulation.

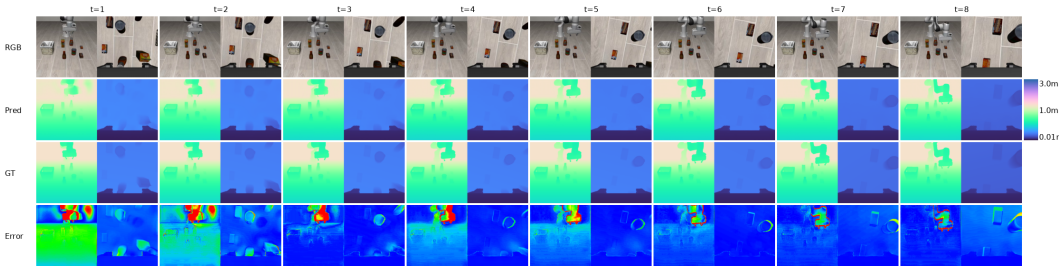


Figure 9: **LIBERO-Object depth rollout (original viewpoint).** Predicted future depth maps for object-centric manipulation.

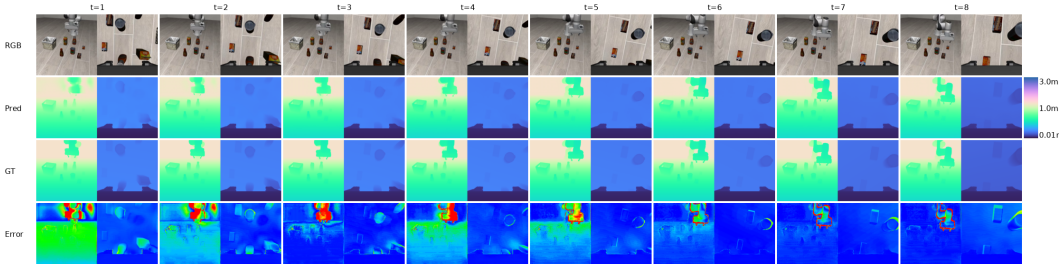


Figure 10: **LIBERO-Object depth rollout (left perturbation).** Predicted future depth maps for object-centric manipulation.

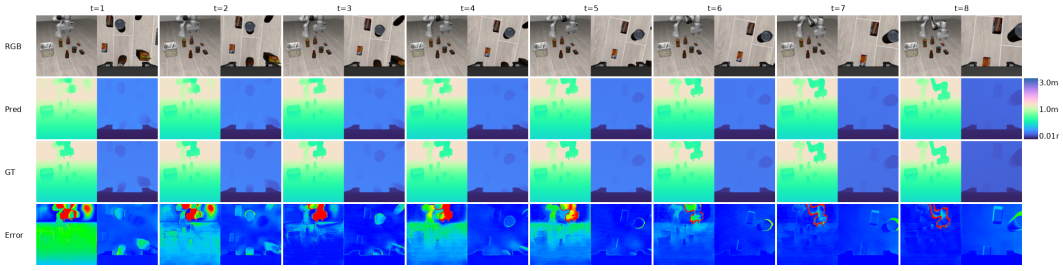


Figure 11: **LIBERO-Object depth rollout (right perturbation).** Predicted future depth maps for object-centric manipulation.

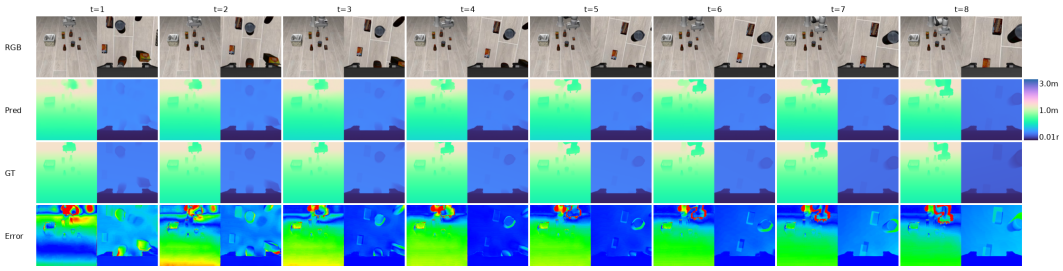


Figure 12: **LIBERO-Object depth rollout (up perturbation).** Predicted future depth maps for object-centric manipulation.

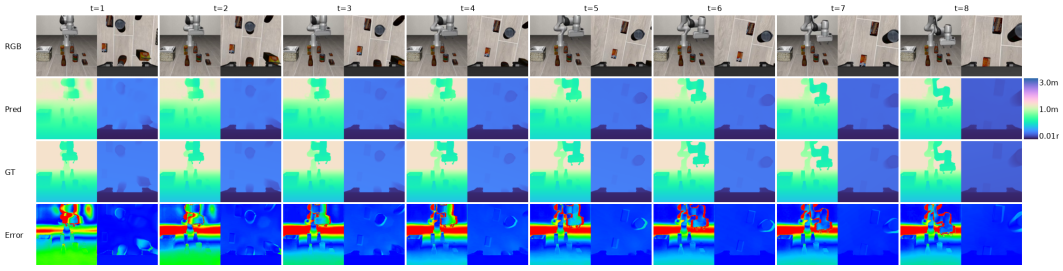


Figure 13: **LIBERO-Object depth rollout (down perturbation).** Predicted future depth maps for object-centric manipulation.

C.3 LIBERO-GOAL

Figures 14–15 show the depth rollout predictions for the LIBERO-Goal task suite.

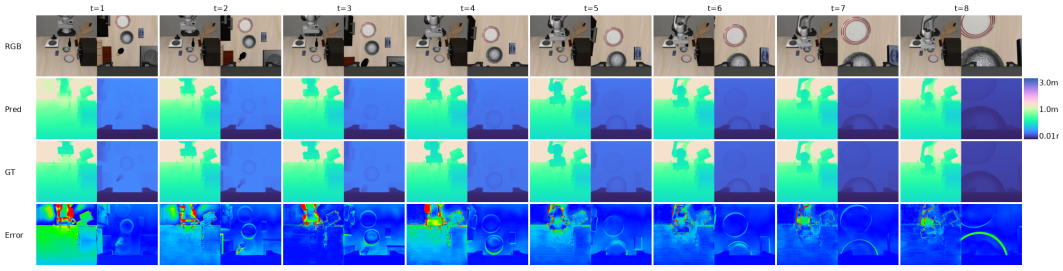


Figure 14: **LIBERO-Goal depth rollout (original viewpoint).** Predicted future depth maps for goal-conditioned manipulation.

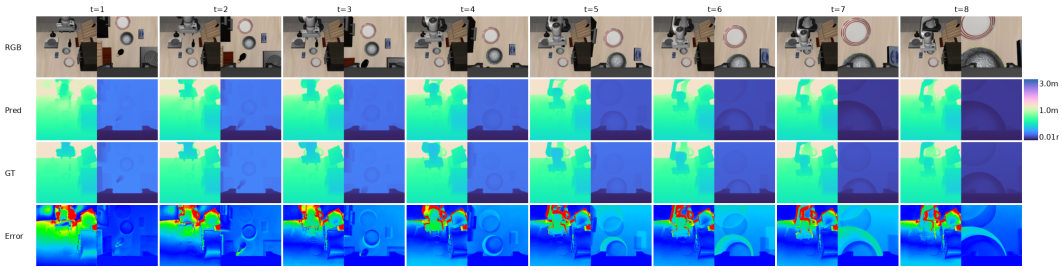


Figure 15: **LIBERO-Goal depth rollout (up perturbation).** Predicted future depth maps for goal-conditioned manipulation.

C.4 LIBERO-10

Figure 16 presents the depth rollout prediction for the LIBERO-10 benchmark.

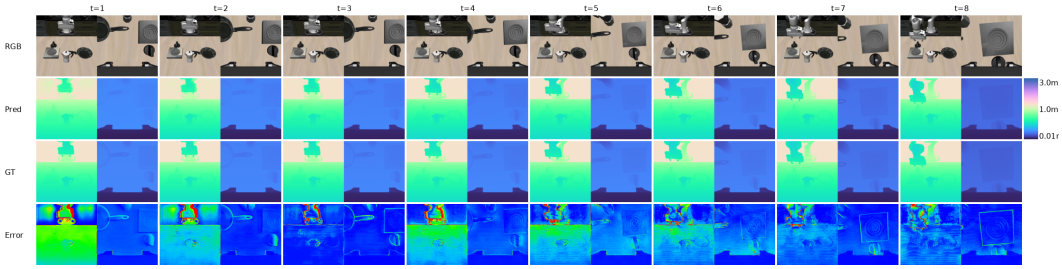


Figure 16: **LIBERO-10 depth rollout visualization.** Predicted future depth maps for a LIBERO-10 manipulation task.