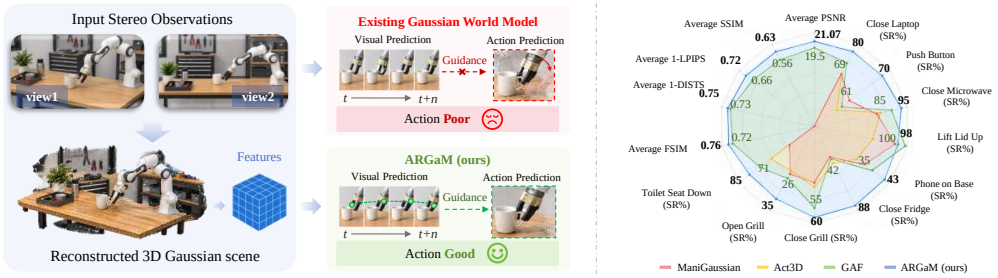
ARGaM: Action-Recoverable Gaussian Motion for Robotic Manipulation

Anonymous authors
Paper under double-blind review

Abstract

Gaussian world models predict future scene evolution in explicit 3D representations, offering geometric cues for robotic manipulation. However, visually accurate predictions alone do not ensure that the underlying motion can reliably guide action trajectories. To address this, we propose ARGaM, which learns Gaussian dynamics autoregressively with temporally aligned visual and action supervision, and refines directly regressed trajectories using a separately recovered geometric motion prior. The dynamics model predicts successive Gaussian states and a coarse action trajectory from shared temporal features, supervised by future-image reconstruction and demonstrated actions across the prediction horizon. An additional motion loss supervises selected gripper Gaussian displacements using demonstration-derived targets, bringing explicit motion constraints into the predicted geometry. Geometric registration then produces a motion prior from the predicted Gaussian correspondences. A residual diffusion model combines this prior, its recovery confidence, the coarse trajectory, and rendered future visual context to predict corrections to the coarse actions. On nine RLBench tasks, ARGaM achieves a mean closed-loop success rate of 72.7%, improving over the strongest evaluated baseline (60.4%) by 12.3 percentage points. On six real-robot tasks, ARGaM succeeds in 48 of 60 trials (80.0%), compared with 35 of 60 (58.3%) for the same baseline, an improvement of 21.7 percentage points.



(a) Action-Recoverable World Modeling **(b) Overall Performance Comparison**

Figure 1: **Motivation and performance overview.** (a) Conceptual illustration of predicted Gaussian scene motion and its use in action generation. (b) ARGaM pushes the limits on nine RLBench tasks when compared with existing methods.

1 Introduction

Gaussian world models represent future scene evolution with explicit three-dimensional primitives, providing both rendered observations and geometric correspondences for robotic manipulation (Kerbl et al., 2023; Wu et al., 2024; Lu et al., 2025; Chai et al., 2026). Their value for control depends on the motion encoded in these predictions and how that motion informs an action trajectory (Hafner et al., 2020; 2021; Hansen et al., 2024; Ko et al., 2024). This motivates a question beyond future-image prediction: how should Gaussian dynamics be trained so that their motion can guide trajectory generation?

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.

Our central idea is to train future Gaussian motion with action targets so that it can serve as a geometric prior for trajectory refinement. Image reconstruction supervises projected appearance and action regression supervises trajectory outputs, but neither provides explicit displacement targets for the Gaussian correspondences used in motion readout. We pursue *action recoverability* as a design goal and the predicted geometry should retain motion that can be read out to guide action generation (Du et al., 2023; Bharadhwaj et al., 2025; Zhi et al., 2025; Zhou et al., 2026). Demonstrations support this goal at two levels, by supervising complete action trajectories and supplying additional local displacement targets for geometric recovery.

We introduce ARGaM through three connected components. **Autoregressive Gaussian dynamics** (Sec. 3.2) learn successive scene states and a coarse action trajectory from shared temporal features under aligned visual and action supervision. **Correspondence-aware motion recovery** (Sec. 3.3) combines demonstration-derived Gaussian displacement supervision with geometric readout of a motion prior. **Residual action refinement** (Sec. 3.4) corrects the directly regressed trajectory, conditioned on this prior, its confidence, and rendered future context. Thus, action targets shape temporal prediction, and recovered geometry supplies an additional condition for correcting the learned trajectory (Fig. 1 (a)).

ARGaM achieves 72.7% mean success on nine RLBench tasks, compared with 60.4% for the strongest evaluated baseline, and performs better on eight tasks (Fig. 1 (b)). On six real-robot tasks, it succeeds in 48 of 60 trials, compared with 35. Controlled supervision ablations show lower coarse and refined position errors, while component ablations support the utility of geometric conditioning. These complementary evaluations assess how the proposed training and refinement design supports action generation.

Our contributions are:

- We introduce autoregressive Gaussian prediction with temporally aligned scene and action supervision, augmented by demonstration-derived displacement targets for the local geometry used in motion recovery.
- We develop residual refinement that combines a directly learned trajectory with a separately recovered geometric motion prior, using recovery confidence and predicted visual context to condition the correction.
- Experiments on nine RLBench tasks and six real-robot tasks show improved manipulation success. Ablations connect temporal and local motion supervision to lower action errors, and geometric conditioning to higher task success.

Together, these results support learning action-supervised Gaussian dynamics and using their recovered motion to refine trajectories.

2 Related Work

2.1 World Models for Robotic Manipulation

World models support robot control through visual prediction, model-predictive control, and latent imagination (Finn & Levine, 2017; Ebert et al., 2018; Wu et al., 2023). Generative approaches use predicted futures for inverse dynamics (Du et al., 2023), correspondence-based action recovery (Ko et al., 2024), and behavior composition (Zhou et al., 2024). Predictive representations and generated demonstrations also support manipulation (Zeng et al., 2024; Hu et al., 2025; Bharadhwaj et al., 2025; Zhen et al., 2025), while recent systems integrate action-conditioned simulation or joint observation action modeling (Zhu et al., 2025; Cen et al., 2025; Huang et al., 2026; Zhu et al., 2026; Guo et al., 2026; Liao et al., 2026). These approaches commonly represent future states through images or learned features. ARGaM focuses on explicit Gaussian dynamics, where action supervision can constrain the local geometry used for motion readout as well as the learned trajectory.

2.2 3D and 4D Gaussian World Models

3D Gaussian Splatting represents geometry and appearance with differentiable primitives (Kerbl et al., 2023). Extensions model dynamic scenes through persistent Gaussians, deformation, 4D representations, and physical dynamics (Luiten et al., 2024; Wu et al., 2024; Yang et al., 2024a;b;

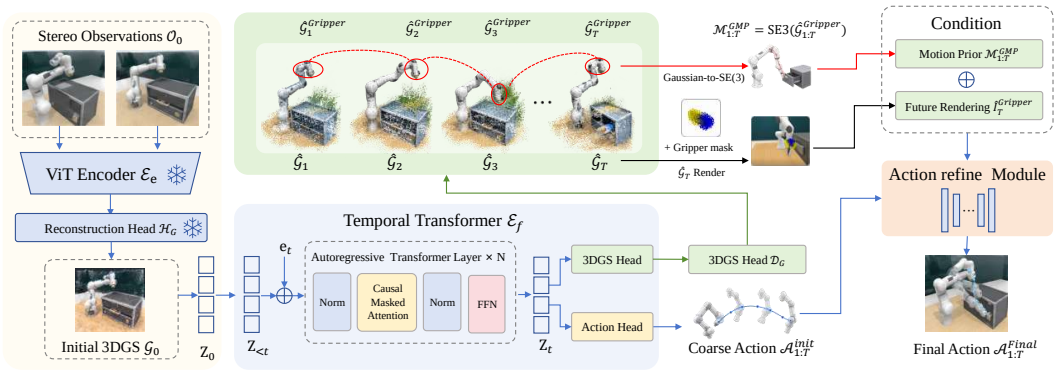


Figure 2: ARGaM autoregressively predicts future Gaussian states together with a coarse action. Temporal Gaussian correspondences are then used to recover a geometry-derived motion prior through reliability-weighted rigid registration. Finally, residual diffusion refines the coarse action conditioned on the recovered motion prior and the predicted gripper-aware future state.

Lin et al., 2024; Xie et al., 2024). Robotic applications use Gaussian futures for policy learning and planning (Lu et al., 2025; Zhang et al., 2026; Kreber et al., 2026); related methods exploit explicit 3D motion and correspondence for action generation (Zhi et al., 2025; Zhou et al., 2026). The closest method, GAF, predicts future Gaussians under visual supervision, obtains an initial action through rigid registration, and refines it with diffusion (Chai et al., 2026). ARGaM adds temporally aligned scene–action learning and local Gaussian displacement supervision. It directly regresses the initial trajectory and uses geometric readout as an additional refinement condition, separating learned initialization from recovered motion guidance.

3 METHOD

We present **ARGaM** (*Action-Recoverable Gaussian Motion*), a 4D Gaussian world model that couples future scene prediction with robot action generation. ARGaM autoregressively predicts future Gaussian states together with a coarse action, recovers a geometry-derived motion prior from persistent Gaussian correspondences, and refines the coarse action by reliability-guided residual diffusion.

3.1 PROBLEM FORMULATION AND METHOD OVERVIEW

We study *action-recoverable 4D Gaussian world modeling*, where predicted Gaussian dynamics support a learned action trajectory and provide a geometric motion prior for refinement. As illustrated in Fig. 2, given an initial stereo observation $\mathcal{O}_0 = \{I_0^v, K_0^v, T_0^v\}_{v=1,2}$, where I_0^v, K_0^v, and T_0^v denote the RGB observation, camera intrinsics, and camera extrinsics of view v, a pretrained Gaussian reconstruction model reconstructs the initial world:

$$\mathcal{G}_0 = \mathcal{H}_G(\mathcal{E}_e(\mathcal{O}_0)) = \{g_i^0\}_{i=1}^N \tag{1}$$

The reconstruction model remains frozen throughout ARGaM training. Each Gaussian g_i^0 is parameterized by centroid $\boldsymbol{\mu}_i^0 \in \mathbb{R}^3$, spherical harmonics coefficients $\mathbf{c}_i^0$, opacity $\sigma_i^0 \in [0,1]$, rotation quaternion $\mathbf{r}_i^0 \in \mathbb{R}^4$, and anisotropic scale $\mathbf{s}_i^0 \in \mathbb{R}^3$.

Starting from $\mathcal{G}_0$, ARGaM predicts future Gaussian states $\hat{\mathcal{G}}_{1:T}$ and directly regresses a coarse action trajectory $\mathcal{A}_{1:T}^{\text{init}}$ from shared temporal features, where $\mathcal{A}_t \in \mathbb{R}^{action-dim}$ denotes the robot action at time step t. Persistent correspondences among selected gripper Gaussians provide a Gaussian-derived Motion Prior (GMP), $\mathcal{M}_{1:T}^{\text{GMP}}$, through weighted rigid registration. The final predicted Gaussian state is rendered into a gripper-aware observation $\hat{I}_T^{\text{gripper}}$. A separately trained residual diffusion module conditions on these cues to produce the final trajectory $\mathcal{A}_{1:T}^{\text{final}}$.

3.2 AUTOREGRESSIVE GAUSSIAN WORLD DYNAMICS

We refer to this temporal prediction module as Autoregressive Gaussian Dynamics (AGD). ARGaM models future scene evolution directly in the explicit Gaussian space. We construct the initial Gaus-

sian tokens $\mathbf{Z}_0$ from $\mathcal{G}_0$, and employ a temporal Transformer $\mathcal{E}_f$ to autoregressively propagate Gaussian dynamics:

$$\mathbf{Z}_t = \mathcal{E}_f\left(\mathbf{Z}_{<t}, \mathbf{e}_t^{\text{time}}\right), \qquad t = 1, \ldots, T \tag{2}$$

where $\mathbf{e}_t^{\text{time}}$ encodes the future step and $\mathbf{Z}_t = \{\mathbf{z}_i^t\}_{i=1}^N$ contains the dynamic features of all Gaussian primitives at future step t. The Gaussian dynamics head and detokenizer $\mathcal{D}_G$ decode $\mathbf{Z}_{1:T}$ into the future Gaussian evolution $\hat{\mathcal{G}}_{1:T}$. In parallel, an action head projects the shared temporal representation into a coarse action $\mathcal{A}_{1:T}^{\text{init}}$.

The predicted Gaussian states are differentiably rendered into the observed views and supervised by

$$\mathcal{L}_{\text{world}} = \sum_{t \in \mathcal{S}} \alpha_t \left(\mathcal{L}_{\text{rec}}^t + \lambda_{\text{DSSIM}} \mathcal{L}_{\text{DSSIM}}^t\right) \tag{3}$$

where $\mathcal{S}$ indexes supervised future frames and α_t denotes their normalized temporal weights. Here, $\mathcal{L}_{\text{rec}}^t = \mathcal{L}_{\text{MSE}}^t + \mathcal{L}_{\text{LPIPS}}^t$ and $\mathcal{L}_{\text{DSSIM}}^t = 1 - \text{SSIM}(\hat{I}_t, I_t)$. The coarse action prediction is directly supervised by the demonstrated trajectory:

$$\mathcal{L}_{\text{init}} = \frac{1}{T} \sum_{t=1}^{T} \left\| \mathcal{A}_t^{\text{init}} - \mathcal{A}_t^{\text{GT}} \right\|_1 \tag{4}$$

3.3 Correspondence-aware Gaussian Motion Recovery

The persistent identities of Gaussian primitives provide explicit temporal correspondences from which gripper motion can be geometrically recovered. We project the initial gripper mask into the reconstructed Gaussian world and select primitives associated with the approximately rigid gripper region, denoted as $\mathcal{G}_{\text{gripper}}^0$. Since each $g_i^0 \in \mathcal{G}_{\text{gripper}}^0$ retains its identity through temporal prediction, its predicted counterpart $\hat{g}_i^t$ forms the correspondence pair $(\boldsymbol{\mu}_i^0, \hat{\boldsymbol{\mu}}_i^t)$. We restrict registration to this local rigid subset rather than the complete scene, since articulated objects and movable gripper parts generally do not satisfy a single rigid-motion assumption.

At each future step t, we recover the gripper motion through reliability-weighted rigid registration:

$$(\mathbf{R}_t, \mathbf{p}_t) = \arg\min_{\mathbf{R} \in SO(3),\, \mathbf{p}} \sum_{g_i^0 \in \mathcal{G}_{\text{gripper}}^0} w_i^t \left\| \mathbf{R}\boldsymbol{\mu}_i^0 + \mathbf{p} - \hat{\boldsymbol{\mu}}_i^t \right\|_2^2 \tag{5}$$

where w_i^t denotes the confidence of the i-th Gaussian correspondence and suppresses unreliable matches during weighted Kabsch alignment (Kabsch, 1976). The estimated $(\mathbf{R}_t, \mathbf{p}_t)$ defines the cumulative rigid transformation $\Delta\mathbf{T}_{0\to t}^{\text{GMP}} \in SE(3)$. When registration is performed in the Gaussian reference frame, the recovered transform is converted to the current gripper reference frame before being used for action refinement.

We further characterize the recovered trajectory by its cumulative motion, step-wise motion, and remaining motion to the final predicted state. In particular, the incremental motion between consecutive predictions is represented in the Lie algebra of $SE(3)$ as $\boldsymbol{\xi}_t^{\text{step}} = \text{Log}\big((\Delta\mathbf{T}_{0\to t-1}^{\text{GMP}})^{-1} \Delta\mathbf{T}_{0\to t}^{\text{GMP}}\big)$, with $\Delta\mathbf{T}_{0\to 0}^{\text{GMP}} = \mathbf{I}$. Together with the cumulative and remaining motions, we attach registration-quality descriptors measuring fitting consistency, Gaussian support, geometric confidence, and temporal cycle consistency. These quantities form a motion token $\mathbf{f}_t^{\text{GMP}}$ containing both recovered $SE(3)$ motion and its internal geometric reliability. The resulting sequence $\mathcal{M}_{1:T}^{\text{GMP}} = \{\mathbf{f}_t^{\text{GMP}}\}_{t=1}^T$ is referred to as the **Gaussian-derived Motion Prior (GMP)** and is used as a complementary geometric condition for action refinement.

To make the predicted correspondences informative for this geometric readout, we introduce **Action-Guided Gaussian Motion (AGM)** supervision in addition to future-image and coarse-trajectory supervision. Let $(\mathbf{R}_t^{\text{GT}}, \mathbf{p}_t^{\text{GT}})$ denote the demonstrated gripper transformation from the initial state to future step t, expressed in the Gaussian coordinate frame. For each selected gripper Gaussian, the corresponding target displacement is $\Delta\boldsymbol{\mu}_{i,t}^{\text{GT}} = \mathbf{R}_t^{\text{GT}}\boldsymbol{\mu}_i^0 + \mathbf{p}_t^{\text{GT}} - \boldsymbol{\mu}_i^0$, while the predicted displacement is $\Delta\boldsymbol{\mu}_{i,t}^{\text{pred}} = \hat{\boldsymbol{\mu}}_i^t - \boldsymbol{\mu}_i^0$. The corresponding AGM objective is defined as

$$\mathcal{L}_{\text{AGM}} = \frac{1}{T|\mathcal{G}_{\text{gripper}}^0|} \sum_{t=1}^{T} \sum_{g_i^0 \in \mathcal{G}_{\text{gripper}}^0} \text{SmoothL1}\left(\Delta\boldsymbol{\mu}_{i,t}^{\text{pred}}, \Delta\boldsymbol{\mu}_{i,t}^{\text{GT}}\right) \tag{6}$$

Table 1: Closed-loop manipulation success rate (%) on nine RLBench tasks. ARGaM is evaluated on 100 unseen episodes per task. Best results are shown in bold.

Method	Toi.	Open G.	Close G.	Close F.	Phone B.	Lift L.	Mic.	But.	Lap.	Avg.
ACT	46	15	33	18	25	89	59	57	61	44.8
R&D-AI	53	17	38	22	25	91	70	61	55	48.0
DP	39	20	36	16	23	81	62	61	64	44.7
Act3D	60	22	41	47	26	91	73	60	58	53.1
ManiGaussian	34	24	38	41	30	97	67	63	57	50.1
GAF	71	26	55	42	35	**100**	85	61	69	60.4
ARGaM (Ours)	**85**	**35**	**60**	**88**	**43**	98	**95**	**70**	**80**	**72.7**

AGM directly constrains the local Gaussian displacements used for geometric motion recovery, complementing image-level supervision and encouraging action-consistent temporal correspondences.

3.4 Reliability-guided Residual Diffusion Refinement

Although the Gaussian-derived Motion Prior (GMP) provides geometry-grounded motion guidance, prediction errors and imperfect correspondences may introduce uncertainty. ARGaM therefore does not directly execute GMP, but uses it as a complementary condition to refine the coarse action trajectory $\mathcal{A}_{1:T}^{\text{init}}$ through residual diffusion.

Specifically, the recovered motion prior $\mathcal{M}_{1:T}^{\text{GMP}}$ provides geometric guidance, while the final predicted Gaussian state $\hat{\mathcal{G}}_T$ is rendered with the gripper mask to obtain a gripper-aware future observation $\hat{I}_T^{\text{gripper}}$ as visual target guidance. The contribution of GMP is modulated by its recovery reliability before being fused with $\mathcal{A}_{1:T}^{\text{init}}$ and $\hat{I}_T^{\text{gripper}}$. The diffusion model predicts a residual trajectory $\Delta\mathcal{A}_{1:T}$, yielding

$$\mathcal{A}_{1:T}^{\text{final}} = \mathcal{A}_{1:T}^{\text{init}} + \Delta\mathcal{A}_{1:T}. \tag{7}$$

Training Objectives. The initial Gaussian reconstruction model is pretrained and kept frozen throughout training. ARGaM is optimized in two stages. In the first stage, we train the autoregressive Gaussian dynamics and coarse action prediction with

$$\mathcal{L}_{\text{stage1}} = \mathcal{L}_{\text{world}} + \lambda_{\text{init}}\mathcal{L}_{\text{init}} + \lambda_{\text{AGM}}\mathcal{L}_{\text{AGM}}. \tag{8}$$

Here, $\mathcal{L}_{\text{world}}$ supervises future Gaussian prediction, $\mathcal{L}_{\text{init}}$ supervises the coarse action trajectory, and $\mathcal{L}_{\text{AGM}}$ encourages action-consistent Gaussian motion. The GMP is geometrically recovered from the predicted Gaussian evolution through correspondence-aware Kabsch alignment and does not require an additional training objective.

In the second stage, only the residual diffusion refiner is optimized. We define the target residual as $\Delta\mathcal{A}_{1:T}^{\text{GT}} = \mathcal{A}_{1:T}^{\text{GT}} - \mathcal{A}_{1:T}^{\text{init}}$ and the conditioning set as $\mathcal{C} = \{\mathcal{A}_{1:T}^{\text{init}}, \mathcal{M}_{1:T}^{\text{GMP}}, \hat{I}_T^{\text{gripper}}\}$. Following the standard diffusion denoising objective (Ho et al., 2020), we optimize

$$\mathcal{L}_{\text{diff}} = \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta \left(\Delta\mathcal{A}_{1:T}^{(k)}, k; \mathcal{C} \right) \right\|_2^2 \tag{9}$$

Here, k is the diffusion step, $\Delta\mathcal{A}_{1:T}^{(k)}$ is the noised target residual, and $\boldsymbol{\epsilon}$ is the added Gaussian noise.

Together, these components connect future Gaussian prediction, geometric motion recovery, and residual action refinement within a unified action-recoverable world modeling framework.

4 Experiments

We evaluate ARGaM across simulation and real-world manipulation from four complementary perspectives: closed-loop task success, future prediction quality, action accuracy, and long-horizon robustness. We further conduct controlled ablations to analyze the contributions of autoregressive temporal modeling, action-guided Gaussian motion supervision, and geometric motion conditioning to the overall manipulation performance.

Table 2: **Future rendering quality on representative RLBench tasks.** Metrics are computed between rendered Gaussian predictions and ground-truth future observations.

	Close Microwave						Close Grill				
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FSIM↑		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FSIM↑
GAF	18.56	0.48	0.44	0.26	0.71	GAF	16.41	0.39	0.48	0.37	0.64
ARGaM	**21.37**	**0.59**	**0.31**	**0.23**	**0.73**	**ARGaM**	**19.36**	**0.57**	**0.29**	**0.30**	**0.69**
	Close Fridge						Open Grill				
	PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FSIM↑		PSNR↑	SSIM↑	LPIPS↓	DISTS↓	FSIM↑
GAF	17.53	0.50	0.44	0.34	0.75	GAF	17.59	0.41	0.44	0.30	0.65
ARGaM	**20.42**	**0.65**	**0.32**	**0.27**	**0.80**	**ARGaM**	**23.14**	**0.71**	**0.18**	**0.20**	**0.83**

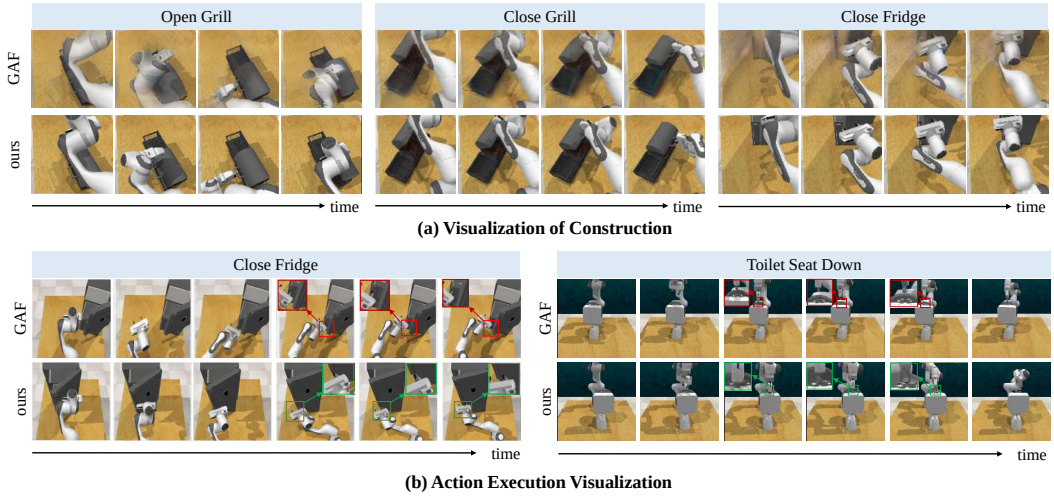


Figure 3: **Prediction and execution examples in simulation.** (a) Rendered Gaussian scene predictions over successive frames on *Open Grill*, *Close Grill*, and *Close Fridge*. (b) Execution sequences on *Close Fridge* and *Toilet Seat Down*, with highlighted regions showing gripper–object interaction.

4.1 Experimental Setup

Benchmarks and observations. We evaluate ARGaM on nine RLBench tasks (James et al., 2020), with 20 training demonstrations and 100 evaluation episodes per task using unseen object configurations and poses. Inputs comprise two external RGB views, the calibrated external pair initializes the Gaussian scene. The real-robot protocol is described in Sec. 4.3.

Comparisons and metrics. We compare with Action Chunking with Transformers (ACT) (Zhao et al., 2023), Diffusion Policy (DP) (Chi et al., 2023), Render and Diffuse (R&D-AI) (Vosylius et al., 2024), Act3D (Gervet et al., 2023), ManiGaussian (Lu et al., 2024), and Gaussian Action Field (GAF) (Chai et al., 2026). We report per-task success and its unweighted mean (Avg.), with differences in percentage points. Rendered futures are evaluated using SSIM (Wang et al., 2004), LPIPS (Zhang et al., 2018), DISTS (Ding et al., 2022) and FSIM (Zhang et al., 2011); action accuracy is measured by end-effector position error before and after refinement. The task list, training hyperparameters, image resolutions, and execution details are provided in Appendix.

4.2 Main Results in Simulation

Closed-loop manipulation. Table 1 reports that ARGaM achieves the best overall performance among all baseline methods, with an average success rate of **80.0%**, outperforming ACT (41.7%), ManiGaussian (35.0%), Act3D (50.0%), and GAF (58.3%) by **38.3**, **45.0**, **30.0**, and **21.7** percentage points (pp). ARGaM achieves higher success on eight of nine tasks, with the largest improvement on *Close Fridge* (42% to 88%). On *Lift Lid Up*, both methods are near the ceiling (98% and 100%). The

Table 3: **Real-world manipulation success rates (%).** Results on six physical manipulation tasks.

Method	Lift Bowl	Move Apple	Push Button	Move Block	Close Door	Pick Cup	Average
ACT	20	30	50	50	40	60	41.7
ManiGaussian	0	20	60	30	50	50	35.0
Act3D	40	40	60	50	50	60	50.0
GAF	50	60	70	60	40	70	58.3
ARGaM (Ours)	**70**	**90**	**90**	**60**	**70**	**100**	**80.0**

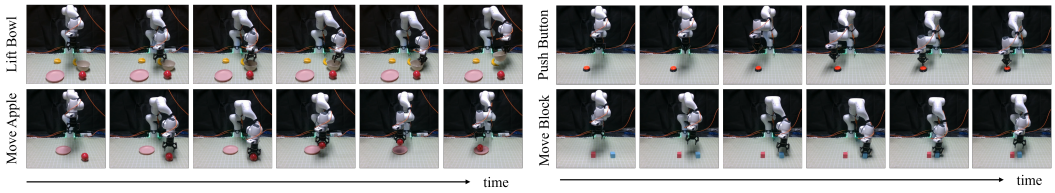


Figure 4: **Real-world manipulation.** Representative execution sequences across the six physical manipulation tasks.

gains across articulated-object interaction, relocation, and button pressing support the full policy's effectiveness across the tested task types.

Future prediction quality. Table 2 evaluates peak signal-to-noise ratio (PSNR), structural similarity (SSIM), learned perceptual image patch similarity (LPIPS), deep image structure and texture similarity (DISTS), and feature similarity (FSIM). ARGaM improves all five metrics over GAF on each of the four tasks. On *Close Microwave*, PSNR increases from 18.56 to 21.37 dB and SSIM from 0.48 to 0.59, while LPIPS decreases from 0.44 to 0.31. PSNR gains range from 2.81 to 5.55 dB. Agreement across pixel, structural, and perceptual metrics provides complementary evidence for the quality of the predicted visual context used during refinement.

Qualitative comparison. Figure 3 (a) shows better preservation of robot and object structure across predicted frames. The execution sequences in Fig. 3 (b) illustrate gripper–object interaction on *Close Fridge* and *Toilet Seat Down*, complementing the task-success comparisons.

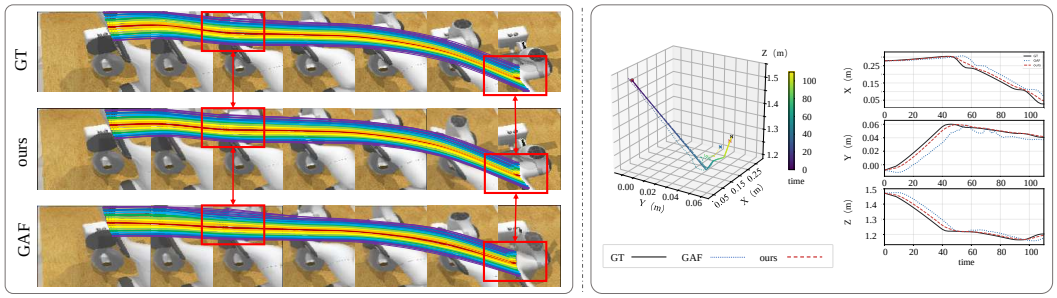
4.3 Real-World Manipulation

Table 3 compares ARGaM with ACT, ManiGaussian, Act3D, and GAF on six real-robot tasks spanning object relocation, button pressing, and articulated-object interaction. Each task includes ten trials with randomized object poses and initial robot configurations. We report per-task success rates and their mean across the six tasks.

ARGaM achieves the highest mean success rate among all evaluated methods, succeeding in 48 of 60 trials (80.0%), compared with 58.3% for the strongest baseline. It leads on five tasks and ties for the best performance on *Move Block*, showing improvements across multiple interaction types. Figure 4 illustrates representative physical executions. Together, these results demonstrate the effectiveness of ARGaM in real-world manipulation, extending its performance advantages from simulation to the evaluated physical tasks.

4.4 Motion and Horizon Analysis

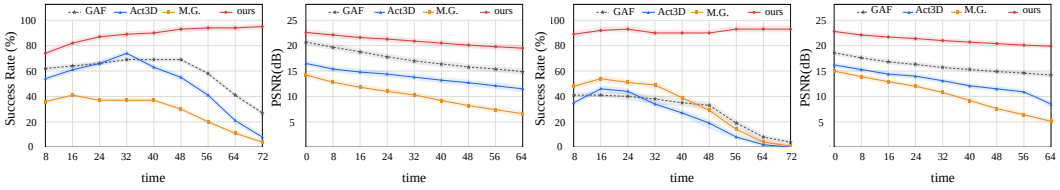
Qualitative motion and action alignment. Figure 5 compares predicted gripper motion and final position trajectories with ground truth (GT). In panel (a), the rows are GT, GAF, and ARGaM; ARGaM's optical-flow tracks follow the demonstrated motion more closely in the highlighted regions. In panel (b), its final trajectory also follows the GT path more closely in the 3D and per-axis plots. Agreement with GT in both views illustrates the intended outcome of supervising future motion and using geometric guidance for action refinement. Tables 6 and 5 quantify the associated action-accuracy and conditioning benefits.



(a) 4D Optical Flow Comparison

(b) End-Effector Trajectory Prediction Comparison

Figure 5: **Qualitative alignment of predicted gripper motion and final actions.** (a) Optical-flow visualizations for the same sequence, ordered from top to bottom as ground truth, GAF, and ARGaM. Red boxes highlight the gripper regions used to compare projected motion. (b) Final end-effector position trajectories from GAF and ARGaM compared with ground truth in a 3D view and per-axis temporal plots.



(a) Toilet Seat Down **(b) Toilet Seat Down** **(c) Close Fridge** **(d) Close Fridge**

Figure 6: **Long-horizon analysis.** Manipulation performance and future prediction quality as the future prediction horizon increases on Toilet Seat Down and Close Fridge. (a,c) Manipulation success rate and (b,d) rendering PSNR. The horizontal axis denotes future prediction length; M.G. denotes ManiGaussian.

Long-horizon behavior. Figure 6 compares performance as the future prediction length increases on Toilet Seat Down and Close Fridge. ARGaM maintains higher success across the tested horizons, with a pronounced advantage at longer prediction lengths. Its rendering PSNR also remains higher as predictions extend farther into the future. These evaluations extend the standard eight-step setting to longer requested futures, probing the complete prediction-and-action pipeline as the prediction length grows. The concurrent advantages in rendering quality and task success complement the component ablations over the evaluated horizon range.

4.5 Ablation Studies

To analyze the contributions of different supervision objectives and action-generation components, we conduct controlled ablations and evaluate their effects on task success and action accuracy before and after refinement.

Effect of Gaussian motion supervision. Table 4 evaluates Action-Guided Gaussian Motion (AGM) and image-structural supervision (D-SSIM) on *Toilet Seat Down*. AGM alone increases success from 76% to 81%, while adding it to D-SSIM raises success from 79% to 85%. In the latter comparison, PSNR rises from 21.16 to 23.00 dB, while LPIPS and DISTS decrease from 0.25 to 0.20 and 0.25 to 0.17, respectively, indicating improved reconstruction and perceptual agreement. The combined objective achieves the best results across all five rendering metrics and task success. These results show that demonstration-derived displacement targets complement rendered-appearance supervision, improving Gaussian motion for geometric readout, future prediction, and manipulation performance on this task.

Geometric guidance for action refinement. Table 5 progressively adds Autoregressive Gaussian Dynamics (AGD), the directly regressed coarse-action branch (Init-A), and the Gaussian-derived

Table 4: **World-model supervision on Toilet Seat Down.** Local Gaussian motion supervision (AGM) and image-structural supervision (D-SSIM), individually and together. Success is measured on this single task.

Method	AGM	D-SSIM	PSNR ↑	SSIM ↑	LPIPS ↓	DISTS ↓	FSIM ↑	Success (%) ↑
Baseline	×	×	20.14	0.57	0.24	0.32	0.63	76
+ D-SSIM supervision	×	✓	21.16	0.60	0.25	0.25	0.70	79
+ AGM supervision	✓	×	22.00	0.56	0.22	0.27	0.69	81
Full model	✓	✓	**23.00**	**0.67**	**0.20**	**0.17**	**0.73**	85

Table 5: **Ablation of action generation pipeline.** Each component is progressively added to evaluate its contribution to closed-loop manipulation performance. Results are reported as success rates (%) on RLBench tasks.

Method	Toi.	Open G.	Close G.	Close F.	Phone B.	Lift L.	Mic.	But.	Lap.	Avg.
Base	65	26	15	36	35	87	85	34	43	47.3
+AGD	75	29	54	60	37	75	87	55	58	58.9
+Init-A	79	32	57	68	39	74	87	62	61	62.1
+GMP	**85**	**35**	**60**	**88**	43	**98**	**95**	70	80	72.7

Motion Prior (GMP). AGD increases mean success from 47.3% to 58.9%, while Init-A further improves it to 62.1%. Adding GMP as a refinement condition raises success to 72.7%, with gains on all nine tasks, including 68% to 88% on *Close Fridge*. This improvement supports the complementary roles of learned action initialization and geometric motion conditioning: the predicted trajectory provides a starting point, while recovered Gaussian motion supplies additional guidance for refinement.

Gaussian dynamics and action accuracy. Table 6 evaluates the effect of Gaussian dynamics learning on action accuracy, with the coarse-action head, GMP conditioning, and residual refiner held fixed. Introducing Autoregressive Gaussian Dynamics (AGD) with finer temporal supervision reduces coarse and refined position errors from 3.113 to 2.794 mm and from 2.272 to 2.053 mm, respectively. Adding AGM further reduces these errors to 2.680 and 1.832 mm, while success increases from 58.9% to 72.7%. These results show that supervising predicted Gaussian motion improves both coarse and refined action accuracy within the same action-generation architecture. These results show that supervising predicted Gaussian motion improves both coarse and refined action accuracy within the same action-generation architecture. The consistent improvement after residual refinement further indicates that more accurate Gaussian dynamics provide a better geometric prior for downstream action correction.

Table 6: **Action accuracy and task success.**

Method	Coarse (mm)↓	Refined (mm)↓	Success (%)↑
Baseline	3.113	2.272	47.3
+AGD	2.794	2.053	58.9
+AGM	**2.680**	**1.832**	**72.7**

5 Conclusion

We presented ARGaM, which trains autoregressive Gaussian dynamics with visual, trajectory, and local motion supervision, then refines learned actions using a recovered geometric prior. Simulation and real-robot results show improved task success. Supervision ablations support lower action-prediction errors, and component ablations support the utility of geometric conditioning. These results demonstrate the value of the proposed training and refinement design on manipulation tasks.

Limitations. ARGaM relies on persistent Gaussian identities and calibrated multi-view geometry. Motion recovery assumes approximate rigidity of the selected gripper region; occlusion and inaccurate correspondences can weaken the geometric prior. Evaluation covers the reported manipulation tasks and prediction horizons. Broader task coverage and substantially longer open-loop execution remain directions for future work.

REFERENCES

Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2Act: Human video generation in novel scenarios enables generalizable robot manipulation. In *Proceedings of the 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 3936–3951, 2025.

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

Ying Chai, Litao Deng, Ruizhi Shao, Jiajun Zhang, Kangchen Lv, Liangjun Xing, Xiang Li, Hongwen Zhang, and Yebin Liu. GAF: Gaussian action field as a 4D representation for dynamic world modeling in robotic manipulation. In *IEEE International Conference on Robotics and Automation*, 2026.

Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin C. M. Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Proceedings of Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.026.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2786–2793, 2017. doi: 10.1109/ICRA.2017.7989324.

Theophile Gervet, Zhou Xian, Nikolaos Gkanatsios, and Katerina Fragkiadaki. Act3d: 3d feature field transformers for multi-task robotic manipulation. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 3949–3965. PMLR, 2023. URL `https://proceedings.mlr.press/v229/gervet23a.html`.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-World: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, 2026.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Danijar Hafner, Timothy P. Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. In *International Conference on Learning Representations*, 2021.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 24328–24346, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2World: Crafting video diffusion models to interactive world models. In *International Conference on Learning Representations*, 2026.

Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. RLBench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 5(2):3019–3026, 2020.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, 1976. doi: 10.1107/S0567739476001873.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023. doi: 10.1145/3592433.

Po-Chen Ko, Jiayuan Mao, Yilun Du, Shao-Hua Sun, and Joshua B. Tenenbaum. Learning to act from actionless videos through dense correspondences. In *International Conference on Learning Representations*, 2024.

Jens U. Kreber, Lukas Mack, and Joerg Stueckler. Learning action-conditional and object-centric gaussian splatting world models for rigid objects. *arXiv preprint arXiv:2606.01950*, 2026.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Jingbin Cai, Si Liu, Jianlan Luo, Liliang Chen, Shuicheng Yan, Maoqing Yao, and Guanghui Ren. Genie Envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, 2026.

Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-Flow: 4D reconstruction with dynamic 3D gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21136–21145, 2024.

Guanxing Lu, Shiyi Zhang, Ziwei Wang, Changliu Liu, Jiwen Lu, and Yansong Tang. ManiGaussian: Dynamic gaussian splatting for multi-task robotic manipulation. In *European Conference on Computer Vision*, volume 15093 of *Lecture Notes in Computer Science*, pp. 349–366. Springer, 2024. doi: 10.1007/978-3-031-72761-0_20.

Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. GWM: Towards scalable gaussian world models for robotic manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9263–9274, 2025.

Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3D gaussians: Tracking by persistent dynamic view synthesis. In *International Conference on 3D Vision (3DV)*, pp. 800–809, 2024. doi: 10.1109/3DV62453.2024.00044.

Vitalis Vosylius, Younggyo Seo, Jafar Uruç, and Stephen James. Render and diffuse: Aligning image and action spaces for diffusion-based behaviour cloning. In *Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.051. URL `https://www.roboticsproceedings.org/rss20/p051.html`.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4D gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20310–20320, 2024.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. DayDreamer: World models for physical robot learning. In *Proceedings of the 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pp. 2226–2240. PMLR, 2023.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. PhysGaussian: Physics-integrated 3D gaussians for generative dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4389–4398, 2024.

Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4D gaussian splatting. In *International Conference on Learning Representations*, 2024a.

Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20331–20341, 2024b.

Jia Zeng, Qingwen Bu, Bangjun Wang, Wenke Xia, Li Chen, Hao Dong, Haoming Song, Dong Wang, Di Hu, Ping Luo, Heming Cui, Bin Zhao, Xuelong Li, Yu Qiao, and Hongyang Li. Learning manipulation by predicting interaction. In *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011. doi: 10.1109/TIP.2011.2109730.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Zijian Zhang, Yuqing Jiang, Qian Cheng, Xiaofan Li, Si Liu, Ding Zhao, Ping Luo, Weitao Zhou, and Haibao Yu. GaussianDream: A feed-forward 3D gaussian world model for robotic manipulation. *arXiv preprint arXiv:2605.20752*, 2026.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems*, 2023. doi: 10.15607/RSS.2023.XIX.016. URL `https://www.roboticsproceedings.org/rss19/p016.html`.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Learning 4D embodied world models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5337–5347, 2025.

Hongyan Zhi, Peihao Chen, Siyuan Zhou, Yubo Dong, Quanxi Wu, Lei Han, and Mingkui Tan. 3DFlowAction: Learning cross-embodiment manipulation from 3D flow world model. *arXiv preprint arXiv:2506.06199*, 2025.

Kaichen Zhou, Yuzhen Chen, Fangneng Zhan, Hang Hua, Grace Chen, Xinhai Chang, Ao Qu, Yilun Du, Zhuang Liu, Paul Pu Liang, and Mengyu Wang. GEM-4D: Geometry-enhanced video world models for robot manipulation. *arXiv preprint arXiv:2605.22882*, 2026.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. RoboDreamer: Learning compositional world models for robot imagination. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61885–61896, 2024.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844, 2025. doi: 10.1109/ICCV51701.2025.00917.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, 2026.