WORLDFLOW: JOINT VIDEO AND 3D MOTION PREDICTION VIA LARGE-SCALE PRETRAINING

Anonymous authors
Paper under double-blind review

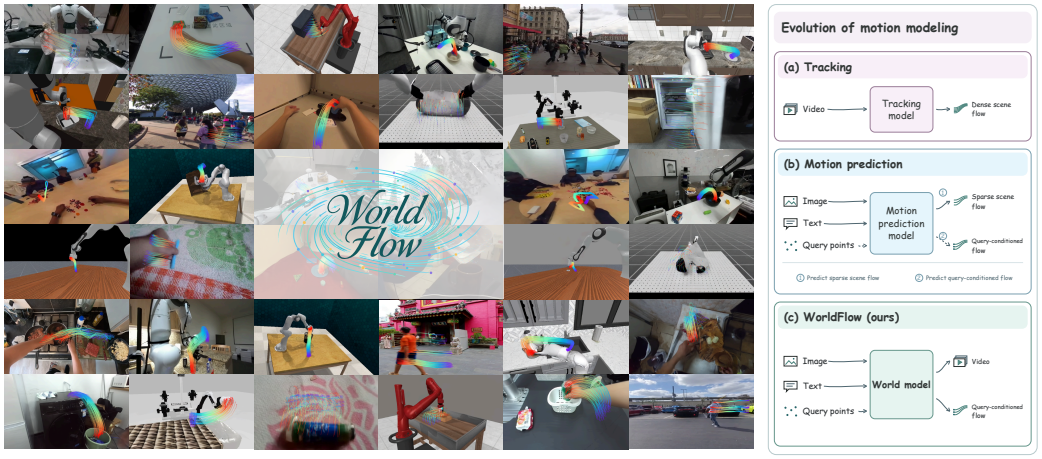

Figure 1: **3D Flow Predictions and motion modeling paradigms. Left:** 3D motion flow predicted by WorldFlow across simulated, robotic manipulation, human-object interaction, and in-the-wild environments. **Right:** (a) Tracking models recover dense scene flow from video. (b) Motion prediction models forecast future 3D flow from an image and text. Existing formulations can be broadly divided into sparse scene flow prediction and query-conditioned flow prediction. (c) WORLDFLOW is a world model conditioned on an image, text, and query points that jointly generates future video and query-conditioned 3D flow.

ABSTRACT

Predicting future physical states stands as a central challenge in embodied intelligence. Existing video world models leave 3D dynamics implicit, action-based formulations remain tied to specific robot configurations, and 3D motion prediction models use flow-only supervision, providing a limited learning signal for physical dynamics. To circumvent these limitations, we introduce WORLDFLOW, a generative world model that jointly predicts future video and query-based 3D trajectories. By parameterizing 3D motion as continuous displacement tokens aligned with the video grid, a unified Mixture-of-Transformers (MoT) backbone jointly denoises both modalities under a rectified-flow objective. This formulation couples explicit geometric dynamics with visual evolution, effectively unlocking the physical commonsense embedded within pretrained video models. To empirically scale and validate this unified architecture, we train WORLDFLOW on a heterogeneous mixture of 1.44 million object-level samples spanning simulated and real-world domains. We further introduce WorldFlowBench, a benchmark comprising three evaluation sets, including SpatialCog for assessing spatial reasoning. Across all three evaluation sets, WorldFlow achieves the lowest ADE and FDE among the evaluated learning-based methods. Crucially, on SpatialCog, our joint model reduces Average Displacement Error (ADE) by 55.9% and Final Displacement Error (FDE) by 68.5% relative to a flow-only ablation, demonstrating that generative visual priors fundamentally regularize complex geometric forecasting. Our code, pretrained models, and benchmarks will be made publicly available.

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.

1 INTRODUCTION

The philosophy of Heraclitus is often distilled into the aphorism *panta rhei*—everything flows. In the physical world, appearances and spatial configurations continually change, while objects maintain continuity over time. An embodied world model should capture both aspects. Recent generative video models (Team Wan et al., 2025; Team Seedance et al., 2026) predict increasingly coherent visual futures, but express these predictions in pixels, leaving the underlying 3D dynamics implicit. World action models (Zhou et al., 2026; Yuan et al., 2026; NVIDIA, 2026; Ye et al., 2026) connect future prediction to robotic behavior by incorporating actions into world modeling. However, action supervision is costly to collect at scale, and action representations are often tied to particular robot embodiments, complicating cross-platform transfer. These limitations motivate extending world modeling into the space of explicit 3D motion flow: a geometric representation of physical change that can connect visual prediction to planning and control across embodiments.

A complementary line of research approaches this space through motion modeling, progressing from tracking motion in observed videos (Lu et al., 2026; Luo et al., 2026) to predicting future trajectories from current observations (Zhang et al., 2026; Lee et al., 2026b). For generative motion prediction, producing dense 3D scene flow is computationally expensive. Existing scene-level methods instead predict motion on coarse spatial grids (Lee et al., 2026a; Li et al., 2026a). However, sparse scene flow still allocates motion tokens to task-irrelevant regions, while its reduced spatial resolution limits the fidelity of fine motion boundaries. Query-conditioned prediction offers a more targeted alternative, allocating motion tokens to specified 3D locations in task-critical regions while retaining global scene context as conditioning information. Yet learning generalizable dynamics at scale faces two challenges: high-quality 3D trajectory annotations are costly to acquire, while flow-only supervision provides a limited learning signal for physical dynamics, potentially limiting generalization across scenes and interactions. Jointly modeling visual evolution and explicit motion offers a natural way to enrich this supervision, bringing motion prediction into a generative world model.

To this end, we introduce WORLDFLOW, a generative world model that jointly predicts future video and query-conditioned 3D object trajectories through a unified Mixture-of-Transformers (MoT) backbone. Given an observed image, a language instruction, and sparse 3D query points, WORLDFLOW forecasts motion at the queried locations while using the full scene as visual context. We represent trajectories as continuous tokens encoding normalized frame-to-frame 3D displacements, enriched with initial point geometry and spatiotemporally aligned with video tokens. A unified rectified-flow objective jointly denoises both modalities, coupling trajectory learning with dense visual supervision of the same future.

To support training at scale, we assemble a heterogeneous corpus spanning simulated and real-world domains, comprising approximately 1.44 million object-level samples—equivalent to 1,224 hours of standardized 49-frame sequences at 16 FPS. We combine existing motion annotations with automated trajectory extraction, unify their representations, and apply source-specific quality control and fixed domain-level normalization for mixed-data training. We further introduce WorldFlowBench, comprising three evaluation sets, including SpatialCog for spatial reasoning in future motion prediction. On WorldFlowBench, WORLDFLOW achieves the lowest ADE and FDE across all three evaluation sets among non-static methods. Compared with its flow-only variant, joint video-flow generation reduces ADE by 48.5% on Sim7 and 55.9% on SpatialCog, with a 68.5% reduction in FDE on SpatialCog. These results support the value of complementary visual supervision for accurate 3D motion forecasting.

Our contributions are:

- We introduce WORLDFLOW, a generative world model that jointly predicts video and query-conditioned 3D object motion through a unified MoT backbone, combining explicit trajectory supervision with dense visual supervision.
- We construct a heterogeneous corpus of approximately 1.44 million object-level samples from simulated and real-world data, using unified motion representations and source-specific quality control to support large-scale training.
- We introduce WorldFlowBench, comprising three evaluation sets including SpatialCog for spatial reasoning, and demonstrate improvements over existing motion prediction methods and video-generation-then-tracking pipelines.

2 Related Work

2.1 Predictive World Models: From Pixels to Actions

A fundamental objective of visual world modeling is to predict future physical states from current observations. In the visual domain, large-scale video generators establish rich appearance and motion priors to enable coherent future synthesis (Team Wan et al., 2025; Team Seedance et al., 2026), driven by architectural and training efficiencies (Huang et al., 2025; Chen et al., 2025b), and extending toward physical-world prediction, interactive simulation, and semantic control (NVIDIA et al., 2025; Robbyant Team et al., 2026; Chen et al., 2026; Bian et al., 2025). To bridge visual futures with robotic behavior, World Action Models (WAMs) directly incorporate the robot's action space into predictive modeling, jointly forecasting future observations and actions to support end-to-end policy learning and interactive closed-loop control (Ye et al., 2026; Li et al., 2026c), while expanding toward in-context task generalization (Zhou et al., 2026), fast inference (Yuan et al., 2026), and cross-embodiment transfer (Li et al., 2026a). Paralleling these efforts, unified multimodal models integrate visual understanding with multi-modal generation (Deng et al., 2025), paving the way for comprehensive frameworks that jointly process language, pixels, and actions for physical-world modeling (NVIDIA, 2026).

2.2 3D Motion Modeling: From Tracking to Forecasting

Point tracking stands as a classic paradigm focused on resolving dense, pixel-level correspondences and spatial linkages over observable video sequences (Lu et al., 2026; Karhade et al., 2026; Karaev et al., 2024; 2025; Xiao et al., 2025; Zhang et al., 2025; Luo et al., 2026). Although these trackers provide direct measurements of surface history, they operate strictly on immediate visual evidence and remain susceptible to severe error accumulation under complex occlusions or non-rigid deformations. Concurrently, predictive motion forecasting has recently emerged to offer fresh research motivations tailored for downstream embodied and robotic applications. As a generative geometric interface, 3D motion flow bypasses immediate frame matching to predict future physical dynamics directly across objects and embodiments. Early work, General Flow (Yuan et al., 2024), learned language-conditioned 3D point trajectories from human RGB-D videos as transferable manipulation affordances. NovaFlow (Li et al., 2025) and Dream2Flow (Dharmarajan et al., 2025) leveraged video generation to obtain motion priors, recovering actionable 3D object flow through tracking and geometric reconstruction. Alongside these video-based pipelines, direct motion models predict future geometry without synthesizing videos. TraceGen (Lee et al., 2026a) models compact 3D traces, while PointWorld (Huang et al., 2026) predicts action-conditioned full-scene flow and ObjectForesight (Soraki et al., 2026) forecasts rigid-object pose trajectories. More recently, RoboFlow4D (Lin et al., 2026), μ_0 (Lee et al., 2026b), and MolmoMotion (Zhang et al., 2026) advanced flow-guided planning, interaction-trace generation, and language-conditioned object-point forecasting, respectively. Flow representations have also become integral to policy learning and visual world modeling: LaMP (Wang et al., 2026) introduces a latent 3D motion prior for action generation, EgoWAM (Li et al., 2026a) studies 3D flow as a world-prediction target for human-to-robot transfer, and Hydra-0 (Li et al., 2026b) uses image-plane action flow to condition video prediction and support robot control.

3 WorldFlow

3.1 Problem Formulation

Given a language instruction c, an observed RGB frame I_0, and 3D query points $\mathcal{Q} = \{\mathbf{p}_{0,i}\}_{i=1}^{N}$ with corresponding image coordinates $\mathcal{S}_0 = \{\mathbf{s}_{0,i}\}_{i=1}^{N}$, WorldFlow jointly predicts future RGB frames $I_{1:T}$ and 3D flow $f_{1:T}$. All 3D positions are expressed in the initial camera coordinate system. We define flow as the frame-to-frame displacement of each query point:

$$f_{t,i} = \mathbf{p}_{t,i} - \mathbf{p}_{t-1,i}, \qquad t = 1, \ldots, T. \tag{1}$$

The corresponding trajectories are recovered by accumulation:

$$\mathbf{p}_{t,i} = \mathbf{p}_{0,i} + \sum_{\tau=1}^{t} f_{\tau,i}. \tag{2}$$

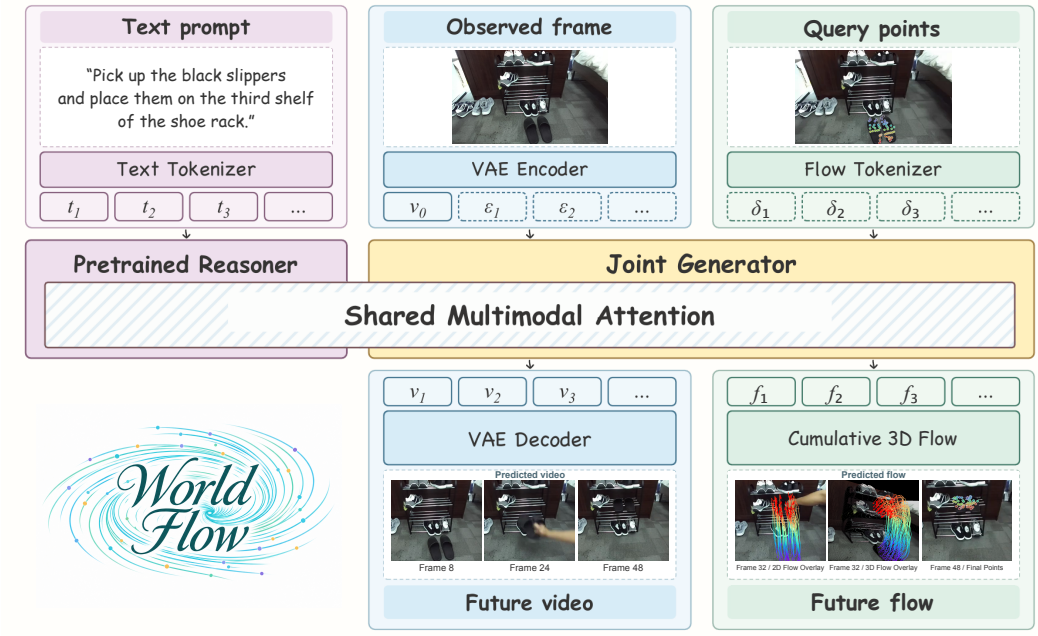

Figure 2: Overview of WORLDFLOW. Conditioned on text tokens t_j through the pretrained reasoner, clean observed-frame latents v_0, and query-point geometry, the generator jointly denoises video and flow from noise ϵ_j and δ_k through shared multimodal attention. Generated video latents v_j are decoded by the VAE, while generated 3D flow f_t is rescaled and accumulated from the initial query positions to recover metric trajectories. The 2D flow overlay shows generated 3D trajectories projected onto generated video frames. The 3D flow overlay shows generated 3D trajectories overlaid on a scene point cloud reconstructed from the initial frame's ground-truth depth.

Our objective is to model the conditional joint distribution

$$p_\theta \left(I_{1:T}, f_{1:T} \mid I_0, \mathcal{Q}, \mathcal{S}_0, c \right). \tag{3}$$

3.2 ARCHITECTURE

Our central design introduces 3D flow tokens as an additional generative modality for joint video-motion prediction. Built on the Cosmos-Nano framework (NVIDIA, 2026), WORLDFLOW retains the pretrained reasoner and extends the generator to jointly synthesize future video and 3D flow (Figure 2). The reasoner processes the language instruction, while the video VAE encodes the observed frame into clean latents. The flow tokenizer embeds noisy displacements together with initial query-point geometry and aligns their spatiotemporal positional coordinates with the video token grid (Section 3.3).

Within the Mixture-of-Transformers (MoT) architecture, video and flow tokens interact through full attention and attend to the reasoner's causal context. At inference, future video latents and flow are initialized from noise and jointly denoised, conditioned on the clean observed-frame latents, language context, and query geometry. The VAE decodes the generated video latents into RGB frames, while the predicted flow is rescaled and accumulated to recover 3D trajectories. This joint generation process produces explicit motion alongside video without a separate tracking stage.

3.3 3D FLOW TOKENIZATION

We represent each query point's displacement at each future frame as a continuous-valued token, augmented with initial 3D geometry and spatiotemporal positional encoding.

Normalization and sequence construction. To balance motion scales across domains, we normalize displacements by a domain-specific scalar estimated from training data and fixed throughout

training:

$$\bar{f}_{t,i} = \frac{f_{t,i}}{a_d}, \qquad a_d > 0. \tag{4}$$

The normalized displacements are arranged in frame-major order, yielding TN three-dimensional vectors for token construction.

Query point geometry encoder. We encode each query point's absolute initial position and its position relative to the query-set centroid using two parallel encoders. Each applies coordinate normalization, fixed Fourier features, and an MLP. Their outputs are concatenated and fused:

$$\mathbf{e}_i^{\mathrm{3D}} = \mathrm{MLP}\left([E_{\mathrm{abs}}(\mathbf{p}_{0,i}); E_{\mathrm{rel}}(\mathbf{p}_{0,i} - \boldsymbol{\mu})]\right), \tag{5}$$

where E_{abs} and E_{rel} denote the two encoders, and $\boldsymbol{\mu}$ is the query-set centroid. The geometry embedding is shared across all future tokens of the same query point.

Flow token construction. A linear projection maps noisy normalized displacements into the generator's hidden space, where they are combined with geometry and conditioning embeddings:

$$\mathbf{h}_{t,i}^{(\sigma)} = \mathrm{Linear}(\bar{f}_{t,i}^{(\sigma)}) + \mathbf{e}_i^{\mathrm{3D}} + \mathbf{e}_{\mathrm{cond}}(\sigma). \tag{6}$$

Here, σ denotes the denoising noise level, distinct from the physical frame index t. The conditioning embedding combines a learned flow-modality embedding with the noise-level embedding.

Spatiotemporal alignment. We assign flow tokens rotary positional coordinates using the same temporal and spatial conventions as video tokens. Each flow token is anchored to its future frame and the query point's initial image patch:

$$\mathbf{r}_{t,i} = \left(\frac{t}{q_t}, \left\lfloor \frac{v_{0,i}}{q_s p_s} \right\rfloor, \left\lfloor \frac{u_{0,i}}{q_s p_s} \right\rfloor\right), \tag{7}$$

where $(u_{0,i}, v_{0,i}) = \mathbf{s}_{0,i}$ denotes the initial image coordinates, q_t and q_s are the temporal and spatial VAE compression factors, and p_s is the latent patch size. These coordinates determine the rotary transformations of attention queries and keys:

$$\widetilde{\mathbf{q}}_{t,i} = R(\mathbf{r}_{t,i})\mathbf{q}_{t,i}, \qquad \widetilde{\mathbf{k}}_{t,i} = R(\mathbf{r}_{t,i})\mathbf{k}_{t,i}. \tag{8}$$

The shared positional convention encodes relative spatiotemporal relationships between flow and video tokens within multimodal attention.

3.4 Training Objectives

3.4.1 Joint flow matching

We jointly train video and 3D flow generation using a weighted rectified-flow objective:

$$\mathcal{L}_{\mathrm{FM}} = \lambda_{\mathrm{video}}\mathcal{L}_{\mathrm{video}} + \lambda_{\mathrm{flow}}\mathcal{L}_{\mathrm{flow}}, \tag{9}$$

where $\mathcal{L}_{\mathrm{video}}$ and $\mathcal{L}_{\mathrm{flow}}$ supervise vector-field prediction for video latents and normalized 3D flow, respectively. Both modalities share a per-sample noise level and are jointly denoised. Losses are computed separately for each modality, excluding conditioning positions from supervision.

3.4.2 Auxiliary motion objectives

To complement vector-field supervision with explicit constraints on predicted motion, we apply auxiliary objectives to clean normalized flow reconstructed from the predicted vector field. These objectives penalize frame-to-frame displacement error, cumulative displacement direction error, and relative endpoint error, encouraging both local accuracy and trajectory-level consistency.

Direction errors are weighted by ground-truth displacement magnitude, while points with small net displacement are excluded from endpoint supervision. All auxiliary terms operate in normalized coordinates and use noise-dependent weighting to emphasize lower-noise samples. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{dir}}\mathcal{L}_{\mathrm{dir}} + \lambda_{\mathrm{vmag}}\mathcal{L}_{\mathrm{vmag}} + \lambda_{\mathrm{end}}\mathcal{L}_{\mathrm{end}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{dir}}$, $\mathcal{L}_{\mathrm{vmag}}$, and $\mathcal{L}_{\mathrm{end}}$ denote the direction, displacement, and endpoint losses, respectively. Detailed formulations and derivations are provided in the supplementary material, and their contribution is evaluated through ablation experiments.

4 EXPERIMENTS

4.1 DATASETS

We assemble a heterogeneous corpus of paired videos and 3D point trajectories spanning simulated and real-world interactions. The training corpus contains approximately 1.44 million object-level samples, equivalent to 1,224 hours of standardized 49-frame sequences at 16 FPS. The data are organized into four collections:

- **Sim7** combines seven simulation datasets: ManiSkill (Tao et al., 2024), MIKASA (Cherepanov et al., 2026), MetaWorld (Yu et al., 2020), MetaFold (Chen et al., 2025a), RLBench (James et al., 2020), RoboCasa (Nasiriany et al., 2024), and RoboTwin (Mu et al., 2025), covering diverse robotic interactions and manipulation tasks.
- **Real3** comprises AgiBot World (Bu et al., 2025), EPIC-KITCHENS (Damen et al., 2018), and Something-Something V2 (Goyal et al., 2017), covering real-world robot and human interactions.
- **SpatialCog** is our self-collected dataset of tabletop and household interactions, with an emphasis on spatial relations in instruction-conditioned motion.
- **MolmoMotion-1M** (Zhang et al., 2026) contributes curated 3D motion annotations from six sources: EgoDex (Hoque et al., 2025), Xperience (Ropedia, 2026), YT-VIS (Yang et al., 2019), Stereo4D (Jin et al., 2025), DROID (Khazatsky et al., 2024), and MolmoSpaces (Kim et al., 2026).

Annotation and standardization. For Sim7 and Real3, we generate captions with Gemini-2.5 Pro (Comanici et al., 2025), obtain object masks with SAM 2 (Ravi et al., 2024), and extract 3D point trajectories using Track4World (Lu et al., 2026). SpatialCog pairs manually annotated captions with Track4World trajectories. For MolmoMotion-1M, we reuse the released motion annotations, apply source-specific geometry, scale, and trajectory-validity checks, and standardize the retained sequences into 49-frame clips. All collections use a shared representation consisting of video, a language instruction, initial 3D query points, and frame-to-frame 3D displacements. Additional processing details and per-source statistics are provided in the supplementary material.

Evaluation split. We introduce WorldFlowBench, a benchmark comprising 1,678 held-out evaluation samples: 396 from Sim7, 882 from Real3, and 400 from SpatialCog. The three subsets cover simulated interactions, real-world motion, and spatial-relation reasoning, respectively. All evaluation samples are excluded from training.

4.2 TRAINING CONFIGURATION AND STRATEGY

Implementation details. We train on 49-frame clips at 288×480 resolution with $N = 64$ query points. The video VAE uses temporal and spatial compression factors of $q_t = 4$ and $q_s = 16$, followed by 2×2 latent patchification ($p_s = 2$). Each clip therefore contains a $13 \times 9 \times 15$ video-token grid and 48×64 flow tokens. Training uses 64 NVIDIA H800 GPUs across eight nodes, with a per-GPU batch size of two and no gradient accumulation, yielding a global batch size of 128.

Training strategy. We adopt a two-stage training schedule. The first stage initializes the model from pretrained Cosmos-Nano checkpoints and trains on the SpatialCog training split for 32,000 steps. The learning rate follows linear warmup and cosine decay, with a peak of 10^{-4}. The resulting exponential moving average (EMA) weights initialize the second stage.

The second stage trains on the heterogeneous data mixture for 51,100 steps across three curriculum phases: simulation-focused training, mixed real-world adaptation, and refinement on the SpatialCog training split. During real-world adaptation, sampling gradually shifts from MolmoMotion to Real3. Each phase prioritizes its target domain while retaining cross-domain replay, with smooth transitions between sampling distributions. After warmup, the learning rate reaches 10^{-4}, transitions to an 8×10^{-5} plateau during real-world adaptation, and decays through late adaptation and final refinement.

We expand task prompts into structured instructions specifying data origin, camera motion, actor type, and environment. For classifier-free guidance, text conditioning is dropped with probability 0.1,

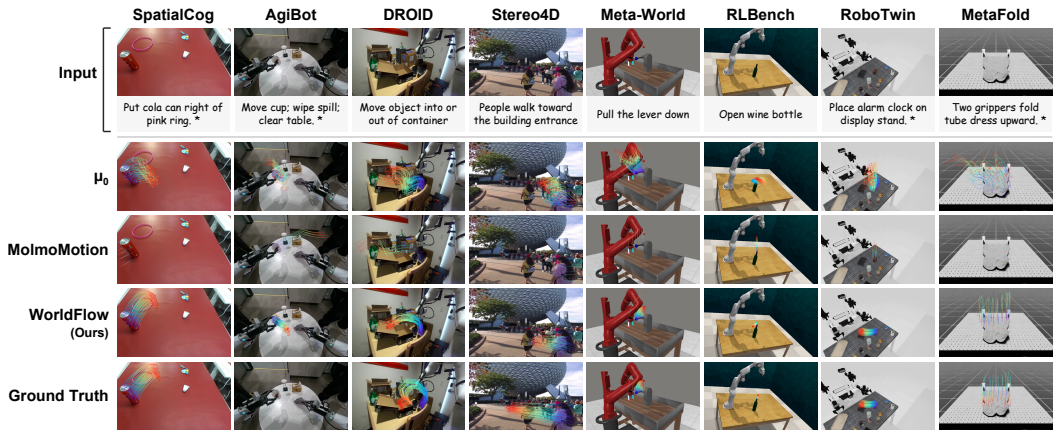

Figure 3: **Qualitative comparison on diverse real and simulated scenes.** We visualize predicted 3D query-point trajectories from μ_0, MolmoMotion, and WORLDFLOW across SpatialCog, AgiBot, DROID, Stereo4D, Meta-World, RLBench, RoboTwin, and MetaFold. The trajectories are projected onto the corresponding input frames for visualization.

while visual and geometric conditioning is retained. At inference, we use 40 rectified-flow sampling steps with a guidance scale of 2.

4.3 FLOW BENCHMARK

4.3.1 EVALUATION METRICS

We evaluate 3D point trajectories in metric coordinates. Average displacement error (ADE) measures the mean Euclidean position error across query points and future frames, while final displacement error (FDE) measures the mean position error at the final frame. Both are reported in meters. PWT measures the fraction of point–frame pairs whose position error falls within a given distance threshold. We report its average over thresholds of 1, 2, and 5 cm to assess trajectory accuracy at fine spatial scales.

4.3.2 QUANTITATIVE EVALUATION ON WORLDFLOWBENCH

We evaluate all learning-based methods under a single-frame protocol, where one RGB observation and a language instruction are provided to forecast a common future horizon without motion history. All predictions are converted into metric 3D point trajectories in the initial camera coordinate system. We additionally report a Static reference that predicts zero flow and therefore keeps every query point at its initial position throughout the prediction horizon.

We compare against two families of learned baselines. The video-generation-then-tracking family includes NovaFlow (Li et al., 2025), as well as pipelines built on Cosmos3-Nano (NVIDIA, 2026) or LingBot-Video-MoE (Robbyant Team et al., 2026), paired with Track4World (T4W) (Lu et al., 2026) or 4RC (Luo et al., 2026) for 3D tracking. The direct flow-generation family includes MolmoMotion-AR-H1 (Zhang et al., 2026), μ_0 (Lee et al., 2026b), and EgoWAM (Li et al., 2026a).

As shown in Table 1, WORLDFLOW achieves the lowest ADE and FDE among all evaluated learning-based methods on all three subsets. Relative to the strongest external learned baseline on each subset, it reduces ADE by approximately 46.9% on Sim7, 4.7% on Real3, and 54.2% on SpatialCog. It also achieves the highest PWT among learning-based methods on Sim7 and SpatialCog. On Real3, WORLDFLOW obtains a PWT of 0.311, compared with the best value of 0.332, while still delivering the lowest ADE and FDE. These results demonstrate consistent improvements in trajectory accuracy across simulated interactions, real-world motion, and spatial-reasoning tasks.

Paradigm	Model	Sim7			Real3			SpatialCog		
		ADE↓	FDE↓	PWT↑	ADE↓	FDE↓	PWT↑	ADE↓	FDE↓	PWT↑
Non-parametric	Static	0.096	0.201	0.573	0.062	0.086	0.464	0.143	0.278	0.412
Video generation + 3D tracking	NovaFlow	0.196	0.318	0.289	0.095	0.129	0.293	0.165	0.295	0.320
	Cosmos3-Nano + T4W	0.101	0.202	0.451	0.088	0.133	0.310	0.141	0.257	0.398
	Cosmos3-Nano + 4RC	0.096	0.196	<u>0.531</u>	0.085	0.122	**0.332**	0.141	0.263	0.401
	LingBot-Video-MoE + T4W	0.105	0.207	0.425	0.097	0.145	0.267	0.134	0.239	0.364
	LingBot-Video-MoE + 4RC	0.104	0.207	0.492	0.092	0.134	0.295	0.131	0.232	0.376
Direct flow generation	MolmoMotion-AR-H1	0.176	0.323	0.280	0.185	0.335	0.222	0.282	0.533	0.173
	μ_0	0.186	0.307	0.069	0.116	0.156	0.148	0.213	0.270	0.055
	EgoWAM	0.198	0.336	0.020	0.130	0.190	0.056	0.249	0.383	0.080
Ablations	WORLDFLOW (flow-only)	0.099	0.206	0.495	0.101	0.152	<u>0.320</u>	0.136	0.248	0.372
	WORLDFLOW + T4W	0.088	0.174	0.424	0.100	0.154	0.263	0.083	0.138	0.403
	WORLDFLOW + 4RC	0.088	0.170	0.483	0.092	0.132	0.291	0.098	0.175	0.411
	WORLDFLOW w/o auxiliary loss	<u>0.063</u>	<u>0.116</u>	0.527	<u>0.083</u>	<u>0.115</u>	0.299	<u>0.062</u>	<u>0.084</u>	<u>0.436</u>
Ours	WORLDFLOW	**0.051**	**0.094**	**0.564**	**0.081**	**0.113**	0.311	**0.060**	**0.078**	**0.449**

Table 1: **3D motion forecasting on WorldFlowBench.** ADE and FDE measure average and final position errors in meters, respectively; lower is better. PWT is the mean point–frame hit rate over distance thresholds of 1, 2, and 5 cm; higher is better. Bold and underline mark the best and second-best results in each column among non-static methods, including ablations. Static is reported as a reference and excluded from ranking. T4W denotes Track4World. For WORLDFLOW + T4W/4RC, trajectories are extracted from the videos generated by WORLDFLOW using the corresponding tracker.

4.3.3 ABLATION STUDIES

Flow-only versus joint generation. To evaluate the benefit of jointly modeling 3D motion and visual evolution, we compare WORLDFLOW with a flow-only variant using the same backbone, training data, and optimization recipe. The variant predicts only 3D flow, whereas the full model jointly generates video and flow. The full model reduces ADE by 48.5%, 19.8%, and 55.9% on Sim7, Real3, and SpatialCog, respectively, with lower FDE on all three subsets. On SpatialCog, FDE decreases from 0.248 to $0.078\,\mathrm{m}$, a 68.5% reduction. PWT also improves on Sim7 and SpatialCog, with a small decrease on Real3. These results support the value of complementary visual supervision for learning accurate 3D motion.

Direct flow generation versus post-hoc tracking. To evaluate flow as an explicit generative modality, we compare directly generated trajectories with those recovered by Track4World and 4RC from videos produced by the same WORLDFLOW checkpoint. This comparison holds the video generator and conditioning inputs fixed, contrasting explicit motion generation with motion recovery from generated pixels. Direct flow generation achieves lower ADE and FDE and higher PWT than both tracking pipelines on all three subsets. On SpatialCog, it achieves an ADE of $0.060\,\mathrm{m}$, compared with $0.083\,\mathrm{m}$ for Track4World and $0.098\,\mathrm{m}$ for 4RC. These results support incorporating explicit 3D flow into the world model's generative output rather than recovering motion through a separate tracking stage.

Auxiliary motion supervision. We compare the full model with a variant trained without auxiliary motion objectives. These objectives supervise displacement, cumulative motion direction, and trajectory endpoints through the reconstructed clean flow. Adding them improves ADE, FDE, and PWT on all three subsets, with the largest ADE reduction on Sim7, from 0.063 to $0.051\,\mathrm{m}$ (19.0%). PWT increases from 0.527 to 0.564 on Sim7, from 0.299 to 0.311 on Real3, and from 0.436 to 0.449 on SpatialCog. These consistent improvements indicate that explicit supervision of clean motion complements the primary flow-matching objective.

4.4 FLOW-GUIDED ROBOTIC MANIPULATION

We evaluate WORLDFLOW as an object-motion predictor within a modular manipulation pipeline using a seven-degree-of-freedom xArm, an XHand dexterous hand, and a ZED Mini RGB-D camera.

Group	Task Caption	WORLDFLOW	μ_0
Pick and Place	Place the watermelon doll into the small pink plate.	**3/3**	0/3
	Place the oolong tea into the large blue tray.	**2/3**	0/3
	Place the potato chips into the dusty pink basket.	**2/3**	0/3
	Place the yellow toy inside the red circle.	**2/3**	1/3
Stacking	Place the green biscuit box on the translucent box.	**3/3**	0/3
	Place the can of Red Bull on the translucent box.	**2/3**	1/3
	Place the beef curry on the paper towel.	**1/3**	0/3
	Place the yellow toy on the paper towel.	0/3	0/3
Relative Placement	Place the sunscreen in front of the tissue.	**1/3**	0/3
	Place the potato chips behind the Oreo box.	**2/3**	1/3
	Place the watermelon doll to the left of the cola.	**3/3**	1/3
	Place the oolong tea to the right of the small blue toy car.	**3/3**	0/3
Overall success rate		**66.7%**	11.1%

Table 2: **Real-world manipulation with different flow predictors.** Task entries report successful executions out of three trials, with 36 trials per method overall. Both methods use matched initial layouts, identical grasp results, and the same trajectory optimization and execution settings; only the flow predictor changes. Bold indicates the higher success rate.

Given an initial observation, a language instruction, and object query points, WORLDFLOW jointly generates video and 3D flow to preview the intended interaction.

Execution pipeline. T(R,O) Grasp (Fei et al., 2025) provides the initial grasp, while predicted flow specifies reference object trajectories. Following Dream2Flow (Dharmarajan et al., 2025), a local patch fitting optimizer fits a locally rigid region near the grasp to these trajectories and computes the corresponding arm motion. Thus, WORLDFLOW specifies object motion, while the grasp and control modules translate it into robot execution.

Evaluation and results. We evaluate 12 tasks spanning pick and place, stacking, and relative placement, with three randomized trials per task and 36 executions per method. Success requires stable placement at the instructed location or spatial relation. We select μ_0 (Lee et al., 2026b) as the baseline based on its competitive offline trajectory prediction performance among the reproducible direct flow prediction methods evaluated in our benchmark. For each paired trial, we replace only the flow predictor, keeping initial layouts, instructions, grasp results, and control settings identical. As shown in Table 2, WORLDFLOW achieves 66.7% success (24/36), compared with 11.1% (4/36) for μ_0, a gain of 55.6 percentage points. Under this controlled setting, WORLDFLOW provides more effective 3D motion references than μ_0 for real-world manipulation, translating improved flow prediction into higher task completion rates.

5 CONCLUSION

We presented WORLDFLOW, a world model that jointly generates future video and metric 3D flow from a single observation, a language instruction, and 3D query points. Continuous flow tokens are grounded in video coordinates and initial point geometry, enabling both modalities to be jointly denoised by a shared rectified-flow generator. Experiments across diverse embodied domains demonstrate improved motion forecasting over direct flow predictors and video-generation-then-tracking pipelines.

The current formulation requires predefined query points and inherits noise from automatically generated trajectory supervision. Future work will explore automatic query selection, multimodal futures, and joint generation of 3D flow and actions to study the benefits of explicit motion representations for policy learning.

AI USE STATEMENT

We used generative AI tools to assist with translation, provide feedback on experimental methodology, and interpret experimental results. Gemini 2.5 Pro was used to generate captions for Sim7 and Real3, as described in the dataset section. We additionally used AI tools to assist with manuscript drafting, language editing, organization, and LaTeX formatting. The authors reviewed the AI-assisted content and checked technical descriptions, numerical comparisons, and conclusions against the methodology and experimental records. The authors take full responsibility for the final content of this work, including all AI-assisted text, claims, and artifacts.

REFERENCES

Yuxuan Bian, Xin Chen, Zenan Li, Tiancheng Zhi, Shen Sang, Linjie Luo, and Qiang Xu. Video-As-Prompt: Unified Semantic Control for Video Generation. *arXiv preprint arXiv:2510.20888*, 2025. URL https://arxiv.org/abs/2510.20888.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xuan Hu, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

Danze Chen, Zeqing Wang, Ziyue Lin, Xingyi Yang, and Yeying Jin. H3-World: Turning Language Understanding into World Control. *arXiv preprint arXiv:2609.01560*, 2026. URL https://arxiv.org/abs/2609.01560.

Haonan Chen, Junxiao Li, Ruihai Wu, Yiwei Liu, Yiwen Hou, Zhixuan Xu, Jingxiang Guo, Chongkai Gao, Zhenyu Wei, Shensi Xu, et al. Metafold: Language-guided multi-category garment folding framework via trajectory generation and foundation model. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4339–4346. IEEE, 2025a.

Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, et al. SANA-Video: Efficient Video Generation with Block Linear Diffusion Transformer. *arXiv preprint arXiv:2509.24695*, 2025b. URL https://arxiv.org/abs/2509.24695.

Egor Cherepanov, Nikita Kachaev, Alexey Kovalev, and Aleksandr Panov. Memory, benchmark & robots: A benchmark for solving complex tasks with reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 9759–9815, 2026.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *European Conference on Computer Vision*, 2018. URL https://arxiv.org/abs/1804.02748.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging Properties in Unified Multimodal Pretraining. *arXiv preprint arXiv:2505.14683*, 2025. URL https://arxiv.org/abs/2505.14683.

Karthik Dharmarajan, Wenlong Huang, Jiajun Wu, Li Fei-Fei, and Ruohan Zhang. Dream2flow: Bridging video generation and open-world manipulation with 3d object flow. *arXiv preprint arXiv:2512.24766*, 2025.

Xin Fei, Zhixuan Xu, Huaicong Fang, Tianrui Zhang, and Lin Shao. T (r, o) grasp: Efficient graph diffusion of robot-object spatial transformation for cross-embodiment dexterous grasping. *arXiv preprint arXiv:2510.12724*, 2025.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE international conference on computer vision (ICCV)*, pp. 5843–5851. IEEE, 2017.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video, 2025. URL https://arxiv.org/abs/2505.11709.

Wenlong Huang, Yu-Wei Chao, Arsalan Mousavian, Ming-Yu Liu, Dieter Fox, Kaichun Mo, and Li Fei-Fei. Pointworld: Scaling 3d world models for in-the-wild robotic manipulation. *arXiv preprint arXiv:2601.03782*, 2026.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self Forcing: Bridging the Train-Test Gap in Autoregressive Video Diffusion. *arXiv preprint arXiv:2506.08009*, 2025. URL https://arxiv.org/abs/2506.08009.

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.

Linyi Jin, Richard Tucker, Zhengqi Li, David Fouhey, Noah Snavely, and Aleksander Holynski. Stereo4d: Learning how things move in 3d from internet stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10497–10509, June 2025.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proc. ECCV*, 2024.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, October 2025.

Jay Karhade, Nikhil Keetha, Yuchen Zhang, Tanisha Gupta, Akash Sharma, Sebastian Scherer, and Deva Ramanan. Any4d: Unified feed-forward metric 4d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14578–14589, 2026.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, et al. DROID: A large-scale in-the-wild robot manipulation dataset, 2024. URL https://droid-dataset.github.io/.

Yejin Kim, Wilbert Pumacay, Omar Rayyan, Max Argus, Winson Han, Eli VanderBilt, Jordi Salvador, Abhay Deshpande, Rose Hendrix, Snehal Jauhri, et al. MolmoSpaces: A large-scale open ecosystem for robot navigation and manipulation. *arXiv preprint arXiv:2602.11337*, 2026. URL https://arxiv.org/abs/2602.11337.

Seungjae Lee, Yoonkyo Jung, Inkook Chun, Yao-Chih Lee, Zikui Cai, Hongjia Huang, Aayush Talreja, Tan Dao, Yongyuan Liang, Jia-Bin Huang, et al. Tracegen: World modeling in 3d trace space enables learning from cross-embodiment videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20721–20731, 2026a.

Seungjae Lee, Yoonkyo Jung, Jusuk Lee, Jonghun Shin, Amir Hossein Shahidzadeh, Yao-Chih Lee, H Jin Kim, Jia-Bin Huang, and Furong Huang. μ_0: A scalable 3d interaction-trace world model. *arXiv preprint arXiv:2606.13769*, 2026b.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. EgoWAM: World Action Models Beyond Pixels with In-the-Wild Egocentric Human Data. *arXiv preprint arXiv:2607.08436*, 2026a. URL https://arxiv.org/abs/2607.08436.

Hongyu Li, Lingfeng Sun, Yafei Hu, Duy Ta, Jennifer Barry, George Konidaris, and Jiahui Fu. Novaflow: Zero-shot manipulation via actionable flow from generated videos. *arXiv preprint arXiv:2510.08568*, 2025.

Hongyu Li, Bowen Wen, Xinghao Zhu, Yixuan Wang, Yilun Du, Yunzhu Li, George Konidaris, Stan Birchfield, Soha Pouya, Chenran Li, et al. Hydra-0: Action flow for generalist world modeling and control. *arXiv preprint arXiv:2608.18077*, 2026b.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, Yujun Shen, and Yinghao Xu. Causal World Modeling for Robot Control. *arXiv preprint arXiv:2601.21998*, 2026c. URL https://arxiv.org/abs/2601.21998.

Sixu Lin, Junliang Chen, Huaiyuan Xu, Zhuohao Li, Guangming Wang, Yixiong Jing, Sheng Xu, Runyi Zhao, Brian Sheil, Lap-Pui Chau, et al. Roboflow4d: A lightweight flow world model toward real-time flow-guided robotic manipulation. *arXiv preprint arXiv:2605.17522*, 2026.

Jiahao Lu, Jiayi Xu, Wenbo Hu, Ruijie Zhu, Chengfeng Zhao, Sai-Kit Yeung, Ying Shan, and Yuan Liu. Track4world: Feedforward world-centric dense 3d tracking of all pixels. In *European Conference on Computer Vision*, pp. 184–204. Springer, 2026.

Yihang Luo, Shangchen Zhou, Yushi Lan, Xingang Pan, and Chen Change Loy. 4rc: 4d reconstruction via conditional querying anytime and anywhere. In *ICML*, 2026.

Yao Mu, Tianxing Chen, Zanxin Chen, Shijia Peng, Zhiqian Lan, Zeyu Gao, Zhixuan Liang, Qiaojun Yu, Yude Zou, Mingkun Xu, et al. Robotwin: Dual-arm robot benchmark with generative digital twins. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27649–27660. IEEE, 2025.

Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: Large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

NVIDIA. Cosmos 3: Omnimodal World Models for Physical AI. *arXiv preprint arXiv:2606.02800*, 2026. URL https://arxiv.org/abs/2606.02800.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, et al. Cosmos World Foundation Model Platform for Physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, et al. Advancing Open-source World Models. *arXiv preprint arXiv:2601.20540*, 2026. URL https://arxiv.org/abs/2601.20540.

Ropedia. Xperience-10M: A large-scale egocentric multimodal dataset with structured 3d/4d annotations. Hugging Face, 2026. URL https://huggingface.co/datasets/ropedia-ai/xperience-10m. Dataset.

Rustin Soraki, Homanga Bharadhwaj, Ali Farhadi, and Roozbeh Mottaghi. Objectforesight: Predicting future 3d object trajectories from human videos. In *European Conference on Computer Vision*, pp. 437–453. Springer, 2026.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

Team Seedance, De Chen, Liyang Chen, Xin Chen, et al. Seedance 2.0: Advancing Video Generation for World Complexity. *arXiv preprint arXiv:2604.14148*, 2026. URL https://arxiv.org/abs/2604.14148.

Team Wan, Ang Wang, Baole Ai, Bin Wen, et al. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Xinkai Wang, Chenyi Wang, Yifu Xu, Mingzhe Ye, Fucheng Zhang, Jialin Tian, Xinyu Zhan, Lifeng Zhu, Cewu Lu, and Lixin Yang. Lamp: Learning vision-language-action policy with 3d scene flow as latent motion prior. In *European Conference on Computer Vision (ECCV)*, 2026.

Yuxi Xiao, Jianyuan Wang, Nan Xue, Nikita Karaev, Yuri Makarov, Bingyi Kang, Xing Zhu, Hujun Bao, Yujun Shen, and Xiaowei Zhou. Spatialtrackerv2: Advancing 3d point tracking with explicit camera motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6726–6737, October 2025.

Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, et al. World Action Models are Zero-shot Policies. *arXiv preprint arXiv:2602.15922*, 2026. URL https://arxiv.org/abs/2602.15922.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, pp. 1094–1100. PMLR, 2020.

Chengbo Yuan, Chuan Wen, Tong Zhang, and Yang Gao. General flow as foundation affordance for scalable robot learning. *arXiv preprint arXiv:2401.11439*, 2024.

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. URL https://arxiv.org/abs/2603.16666.

Bowei Zhang, Lei Ke, Adam W Harley, and Katerina Fragkiadaki. Tapip3d: Tracking any point in persistent 3d geometry. *arXiv preprint arXiv:2504.14717*, 2025.

Jianing Zhang, Chenhao Zheng, Yajun Yang, Max Argus, Rustin Soraki, Winson Han, Taira Anderson, Chun-Liang Li, Shuo Liu, Jiafei Duan, et al. Molmomotion: Forecasting point trajectories in 3d with language instruction. *arXiv preprint arXiv:2606.18558*, 2026.

Jiaming Zhou, Qihang Zhang, Gangwei Xu, Cunxin Fan, Yujie Zhao, Ruilin Wang, Yiming Luo, Shuai Yang, Xing Zhu, Yujun Shen, Junwei Liang, and Yinghao Xu. Zero-WAM: In-Context World-Action Modeling from Human Videos for Open-Ended Task Generalization. *arXiv preprint arXiv:2608.26103*, 2026. URL https://arxiv.org/abs/2608.26103.