WORLD-4D: 4D EMBODIED WORLD MODELS FOR ROBOTIC MANIPULATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Robotic manipulation in the open world requires not only recognizing what a scene looks like, but also anticipating how its 3D structure moves under interaction. We argue that synchronized RGB, depth, and optical flow (RGB-DF) provide a geometry-aware projective representation that makes scene structure and pixel-level motion explicit. Compared to 2D pixel videos, this multi-modal synergy aligns visual appearance with geometric structure and temporal motion, creating a representation space significantly closer to low-level end-effector actions demanded by robotic systems, narrowing the gap between world prediction and policy learning. Building on this insight, we introduce **World-4D**, a generative model that co-produces future RGB frames, depth maps, and optical flow from a single RGB-D image and a language instruction within one unified diffusion process. This 4D world model features a tri-branch architecture that integrates cross-modal attention with frame-wise 3D RoPE, ensuring that appearance, geometry, and motion evolve consistently. To supply training data at scale, we curate **World-4D-200M**, a large-scale dataset containing approximately 254 million frames across egocentric human and robotic manipulation videos with pseudo-labels for depth and optical flow. We further propose **World-4D-Policy**, an inverse dynamics head that consumes the internal 4D representations of World-4D in a single forward pass, bypassing expensive multi-step denoising, to output robot actions in a closed-loop manner. Experiments show that World-4D produces temporally and spatially coherent 4D predictions, and that World-4D-Policy achieves strong performance on real-world bimanual manipulation tasks, outperforming the evaluated baselines on most tasks, with notable gains on tasks demanding spatial precision and temporal coordination.

1 INTRODUCTION

Robotic manipulation in the open world could greatly benefit from visual world models that predict how the environment would evolve given an agent's interactions (Zhao et al., 2026b; 2025a; Li et al., 2026; Agarwal et al., 2025; Ali et al., 2025). While recent generative video models (Ha & Schmidhuber, 2018a; Xiang et al., 2024; Zheng et al., 2024; Wang et al., 2025a) have shown encouraging progress in policy synthesis (Du et al., 2023b; Liang et al., 2024; Zhen et al., 2025), data simulation and generation (Zhu et al., 2024), and long-horizon planning (Du et al., 2023a; Li et al., 2025a), they remain limited by the 2D projective nature of pixels. This inherent limitation leads to a loss of critical spatial relationships, preventing precise 6-DoF pose estimation and depth-aware interaction (Hu et al., 2024; Agarwal et al., 2025; Li et al., 2026). Furthermore, 2D models often lack geometric grounding, leading to temporal inconsistencies such as fluctuating object scales and unphysical shape morphing, which hinders their utility in robust policy learning. Consequently, transitioning generative world modeling from 2D videos to geometry-integrated 4D scene evolution is an essential step toward a solid foundation for embodied intelligence.

Existing 4D scene-modeling approaches fall into two categories. The first builds on Neural Radiance Fields (NeRF) (Mildenhall et al., 2021) or 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023), which can be further divided into optimization-based methods (Zhao et al., 2024a;b; Yu et al., 2024; Bahmani et al., 2024) that are computationally intensive and scene-specific, and feed-forward models (Ren et al., 2024; Wu et al., 2025a) that prioritize speed but typically focus on object-centric generation. These approaches often require multi-view inputs or struggle to scale to

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.

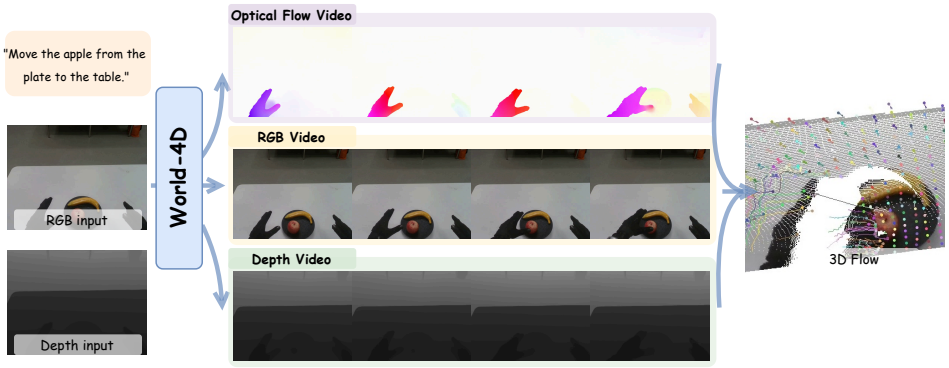


Figure 1: Given an input RGB-D image and description, World-4D generates RGB, depth, and optical flow videos synchronously, which can be further lifted into 3D scene flow (right).

complex scene-level environments. The second category comprises dynamic Structure-from-Motion (SfM) approaches (Wang et al., 2025b; Li et al., 2025b), which reconstruct time-varying point clouds but lack the generative capability to predict future states from a single image. Neither category readily provides a compact, scalable representation that integrates with the strong generative priors of pretrained video diffusion models.

To bridge this gap, we propose a lightweight projective 4D representation by predicting synchronized sequences of RGB, depth, and optical flow (RGB-DF): depth lifts each pixel to a 3D location, and depth together with optical flow can be back-projected into 3D scene flow under standard pinhole-camera assumptions, providing a per-point 3D motion cue (illustrated as the "3D Flow" in Fig. 1). Compared to RGB-only sequences, this representation makes geometry and motion explicit; compared to explicit 3D volumes or 4D Gaussians, it stays in a 2D-aligned format and therefore inherits the scalability and the rich generative priors of large-scale video diffusion models.

Building on this representation, we present **World-4D**, a 4D embodied world model, conditioned on a single RGB-D image and a text instruction, synchronously generates RGB, depth, and optical-flow videos within one shared denoising loop (Fig. 1). Specifically, we extend a pretrained video diffusion model (Wang et al., 2025a) into a tri-branch transformer, where each branch handles one modality with independent transformer with shared cross attention keys/values across modalities, while Joint Cross-Modal Attention modules enforce cross-modal consistency. This design preserves the strong generative priors of the pretrained backbone while allowing each modality to specialize, *i.e.*, textures for RGB, spatial geometry for depth, and motion displacements for optical flow. A key challenge for training World-4D is the absence of large-scale datasets with dense 4D annotations. To address this, we curate **World-4D-200M**, a large-scale hybrid dataset comprising approximately 254 million video frames drawn from egocentric human datasets (Damen et al., 2020; Wang et al., 2024) and robotic manipulation datasets (Wu et al., 2024; Liu et al., 2024; Jiang et al., 2025b; Wu et al., 2025b; Bu et al., 2025), each enriched with pseudo-annotations for depth and optical flow.

The RGB-DF representation offers a critical advantage: it aligns more closely with a robot's action space than raw 2D pixel changes. Consequently, a downstream policy trained on World-4D's internal 4D representations bypasses the heavy structural inference typically required when operating on 2D latents alone. Leveraging this synergy, we introduce **World-4D-Policy**, an inverse dynamics head that extracts robot actions directly from World-4D's predictive 4D features. By utilizing these internal latents in a single forward pass and bypassing the iterative denoising bottleneck, World-4D-Policy enables high-frequency, closed-loop control suitable for real-time interaction. In summary, our work makes the following contributions:

- We introduce a **projective 4D representation** that co-generates RGB, depth, and optical flow, and we show how it admits a natural 3D-scene-flow reading that makes geometry and motion explicit while staying compatible with large-scale video diffusion priors.
- We develop **World-4D**, a tri-branch 4D world model that co-generates spatiotemporally coherent RGB-DF sequences through mutual cross-modal interactions.

- We curate **World-4D-200M**, a large-scale 4D embodied video dataset with depth and optical flow annotations for training 4D embodied world model.
- We propose **World-4D-Policy**, which leverages the internal 4D representations to enable high-frequency, closed-loop robotic control.

2 Related Work

2.1 World Model

Learning a dynamics model of the world that supports downstream action generation has been a long-standing challenge (Ha & Schmidhuber, 2018b; Sutton, 1991). Early work learns world models in low-dimensional state spaces (Achille & Soatto, 2018; Lesort et al., 2018), which are efficient to train but difficult to generalize across visually diverse environments. With advances in generative modeling, a growing body of recent work has explored video models as foundation world models (Kong et al., 2024; Wang et al., 2025a; Yang et al., 2024). However, these models remain in the 2D pixel space, limiting their ability to capture 3D geometric structure and leaving a large representational gap between their predictions and the 3D actions a robot must produce.

3D world models attempt to close this gap by reasoning over meshes or explicit surfaces (Wang et al., 2021; Pfaff et al., 2020; Jiang et al., 2025a; Zhao et al., 2025b; Xia et al., 2025; 2024; Zhen et al., 2025; Guo et al., 2026), radiance fields or Gaussians (Mildenhall et al., 2021; Kerbl et al., 2023; Driess et al., 2023; Xie et al., 2024), or particle systems (Sanchez-Gonzalez et al., 2020; Abou-Chakra et al., 2024; Zhang et al., 2025; Chen et al., 2025). Hybrid approaches additionally reason over hierarchical structures (Kaelbling & Lozano-Pérez, 2011; Wang et al., 2025c; Zhao et al., 2026a). While these representations offer richer geometric reasoning, they often require multi-view inputs, are scene-specific, or lack the scalability of pretrained video priors. Most closely related to our work are (Zhen et al., 2025), which models the 4D scene from RGB-DN (RGB, Depth, and Normal) videos with language-conditioned control, and (Chen et al., 2025), which produces high-quality dynamic point clouds for novel-view video synthesis. Our work shares the goal of scalable 4D prediction but co-generates optical flow alongside RGB and depth. Unlike static-geometry cues such as surface normals (Zhen et al., 2025), optical flow back-projects into 3D scene flow, providing explicit dynamic cues for inverse dynamics—critical for dexterous manipulation where fine-grained object and end-effector trajectories decide success.

2.2 Future Prediction for Embodied Control

A growing line of work bridges generative modeling and control by using 2D future prediction to guide policy learning (Bharadhwaj et al., 2024; Ye et al., 2024; 2026; Bi et al., 2025; Hu et al., 2024). Representative methods include SuSIE (Black et al., 2023), which employs a goal-conditioned keyframe generator (Brooks et al., 2023), and UniPi (Du et al., 2023b), which learns inverse dynamics over generated sequences. Downstream actions are then derived via online planning (Hu et al., 2024; Williams et al., 2017; Hafner et al., 2019b; Pineau et al., 2003), offline policy synthesis (Hafner et al., 2019a; Hansen et al., 2023; Chua et al., 2018; Hafner et al., 2025), or inverse-dynamics models (Du et al., 2023b; Bi et al., 2025). However, because these pipelines operate entirely in 2D pixel space and often require repeated denoising for every action step, they face inherent limitations in both geometric accuracy and control reactivity. In contrast, **World-4D** operates on a unified 4D representation jointly encoding appearance, geometry, and motion, and **World-4D-Policy** consumes its internal predictive features in a single forward pass. By bypassing per-step video decoding and denoising, our approach enables high-frequency, closed-loop control that translates imagined 4D trajectories into precise real-time actions.

3 Method

To address the data scarcity in 4D generative modeling, we first introduce **World-4D-200M** in Sec. 3.1, a large-scale hybrid dataset specifically curated for training feed-forward 4D generative models. Building upon this, Sec. 3.2 presents **World-4D**, a framework capable of co-generating future sequences—including RGB frames, depth maps, and optical flow from a single RGB-D ob-

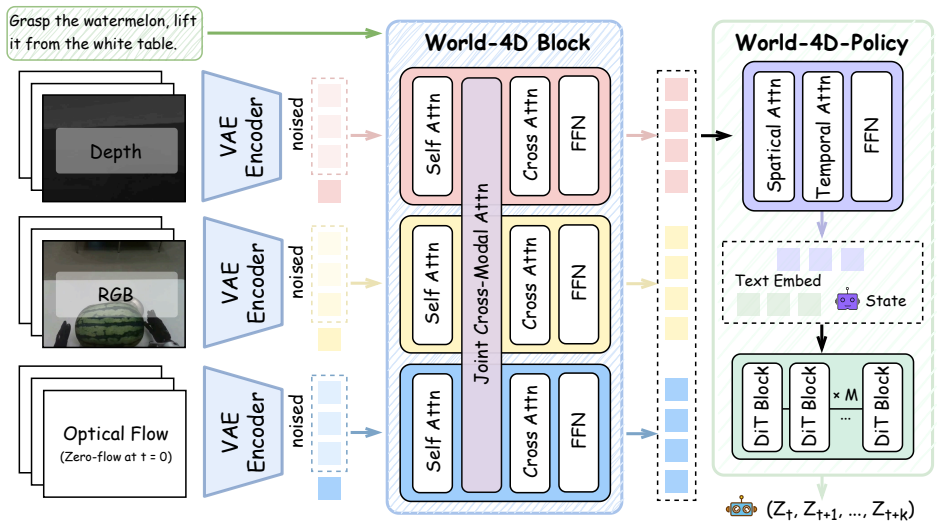


Figure 2: **Overview of World-4D.** Our pipeline leverages the large-scale World-4D-200M dataset to train a generative model capable of predicting future 4D sequences. Given a single RGB-D observation and a language instruction, World-4D co-generates future RGB frames, depth maps, and optical flow. These predictive 4D representations are then aggregated by World-4D-Policy to derive the final robot actions.

servation and a linguistic task description. Finally, in Sec. 3.3, we introduce **World-4D-Policy**, which leverages the predictive 4D representations from World-4D to derive final robotic actions.

3.1 World-4D-200M Dataset

To bridge the gap in large-scale 4D training data, we introduce World-4D-200M, a hybrid dataset comprising approximately 254.4 million video frames from human-centric (Damen et al., 2020; Wang et al., 2024) and robotic manipulation (Wu et al., 2024; Liu et al., 2024; Jiang et al., 2025b; Wu et al., 2025b; Bu et al., 2025) datasets. Each frame is enriched with automatically generated pseudo-annotations: fine-grained instructions (Bai et al., 2025), monocular depth (Lin et al., 2025), and dense optical flow (Morimitsu et al., 2025). More details about dataset composition are provided in Appendix G.

3.2 World-4D

We use RGB-D to denote an RGB frame paired with a depth map. World-4D is agnostic to how the depth map is obtained. In real-robot deployment, the depth is estimated from the RGB image using a monocular depth estimator, and the estimated depth is then used to construct the RGB-D input.

To achieve synchronized generation of RGB-DF sequences, we extend a pretrained video generative model into a tri-branch architecture (see the overview in Fig. 2); preliminaries on the flow-matching video backbone are provided in Appendix C. This representation is not merely a concatenation of channels; it admits a geometrically interpretable unprojection into 3D scene flow, as detailed in Appendix F. We denote the latents for modality $m \in \{\text{rgb}, \text{depth}, \text{flow}\}$ as $\boldsymbol{z}_t^m$, where $t \in [0, 1]$ represents the flow-matching timestep. Each $\boldsymbol{z}_t^m \in \mathbb{R}^{T \times C \times H \times W}$ encapsulates the entire temporal sequence of T frames.

Tri-branch Architecture. To inherit the powerful generative priors of the pretrained model while capturing the distinct characteristics of each modality, we expand the single-branch backbone (Wang et al., 2025a) into a tri-branch structure. This decoupled design allows each modality to model its unique feature distributions, such as complex textures for RGB, spatial geometry for depth, and motion displacements for flow to mitigate representation interference among divergent modalities.

Joint Cross-Modal Attention. To enforce cross-modal consistency, we append a Joint Cross-Modal Attention (JA) module after the intra-modal self-attention of every *third* transformer block (layers

$0, 3, \ldots, 27$), yielding 10 JA modules. Each branch emits one query and *one* shared key/value pair that is reused by all other branches' queries:

$$\boldsymbol{A}_l^m = \mathrm{Attn}\big(\mathrm{RoPE}(\boldsymbol{Q}_l^m), \mathrm{RoPE}(\boldsymbol{K}_l^{\text{cross}}), \boldsymbol{V}_l^{\text{cross}}\big), \quad \hat{\boldsymbol{z}}_l^m = \boldsymbol{z}_l^m + \tanh(g_l^m) \cdot \mathrm{OutProj}_l^m(\boldsymbol{A}_l^m), \tag{1}$$

where $\boldsymbol{K}_l^{\text{cross}} = \mathrm{concat}(\{\boldsymbol{K}_l^j\}_{j\neq m})$ and $\boldsymbol{V}_l^{\text{cross}} = \mathrm{concat}(\{\boldsymbol{V}_l^j\}_{j\neq m})$, so each query attends only to the two complementary modalities. Tokens are reshaped from $[B, T{\cdot}S, d]$ to $[B{\cdot}T, S, d]$, where S is the per-frame token count and d the hidden width, so that mixing is restricted to the *same temporal frame*, and 3D RoPE injects spatial position consistently across branches. Unlike the double zero-initialization of ControlNet (Zhang et al., 2023) which we found to cause a saddle-point deadlock—we zero-initialize only $\mathrm{OutProj}_l^m$ and set the gate $g_l^m{=}1$. The module starts as an exact residual, while $\tanh(1) \neq 0$ keeps gradients flowing into the gate so it can grow, shrink, or change sign. The complete formulation is given in Appendix J.

Phased Training Strategy. To bridge the significant distribution gaps between modalities, we propose a phased training paradigm: **Stage 1: Modality Adaptation.** In this initial stage, we disable the Joint Cross-Modal Attention and train the three branches independently. This allows the depth and flow branches to effectively adapt to their respective geometric and kinetic distributions. **Stage 2: Joint Attention Training.** We insert the JA modules and freeze the entire backbone together with the per-branch self-attention/FFN, so that only the JA modules and the modality embeddings are trainable. **Stage 3: Full-Parameter Joint SFT.** With the joint module already aligned, we unfreeze the entire model and continue on the full World-4D-200M. We additionally apply a Branch Dropout regularizer in Stages 2 and 3, detailed in Appendix J.

Training Objective. All three stages are optimized using the flow matching objective (Lipman et al., 2022). For each modality $m \in \mathcal{M} = \{\text{rgb}, \text{depth}, \text{flow}\}$, we learn a velocity field $\boldsymbol{v}_\theta^m$ that transports Gaussian noise $\boldsymbol{\epsilon}^m$ to data $\boldsymbol{z}_0^m$ along the path $\boldsymbol{z}_t^m = (1-t)\boldsymbol{z}_0^m + t\boldsymbol{\epsilon}^m$. The first frame of each modality is the clean image-to-video conditioning latent (a real RGB frame, a real depth frame, and a zero-flow frame, respectively) and is excluded from supervision; we use the slice $[1{:}]$ to denote frames $1, \ldots, T-1$. The total loss is:

$$\mathcal{L}_{\text{total}} = \sum_{m\in\mathcal{M}} \lambda_m \, \mathbb{E}_{\boldsymbol{z}_0^m, \boldsymbol{\epsilon}^m, t, \boldsymbol{c}} \Big[\big\| \boldsymbol{v}_\theta^m\big(\boldsymbol{z}_t^m, t, \boldsymbol{c}\big)_{[1:]} - \big(\boldsymbol{\epsilon}^m - \boldsymbol{z}_0^m\big)_{[1:]} \big\|_2^2 \Big], \tag{2}$$

where $\boldsymbol{c}$ denotes the text prompt together with the initial RGB-D observation, and $\boldsymbol{\epsilon}^{\text{rgb}} = \boldsymbol{\epsilon}^{\text{depth}} = \boldsymbol{\epsilon}^{\text{flow}}$ is a single Gaussian noise sample shared across the three modalities so that their denoising trajectories stay temporally aligned. The modality weights are $\lambda_{\text{rgb}} = \lambda_{\text{depth}} = 1$ throughout, while $\lambda_{\text{flow}} = 0.5$ in Stage 1 (the flow first frame carries no informative signal at warm-up) and $\lambda_{\text{flow}} = 1.0$ in Stages 2 and 3.

3.3 World-4D-Policy

We leverage World-4D as a *predictive 4D vision encoder*. Given the current RGB-D observation and instruction, we perform a forward pass through the frozen World-4D, which yields a latent trajectory encoding future dynamics—serving as a powerful representation for robotic manipulation. We extract intermediate hidden states across all branches and concatenate them along the channel dimension to form $F_p \in \mathbb{R}^{B\times T\times 3C\times H\times W}$, where C is the latent channel dimension per branch. By concatenating features from these decoupled branches, World-4D-Policy benefits from specialized representations: the RGB branch provides rich visual context, while the independent depth and flow branches offer explicit geometric and kinetic cues, respectively.

Flow Former. To compress 4D features, we use a Flow Former with learnable queries $\boldsymbol{Q}$. It processes F_p via frame-wise spatial cross-attention followed by temporal self-attention:

$$\boldsymbol{Q}_i' = \text{Spat-CrossAttn}(\boldsymbol{Q}_i, F_p[i]), \quad \boldsymbol{Q}'' = \mathrm{FFN}(\text{Temp-SelfAttn}(\boldsymbol{Q}')), i \in \{1, \ldots, T\} \tag{3}$$

where i indexes the frame sequence, and $\boldsymbol{Q}''$ encapsulate the predicted spatio-temporal dynamics.

Action Generation. We employ a flow matching (Lipman et al., 2022) policy to generate actions, following the objective defined in Eq. 2. Here, the velocity field v_ϕ operates on the action space $\boldsymbol{a}$, conditioned on the predictive 4D tokens $\boldsymbol{Q}''$, text embedding l_{emb}, and proprioception p_0. During inference, the action is derived via an ODE solver in $N = 4$ steps, enabling high-frequency closed-loop control through parallel action chunking (see Appendix H for details).

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

4.1.1 WORLD-4D

Our World-4D is built upon the Wan 2.2-TI2V-5B diffusion transformer (Wang et al., 2025a). We extend its single-branch RGB backbone into a tri-branch architecture for synchronous RGB, depth, and optical-flow synthesis, initializing the depth and flow branches by duplicating the pre-trained components. JA modules are inserted every three layers (at layers $0, 3, \ldots, 27$), yielding 10 modules. We adopt the three-stage curriculum of Sec. 3.2 to bridge modality distribution gaps. Full optimizer, EMA, and stage-wise hyper-parameters are in Appendix J.

4.1.2 WORLD-4D-POLICY

We use the frozen World-4D as a 4D vision encoder at 480×640 resolution, producing $T = 21$ latent frames. At each step we run a single-forward-pass feature extraction at normalized flow-matching time $\tau = 0.5$, taking hidden states from block 15 and concatenating the features of the three branches into F_p. A Flow Former compresses F_p, and a flow-matching policy head with 4-step Euler ODE sampling produces action chunks. Please see Appendix J for training details.

4.2 SETUPS AND BASELINES

World Model Metrics. To evaluate the generative performance and physical fidelity of our world model, we conduct benchmarks on a held-out test set of 50 video sequences, randomly sampled from (Wu et al., 2024; Liu et al., 2024; Jiang et al., 2025b). We assess the model across three axes: **(1)** ***Visual Synthesis Quality***: generative fidelity, following (Zhou et al., 2025) (IQ, MS, SC, Subj.) and pixel-level alignment (PSNR, SSIM, LPIPS) with ground truth; **(2)** ***Geometric Accuracy***: evaluating structural integrity via depth estimation metrics (AbsRel, $\delta_1 < 1.25$); **(3)** ***Temporal Motion Consistency***: optical-flow endpoint error (EPE/AEPE) in pixel-displacement space. Although the flow branch is trained on color-coded flow representations, all quantitative motion metrics are computed in flow-vector space after decoding. Detailed metric definitions are provided in Appendix B, and implementation details of the compared 4D world models are in Appendix D.

Real-world Task Metric. The primary evaluation metric is the Success Rate, defined as the percentage of successful task completions over 35 consecutive real-world trials. A trial is considered successful if the robot completes the task within 120 seconds.

Policy Baselines. We compare World-4D-Policy against two groups of baselines, all fine-tuned on the same real-robot demonstrations as ours: (1) policy-learning methods, including Diffusion Policy (DP) (Chi et al., 2025), the generalist foundation policies π_0 (Black et al., 2024) and $\pi_{0.5}$ (Intelligence et al., 2025), and the video-prediction-driven policies VPP (Hu et al., 2024) and Mimic-Video (Pai et al., 2025); (2) generative world models repurposed for control, marked with * in Tab. 2: TesserAct* (Zhen et al., 2025), a 4D world model that generates RGB, depth, and normal videos, and Wan-2.2-TI2V-5B* (Wang et al., 2025a), a video model that generates RGB. For these * variants, we extract their intermediate latent features and feed them into World-4D-Policy, so the comparison isolates the benefit of World-4D's internal 4D (RGB-DF) representation over the baselines' latents.

Hardware Platform. For real-world data collection and policy evaluation, we utilize the TIANJI M6 robot equipped with a WUJI HAND. A RealSense D435i camera is integrated to capture first-person view (FPV) images. Please refer to Appendix A for more details.

Robot Tasks. To demonstrate the generalization ability of our pipeline across diverse manipulation challenges, we evaluate our method on six distinct tasks: **(1)** ***Dual Picking***: Sequentially picking an apple and a banana from a plate with different arms and placing them on the table. **(2)** ***Block Pushing***: The left arm pushes a block to the center, followed by the right arm pushing it to a target zone. **(3)** ***Hand-over***: Transferring a cabbage from the left to the right gripper for placement. **(4)** ***Bimanual Lifting***: Synchronously lifting a watermelon plush into a plate. **(5)** ***Lid Placement***: Precisely picking and aligning a lid to cover a box. **(6)** ***Bowl Stacking***: Picking up one bowl and stacking it onto another. More details about our robot tasks are in Appendix I.

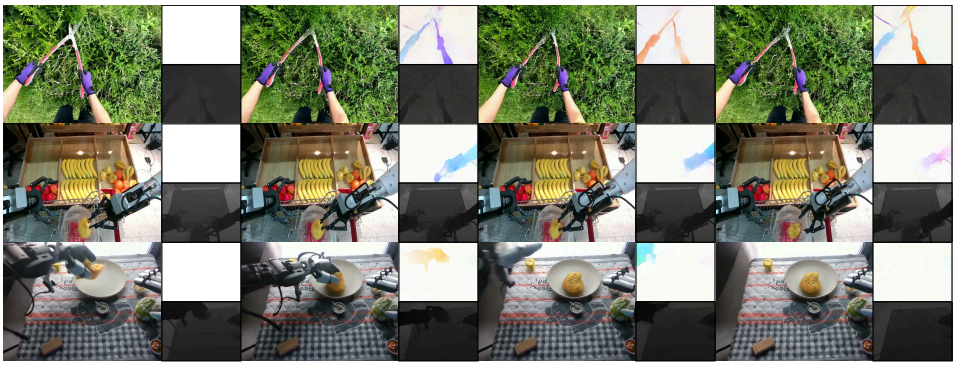


Figure 3: **Qualitative results of World-4D.** Starting from a single RGB-D image and a text prompt, World-4D synchronously generates RGB video, depth maps, and optical flow, preserving sharp geometric structures and producing temporally consistent motion fields.

Table 1: **Quantitative evaluation of 4D world modeling quality.** Metrics span visual synthesis (RGB), geometric structure (Depth), and temporal dynamics (Optical Flow). N/A denotes that the baseline lacks the native capability to generate specific modalities.

Method	RGB (Generation & Fidelity)							Depth		Flow
	IQ ↑	MS ↑	SC ↑	Subj. ↑	SSIM ↑	PSNR ↑	LPIPS ↓	AbsRel ↓	δ_1 ↑	AEPE ↓
Video Generation Models										
CogVideoX (Yang et al., 2024)	0.604	0.976	0.866	0.917	0.534	12.17	0.577	N/A	N/A	N/A
Wan-2.2-TI2V-5B (Wang et al., 2025a)	0.555	0.970	0.886	0.909	0.593	14.54	0.489	N/A	N/A	N/A
Wan-2.1-I2V-14B (Wang et al., 2025a)	**0.684**	0.988	0.891	0.956	0.536	12.72	0.568	N/A	N/A	N/A
4D World Models										
Free4D (Liu et al., 2025)	0.354	0.993	0.787	0.848	0.492	12.40	0.597	0.804	0.179	N/A
TesserAct (Zhen et al., 2025)	0.608	0.992	0.904	0.956	0.693	16.91	0.335	0.699	0.279	N/A
4DNeX (Chen et al., 2025)	0.637	0.994	0.917	0.986	0.649	14.47	0.404	0.423	0.327	N/A
World-4D	0.635	**0.995**	**0.957**	**0.992**	**0.754**	**17.85**	**0.269**	**0.310**	**0.610**	**0.170**
Ablation Study										
Independent Branches	0.613	0.986	0.922	0.971	0.683	17.26	0.346	0.737	0.245	0.247
w/o Modality Adaptation	0.621	0.992	0.952	0.975	0.699	17.85	0.303	0.507	0.479	0.231
w/o 4D Pre-training	0.615	0.982	0.879	0.938	0.651	16.25	0.344	0.797	0.263	0.729
w/o RoPE in JA	0.628	0.990	0.935	0.980	0.710	17.10	0.295	0.420	0.450	0.210
Shared FFN	0.618	0.985	0.902	0.965	0.695	16.50	0.320	0.580	0.380	0.280

4.3 Results

4D World Modeling. As shown in Tab. 1, compared to state-of-the-art video generation models such as Wan (Wang et al., 2025a) and CogVideoX (Yang et al., 2024), World-4D maintains highly competitive visual quality (IQ) while significantly outperforming them in reconstruction fidelity. This indicates that while general video models excel at creative synthesis, our model is better at preserving the structural and textural integrity of the scene during evolution, benefiting from the mutual regularization between depth and flow branches. When compared to existing 4D world models, World-4D demonstrates a clear advantage in both spatial and temporal accuracy. In terms of geometry, our model achieves a δ_1 of 0.610, nearly doubling the performance of 4DNeX (Chen et al., 2025) (0.327) and TesserAct (Zhen et al., 2025) (0.279). Regarding motion, World-4D uniquely provides synchronized optical flow with a low AEPE of 0.170, whereas most baseline 4D models lack the capability to produce explicit motion fields.

As visualized in Fig. 3, the generated depth and flow maps are not only internally consistent but also precisely aligned with the RGB texture changes. This result validates our core hypothesis: jointly modeling RGB, depth, and flow within a single diffusion loop acts as a powerful geometric and motion regularizer. Furthermore, Fig. 4 presents the 3D scene flow lifted from the co-generated depth and optical flow (Appendix F), demonstrating that our projective 4D predictions can be faithfully unprojected into coherent 3D motion trajectories. More qualitative results are provided in Appendix E.

Policy Learning. Tab. 2 compares World-4D-Policy against baselines spanning imitation learning (DP (Chi et al., 2025)), generalist foundation policies (π_0 (Black et al., 2024), $\pi_{0.5}$ (Intelligence

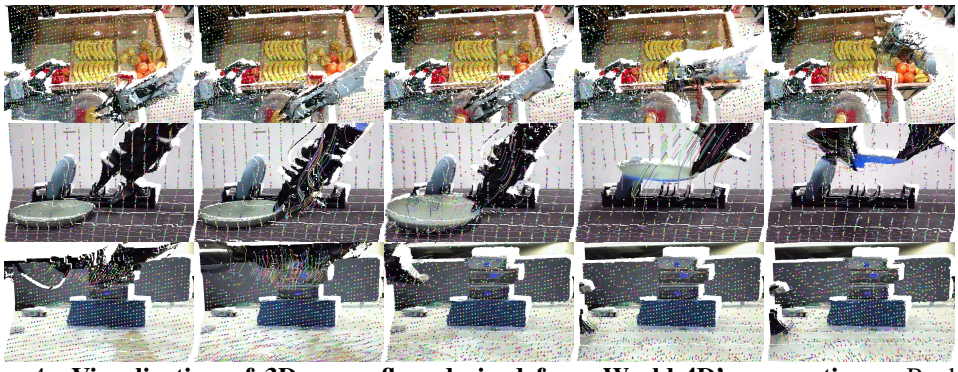


Figure 4: **Visualization of 3D scene flow derived from World-4D's generations.** By back-projecting the co-generated depth and optical flow into camera space (Appendix F), World-4D lifts its projective 4D predictions into explicit 3D scene flow. The trajectories, rendered as colored trails over the depth-ordered point cloud, are temporally coherent and geometrically plausible, confirming that the generated modalities are geometrically consistent with each other rather than mere visual hallucinations.

Table 2: **Success rates (%) on robotic manipulation tasks.** We compare World-4D-Policy against state-of-the-art policy learning baselines and foundation models across six challenging tasks. * denotes generative world models fine-tuned on our real-robot data; their intermediate latent features are fed into the *same* Flow Former for action decoding.

Method	Dual Picking	Block Pushing	Hand-over	Bimanual Lifting	Lid Placement	Bowl Stacking
DP (Chi et al., 2025)	77.14	85.71	17.14	88.57	57.14	57.14
π_0 (Black et al., 2024)	88.57	94.29	2.86	91.43	34.29	51.43
$\pi_{0.5}$ (Intelligence et al., 2025)	**94.29**	**100.00**	0.00	94.29	37.14	42.86
VPP (Hu et al., 2024)	85.71	91.43	5.71	94.29	57.14	51.43
Mimic-Video (Pai et al., 2025)	80.00	88.57	8.57	91.43	51.43	54.29
TesserAct* (Zhen et al., 2025)	91.43	94.29	17.14	94.29	57.14	60.00
Wan-2.2-TI2V-5B* (Wang et al., 2025a)	77.14	91.43	14.29	91.43	57.14	60.00
World-4D-Policy	**94.29**	97.14	**28.57**	**97.14**	**65.71**	**65.71**
Ablation Study						
ResNet (He et al., 2016) as encoder	71.43	88.57	11.43	85.71	51.43	60.00
RGB	77.14	91.43	14.29	91.43	57.14	60.00
RGB + Depth	91.43	91.43	28.57	97.14	60.00	62.86
RGB + Optical Flow	85.71	88.57	20.00	88.57	54.29	62.86

et al., 2025)), video-prediction-driven policies (VPP (Hu et al., 2024), Mimic-Video (Pai et al., 2025)), and world models repurposed for control (TesserAct*, Wan-2.2-TI2V-5B*), where it attains the best or tied-best success rate on five of the six tasks. On the precision-critical *Lid Placement* and *Bowl Stacking*, World-4D-Policy reaches 65.71%, surpassing the strongest baseline by 8.57% and 5.71% respectively. The gap is starkest on *Hand-over*, where foundation policies fail almost outright: their pre-training data is dominated by parallel-jaw grippers, leaving no priors for dexterous hand coordination, and a hand-over further requires reasoning about the relative 3D distance and self-occlusion between two high-DOF end-effectors, which 2D observations alone do not expose.

4.4 Ablation Study

Predictive Representations for Manipulation. The starred rows in Tab. 2 form a controlled comparison that fixes our Flow Former and action head and varies only *what the world model predicts*. Wan-2.2-TI2V-5B* (Wang et al., 2025a) forecasts RGB only, forcing the decoder to infer structure and motion from pixel changes; together with VPP (Hu et al., 2024) and Mimic-Video (Pai et al., 2025), these RGB-only predictors trail World-4D-Policy by 10.79% on average, with the largest gap on the motion-dominated *Hand-over*. TesserAct* additionally predicts depth and surface normals, which improves over the RGB-only predictors on most tasks; yet normals encode only *static* orientation and carry no inter-frame correspondence, so on the precision-critical *Lid Placement* and *Bowl Stacking* it closes none of the gap (57.14% and 60.00%, matching the strongest RGB-only predictor), and on the motion-dominated *Hand-over* it remains 11.43% behind ours. Optical flow

supplies exactly this missing correspondence: paired with depth it back-projects into 3D scene flow (Appendix F), a per-point displacement field of the same nature as the end-effector motion the policy must output. RGB-DF is thus the representation closest to the action space, and because every variant here reads its internal latents through the same decoder, the observed gaps reflect the quality of the representation itself rather than any decoding difference.

Effectiveness of Tri-branch Fusion. Training the three diffusion branches separately (**Independent Branches**) degrades depth (0.737 vs. 0.310 AbsRel) and flow (0.247 vs. 0.170 AEPE) markedly (Tab. 1), confirming that mutual feature interaction is crucial for cross-modal consistency in 4D sequences.

Necessity of Modality Adaptation. Skipping the initial modality-specific adaptation (**w/o Modality Adaptation**) leaves the depth and flow branches unable to bridge the gap between inherited RGB priors and their own distributions, dropping δ_1 from 0.610 to 0.479 (Tab. 1); modality adaptation is thus a prerequisite for effective multi-modal fusion.

Significance of Large-scale 4D Pre-training. Training exclusively on task-specific robot data (**w/o 4D Pre-training**) collapses performance across all dimensions, with AEPE surging from 0.170 to 0.729 (Tab. 1): task-specific data alone lacks the diversity needed to master complex spatio-temporal dynamics.

Effectiveness of Predictive Latents. Replacing the World-4D encoder with a ResNet-18 (He et al., 2016) image encoder (**ResNet (He et al., 2016) as encoder**) drops success across all tasks, most notably *Dual Picking* from 94.29% to 71.43% (Tab. 2), showing that static 2D features are insufficient and that World-4D's predictive dynamics are crucial for robust policy learning.

Impact of 4D Modalities. Using only RGB latents (**RGB**) lacks structural grounding; adding depth (**RGB + Depth**) brings large gains on spatial tasks such as *Hand-over* and *Bimanual Lifting*, while adding optical flow (**RGB + Optical Flow**) supplies complementary motion cues. The full World-4D-Policy combining all three achieves the best overall performance (74.76% success on average), confirming the synergy of visual context, geometry, and kinetics (Tab. 2).

Role of 3D RoPE in Joint Attention. Disabling the shared 3D RoPE inside the JA modules (**w/o RoPE in JA** in Tab. 1) breaks the geometric correspondence between modality features at identical (u, v) coordinates, dropping δ_1 from 0.610 to 0.450 and raising AEPE from 0.170 to 0.210. This confirms that 3D RoPE is the alignment bridge enabling pixel-level cross-modal fusion rather than global semantic averaging.

Per-branch FFN vs. Shared FFN. Replacing our per-branch FFNs with a single shared FFN (**Shared FFN** in Tab. 1) collapses geometric and motion accuracy (AbsRel 0.580 / δ_1 0.380, AEPE 0.280), since the heterogeneous manifolds of RGB textures, depth, and motion fields interfere when forced through one non-linear transform. Modality-specific capacity is thus essential for high-fidelity 4D generation despite its added parameters.

5 Conclusion

We presented **World-4D**, which shifts generative world modeling from 2D pixel sequences to consistent 4D scene evolution through a lightweight RGB-DF representation that bridges the scalability of video diffusion models and the geometric rigor required for robotic manipulation. With the tri-branch architecture trained on our large-scale **World-4D-200M** dataset, World-4D co-generates spatiotemporally coherent RGB, depth, and optical-flow sequences, and **World-4D-Policy** turns these predictive 4D representations into high-frequency, closed-loop control that outperforms existing 2D-based baselines. We believe World-4D provides a promising foundation for general-purpose embodied intelligence in the complex 3D world.

Limitation. Despite the reactive capabilities of World-4D-Policy, our framework has several limitations. First, the 4D sequence generation relies on a diffusion denoising process, which introduces computational overhead. Currently, our implementation achieves an effective control frequency of approximately 9 Hz on an NVIDIA RTX 5090 GPU; while sufficient for many tasks, this latency remains a bottleneck for ultra-high-frequency control. Second, our model is primarily optimized for egocentric perspectives. Extending 4D spatio-temporal consistency to multi-view systems or collaborative multi-robot setups remains an open challenge for future research.

REFERENCES

Jad Abou-Chakra, Feras Dayoub, and Niko Sünderhauf. Particlenerf: A particle-based encoding for online neural radiance fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5975–5984, 2024.

Alessandro Achille and Stefano Soatto. A separation principle for control in the age of deep learning. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(1):287–307, 2018.

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

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7996–8006, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. *arXiv preprint arXiv:2409.16283*, 2024.

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. *arXiv preprint arXiv:2512.13030*, 2025.

Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

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.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025.

Cheng Chi, Zhenjia Xu, Siyuan Feng, Eric Cousineau, Yilun Du, Benjamin Burchfiel, Russ Tedrake, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research*, 44(10-11):1684–1704, 2025.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Proc. of Advances in Neural Information Processing Systems*, 31, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Danny Driess, Zhiao Huang, Yunzhu Li, Russ Tedrake, and Marc Toussaint. Learning multi-object dynamics with compositional neural radiance fields. In *Conference on robot learning*, pp. 1755–1768. PMLR, 2023.

Yilun Du, Mengjiao Yang, Pete Florence, Fei Xia, Ayzaan Wahid, Brian Ichter, Pierre Sermanet, Tianhe Yu, Pieter Abbeel, Joshua B Tenenbaum, et al. Video language planning. *arXiv preprint arXiv:2310.10625*, 2023a.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *Proc. of Advances in Neural Information Processing Systems*, 36:9156–9172, 2023b.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. of Intl. Conf. on Machine Learning*, 2024.

Lijun Guo, Haoyu Zhao, Xingyue Zhao, Rong Fu, Linghao Zhuang, Siteng Huang, Zhongyu Li, and Hua Zou. Articulat3d: Reconstructing articulated digital twins from monocular videos with geometric and motion constraints. *arXiv preprint arXiv:2603.11606*, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Proc. of Advances in Neural Information Processing Systems*, 31, 2018a.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018b.

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

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proc. of Intl. Conf. on Machine Learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Wilson Yan, and Timothy Lillicrap. Training agents inside of scalable world models. *arXiv preprint arXiv:2509.24527*, 2025.

Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. *arXiv preprint arXiv:2310.16828*, 2023.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

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. *arXiv preprint arXiv:2412.14803*, 2024.

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.

Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: Physics-informed reconstruction and simulation of deformable objects from videos. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 7219–7230, 2025a.

Tao Jiang, Tianyuan Yuan, Yicheng Liu, Chenhao Lu, Jianning Cui, Xiao Liu, Shuiqi Cheng, Jiyang Gao, Huazhe Xu, and Hang Zhao. Galaxea open-world dataset and g0 dual-system vla model. *arXiv preprint arXiv:2509.00576*, 2025b.

Leslie Pack Kaelbling and Tomás Lozano-Pérez. Hierarchical task and motion planning in the now. In *2011 IEEE international conference on robotics and automation*, pp. 1470–1477. IEEE, 2011.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pp. 5148–5157, 2021.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

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

Timothée Lesort, Natalia Díaz-Rodríguez, Jean-Franois Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018.

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*, 2025a.

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*, 2026.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 9801–9810, 2023.

Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 10486–10496, 2025b.

Junbang Liang, Ruoshi Liu, Ege Ozguroglu, Sruthi Sudhakar, Achal Dave, Pavel Tokmakov, Shuran Song, and Carl Vondrick. Dreamitate: Real-world visuomotor policy learning via video generation. *arXiv preprint arXiv:2406.16862*, 2024.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Songming Liu, Lingxuan Wu, Bangguo Li, Hengkai Tan, Huayu Chen, Zhengyi Wang, Ke Xu, Hang Su, and Jun Zhu. Rdt-1b: a diffusion foundation model for bimanual manipulation. *arXiv preprint arXiv:2410.07864*, 2024.

Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 25571–25582, 2025.

Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 17810–17820, 2025.

Dujun Nie, Fengjiao Chen, Qi Lv, Jun Kuang, Xiaoyu Li, Xuezhi Cao, and Xunliang Cai. Lary: A latent action representation yielding benchmark for generalizable vision-to-action alignment. In *European Conference on Computer Vision (ECCV)*, 2026.

Jonas Pai, Liam Achenbach, Victoriano Montesinos, Benedek Forrai, Oier Mees, and Elvis Nava. mimic-video: Video-action models for generalizable robot control beyond vlas. *arXiv preprint arXiv:2512.15692*, 2025.

Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. *arXiv preprint arXiv:2010.03409*, 2020.

Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. Point-based value iteration: An anytime algorithm for pomdps. In *Ijcai*, volume 3, pp. 1025–1032, 2003.

Jiawei Ren, Kevin Xie, Ashkan Mirzaei, Hanxue Liang, Xiaohui Zeng, Karsten Kreis, Ziwei Liu, Antonio Torralba, Sanja Fidler, Seung W Kim, et al. L4gm: Large 4d gaussian reconstruction model. *Proc. of Advances in Neural Information Processing Systems*, 37:56828–56858, 2024.

Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter Battaglia. Learning to simulate complex physics with graph networks. In *Proc. of Intl. Conf. on Machine Learning*, pp. 8459–8468. PMLR, 2020.

Richard S Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *ACM Sigart Bulletin*, 2(4):160–163, 1991.

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

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 10510–10522, 2025b.

Qineng Wang, Wenlong Huang, Yu Zhou, Hang Yin, Tianwei Bao, Jianwen Lyu, Weiyu Liu, Ruohan Zhang, Jiajun Wu, Li Fei-Fei, et al. Enact: Evaluating embodied cognition with world modeling of egocentric interaction. *arXiv preprint arXiv:2511.20937*, 2025c.

Xiaofeng Wang, Kang Zhao, Feng Liu, Jiayu Wang, Guosheng Zhao, Xiaoyi Bao, Zheng Zhu, Yingya Zhang, and Xingang Wang. Egovid-5m: A large-scale video-action dataset for egocentric video generation. *arXiv preprint arXiv:2411.08380*, 2024.

Grady Williams, Andrew Aldrich, and Evangelos A Theodorou. Model predictive path integral control: From theory to parallel computation. *Journal of Guidance, Control, and Dynamics*, 40 (2):344–357, 2017.

Kun Wu, Chengkai Hou, Jiaming Liu, Zhengping Che, Xiaozhu Ju, Zhuqin Yang, Meng Li, Yinuo Zhao, Zhiyuan Xu, Guang Yang, et al. Robomind: Benchmark on multi-embodiment intelligence normative data for robot manipulation. *arXiv preprint arXiv:2412.13877*, 2024.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 26057–26068, 2025a.

Shihan Wu, Xuecheng Liu, Shaoxuan Xie, Pengwei Wang, Xinghang Li, Bowen Yang, Zhe Li, Kai Zhu, Hongyu Wu, Yiheng Liu, et al. Robocoin: An open-sourced bimanual robotic data collection for integrated manipulation. *arXiv preprint arXiv:2511.17441*, 2025b.

Hongchi Xia, Zhi-Hao Lin, Wei-Chiu Ma, and Shenlong Wang. Video2game: Real-time interactive realistic and browser-compatible environment from a single video. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4578–4588, 2024.

Hongchi Xia, Entong Su, Marius Memmel, Arhan Jain, Raymond Yu, Numfor Mbiziwo-Tiapo, Ali Farhadi, Abhishek Gupta, Shenlong Wang, and Wei-Chiu Ma. Drawer: Digital reconstruction and articulation with environment realism. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 21771–21782, 2025.

Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024.

Tianyi Xie, Zeshun Zong, Yuxing Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 4389–4398, 2024.

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

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Proc. of Advances in Neural Information Processing Systems*, 37:45256–45280, 2024.

Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Particle-grid neural dynamics for learning deformable object models from rgb-d videos. *arXiv preprint arXiv:2506.15680*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pp. 3813–3824, 2023.

Haoyu Zhao, Chen Yang, Hao Wang, Xingyue Zhao, and Wei Shen. Sg-gs: Photo-realistic animatable human avatars with semantically-guided gaussian splatting. *arXiv preprint arXiv:2408.09665*, 2024a.

Haoyu Zhao, Xingyue Zhao, Lingting Zhu, Weixi Zheng, and Yongchao Xu. Hfgs: 4d gaussian splatting with emphasis on spatial and temporal high-frequency components for endoscopic scene reconstruction. *arXiv preprint arXiv:2405.17872*, 2024b.

Haoyu Zhao, Sixu Lin, Qingwei Ben, Minyue Dai, Hao Fei, Jingbo Wang, Hua Zou, and Junting Dong. Smap: Self-supervised motion adaptation for physically plausible humanoid whole-body control. *arXiv preprint arXiv:2505.19463*, 2025a.

Haoyu Zhao, Hao Wang, Xingyue Zhao, Hao Fei, Hongqiu Wang, Chengjiang Long, and Hua Zou. Physsplat: Efficient physics simulation for 3d scenes via mllm-guided gaussian splatting. In *Proc. of IEEE Intl. Conf. on Computer Vision*, pp. 5242–5252, 2025b.

Haoyu Zhao, Cheng Zeng, Linghao Zhuang, Yaxi Zhao, Shengke Xue, Hao Wang, Xingyue Zhao, Zhongyu Li, Kehan Li, Siteng Huang, et al. High-fidelity simulated data generation for real-world zero-shot robotic manipulation learning with gaussian splatting. *IEEE Robotics and Automation Letters*, 2026a.

Haoyu Zhao, Linghao Zhuang, Xingyue Zhao, Cheng Zeng, Haoran Xu, Yuming Jiang, Jun Cen, Kexiang Wang, Jiayan Guo, Siteng Huang, et al. Towards affordance-aware robotic dexterous grasping with human-like priors. In *Proc. of the AAAI Conf. on Artificial Intelligence*, volume 40, pp. 13126–13134, 2026b.

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Learning 4d embodied world models. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 5337–5347, 2025.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical ai. *arXiv preprint arXiv:2512.01989*, 2025.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv preprint arXiv:2406.14540*, 2024.

AI USE STATEMENT

In this work, we used generative AI tools for editing the research paper to improve readability, formatting references, and suggesting structural improvements (tasks with recommended disclosure). We have not used generative AI tools to generate synthetic datasets, implement methods, propose or refine hypotheses, design research methodology, formulate mathematical claims, or interpret results (tasks with required disclosure), and other required disclosure tasks such as assisting with mathematical proofs are not applicable to this work.

We have reviewed all AI-assisted work. Specifically, all text polished by large language models (LLMs) was manually verified by the authors to ensure scientific accuracy, original meaning, and compliance with the ICLR Code of Ethics were strictly preserved. No core scientific claims, code implementations (e.g., the tri-branch architecture and Flow Former), or data curation processes (e.g., the World-4D-200M dataset construction) were generated by AI. We take full responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

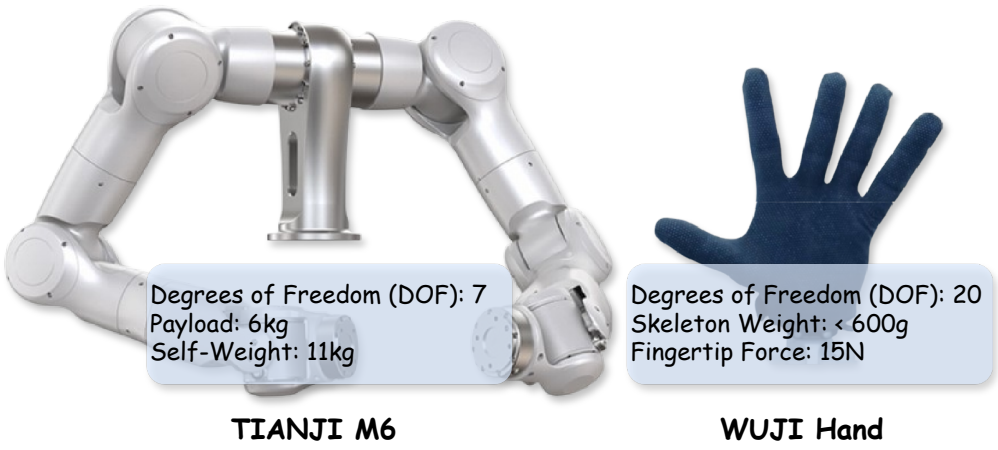


Figure 5: **Standardized Experimental Platform.** All real-world experiments are conducted on a unified hardware configuration consisting of the TIANJI M6 7-DOF robotic arm and the 20-DOF WUJI dexterous hand. This integrated system provides the high-precision control and high-dimensional actuation space required for complex manipulation tasks.

A Real Robot System Setup

Our real robot is built on the TIANJI M6 and WUJI Hand, as shown in Fig. 5. The high-level policy replans at approximately 0.9 Hz, producing an action chunk of $K = 10$ actions. The chunk is executed at an effective action rate of approximately 9 Hz. Command transmission latency is kept between 18 and 30 milliseconds. The low-level interface operates at a frequency of 500 Hz, ensuring smooth real-time control. The communication between the control policy and the low-level interface is realized through LCM (Lightweight Communications and Marshalling).

We collect real-world demonstration data through a teleoperation system, as shown in Fig. 6. Our hardware setup consists of dual Tianji 7-DOF robotic arms and dual Wuji 20-DOF dexterous hands, yielding a total of 54 degrees of freedom.

For arm control, the human operator wears five HTC Vive trackers (mounted on the chest, both wrists, and both upper arms). The system computes wrist-to-chest relative transforms at 100-120 Hz and feeds them into a Pinocchio-based inverse kinematics solver running in a separate process. The resulting joint commands are further smoothed by a Ruckig trajectory generator with velocity, acceleration, and jerk constraints before being sent to the robot arms at 200 Hz.

For hand control, the operator wears Manus data gloves. The raw glove signals are converted to a 21-point MediaPipe hand skeleton format and retargeted to the 20-DOF Wuji hand joint space via a dedicated retargeting module, with an exponential moving average filter applied for motion smoothing.

B Additional Details on Evaluation Metrics

This section provides technical definitions and implementation details for all evaluation metrics used in our experiments.

B.1 Generative Quality

We adopt a subset of task-aligned metrics from PAI-Bench (Zhou et al., 2025) to assess the visual and temporal quality of generated videos:



Figure 6: Operator to collect real world data.

- **Imaging Quality (IQ):** Evaluates low-level visual fidelity including clarity, noise levels, and compression artifacts. We employ the MUSIQ (Ke et al., 2021) predictor trained on the SPAQ dataset to compute a frame-level perceptual quality score.
- **Motion Smoothness (MS):** Quantifies temporal coherence and physical plausibility of motion dynamics. It is computed as the reconstruction error between generated frames and those synthesized via the AMT-S (Li et al., 2023) frame interpolation model, where lower interpolation error indicates smoother motion.
- **Subject Consistency (SC):** Measures the identity stability of the primary subject across the video sequence. We compute the mean pairwise cosine similarity of DINO (Zhang et al., 2022) ViT-B/16 features between the first frame and all subsequent frames.
- **I2V-Subject (Subj.):** For image-to-video generation, this metric evaluates how faithfully the model preserves the identity of the input reference image throughout the generated sequence. It is computed as the DINO feature similarity between the conditioning image and each generated frame.
- **PSNR (Peak Signal-to-Noise Ratio):** Measures pixel-wise reconstruction accuracy between generated and GT frames. Higher values indicate lower distortion. We report the mean PSNR across all frames (excluding the shared first frame) and all video samples.
- **SSIM (Structural Similarity Index):** Evaluates the preservation of structural information by jointly considering luminance, contrast, and structural components between generated and GT frame pairs.
- **LPIPS (Learned Perceptual Image Patch Similarity):** Measures perceptual distance between generated and GT frames using deep features from a pre-trained AlexNet. Lower values indicate higher perceptual similarity, and LPIPS is generally better aligned with human visual judgments than pixel-level metrics.

B.2 Geometric Accuracy

For methods that jointly predict depth maps, we evaluate structural fidelity of the estimated geometry using scale-invariant depth metrics:

- **Absolute Relative Error (AbsRel ↓):** Measures the mean relative deviation between the predicted depth $\hat{d}$ and ground-truth depth d^*. Since different methods may produce depth

in arbitrary scales, we first perform scale-invariant alignment via median scaling: $s = \text{median}(d^*)/\text{median}(\hat{d})$, then compute:

$$\text{AbsRel} = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \frac{|s \cdot \hat{d}_p - d_p^*|}{d_p^*}, \tag{4}$$

where $\mathcal{V}$ denotes the set of valid pixels ($d^* > 0$).

- **Threshold Accuracy ($\boldsymbol{\delta_1 < 1.25} \uparrow$):** Reports the percentage of pixels whose depth ratio falls within a tolerance threshold:

$$\delta_1 = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \mathbb{1}\left[\max\left(\frac{s \cdot \hat{d}_p}{d_p^*}, \frac{d_p^*}{s \cdot \hat{d}_p} \right) < 1.25 \right]. \tag{5}$$

Higher values indicate better geometric alignment with the ground truth.

B.3 Motion Consistency

For methods that predict optical flow, we evaluate the temporal dynamics accuracy:

- **Average End-Point Error (AEPE $\downarrow$):** Measures the accuracy of predicted optical flow fields against ground-truth flow. For each pixel p with predicted flow $\hat{\mathbf{f}}_p = (\hat{u}_p, \hat{v}_p)$ and ground-truth flow $\mathbf{f}_p^* = (u_p^*, v_p^*)$, the end-point error is defined as:

$$\text{AEPE} = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \left\| \hat{\mathbf{f}}_p - \mathbf{f}_p^* \right\|_2 = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \sqrt{(\hat{u}_p - u_p^*)^2 + (\hat{v}_p - v_p^*)^2}. \tag{6}$$

Although the flow branch operates on color-coded flow representations during training (to reuse the pretrained RGB video VAE), all quantitative flow metrics are computed in flow-vector space. Specifically, generated color-coded flow videos are first decoded back to flow vectors using the inverse color-wheel mapping before computing AEPE. The metric is averaged over all valid frames (excluding the first frame, which has no temporal reference) and all video samples. Lower AEPE indicates more accurate motion prediction.

C Preliminaries

Our study builds upon the Wan family of video generation models (Wang et al., 2025a), a latent video diffusion transformer capable of generating temporally coherent video from a single input image or text prompt. The model consists of a 3D variational autoencoder $(\mathcal{E}, \mathcal{D})$ and a transformer-based diffusion model parameterized by Θ. Given an input latent $z_0 = \mathcal{E}(V_0)$, the forward process follows the rectified flow formulation (Esser et al., 2024), where the noised latent is generated by linear interpolation:

$$z_t = (1 - t)\, z_0 + t\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{7}$$

with timestep $t \in [0, 1]$. The denoising process learns a velocity field $v_\Theta(z_t, t)$ that guides the transformation of noise back to data. The model is trained using a conditional flow matching (Lipman et al., 2022), with objective:

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t, z_0, \epsilon}\Big[\| v_\Theta(z_t, t) - u_t(z_0 \mid \epsilon) \|_2^2 \Big] \tag{8}$$

where u_t is the target velocity derived analytically from the forward process. At inference, a sequence of latent frames is recovered by integrating v_Θ over time.

In the image-to-video (I2V) setting, the model is conditioned on an initial image I_0 encoded as $z_{\text{img}} = \mathcal{E}(I_0)$. The transformer-based denoiser $\mathcal{F}_\Theta$ autoregressively predicts video latents $\{z^{(f)}\}_{f=1}^{F}$, starting from z_{img} and producing temporally consistent sequences. In the text-to-video (T2V) setting, the model is instead conditioned on a text prompt p encoded as $z_{\text{text}} = \mathcal{T}(p)$ and starts the autoregressive generation from noise. The final video is reconstructed as $\hat{V} = \mathcal{D}(z^{(1)}, \ldots, z^{(F)})$.

D Baseline Implementation Details

Free4D. Free4D (Liu et al., 2025) lifts a single image into a dynamic 4D Gaussian Splatting (4DGS) representation. Given the first frame of each ground-truth video as input, we run the full Free4D pipeline: (1) the built-in ViewCrafter module with DUSt3R-based monocular depth estimation synthesizes 25 novel views from the input image via a video diffusion model at 576×1024 resolution; (2) COLMAP sparse reconstruction estimates camera poses and produces a sparse point cloud for Gaussian initialization; (3) the 4DGS model with HexPlane-based deformation fields is optimized for 10,000 iterations (3,000 static initialization + 7,000 joint optimization), with temporal resolution $[64, 64, 64, 150]$; (4) RGB and depth videos are rendered from the original camera viewpoint across all timesteps. Since the rendered depth is in arbitrary scale, we apply per-frame median scaling alignment before computing depth metrics.

4DNeX. 4DNeX (Chen et al., 2025) is a feed-forward 4D scene generation framework that repurposes the Wan2.1-I2V-14B (Wang et al., 2025a), fine-tuned with learnable domain embeddings and LoRA adapters, to jointly produce RGB appearance and per-pixel XYZ point-cloud geometry from a single image and a text prompt. We adopt the official variant, using the provided 4dnex-lora weights (rank 64, fused at scale 0.5). For each sequence we extract the first frame of the ground-truth video as the conditioning image and use the corresponding caption, appending the official `POINTMAP_STYLE.` suffix as required by the released model. For each sample we run 50 denoising steps with classifier-free guidance scale 5.0 and seed 42, generating 49 frames at 24 fps. We obtain the depth video by taking the z-channel of the predicted pointmap, min–max normalized per sequence to $[0, 255]$.

TesserAct. TesserAct (Zhen et al., 2025) is built upon CogVideoX-5b-I2V (Yang et al., 2024) and fine-tuned to jointly generate RGB, depth, and surface normal videos from a single initial frame and a language instruction. We adopt the official checkpoint. Since TesserAct requires depth and normal maps as additional conditioning inputs, we first extract the initial RGB frame from each ground-truth video, then apply network to estimate monocular depth and surface normals. The three modalities are concatenated along the channel dimension to form a 9-channel input. We generate 49 frames at 640×480 resolution using 50 DDPM denoising steps with guidance scale 7.5, image guidance scale 1.5. The model outputs RGB, depth, and normal videos concatenated along the width axis; we split along width to obtain separate RGB and depth predictions for evaluation.

E Additional Qualitative Visualizations of World-4D

To further showcase the generation quality of World-4D, we provide extended paired visualizations of **RGB**, **Depth**, and **Optical Flow** generated by World-4D (Fig. 7). World-4D synchronously predicts future RGB, depth maps, and optical flow from a single RGB-D observation. These results demonstrate:

(i) **Cross-modal Consistency:** The geometric structures in depth maps and motion boundaries in optical flow are precisely aligned with the RGB textures.

(ii) **Physical Fidelity:** The model accurately captures complex 4D dynamics, such as object displacements and multi-contact interactions, in both human-centric and robot-specific environments.

(iii) **Temporal Coherence:** The generated sequences maintain stability over time without significant flickering or structural morphing.

F 3D Scene Reconstruction from Multi-Modal Videos

A key advantage of the RGB-DF (*i.e.*, RGB, depth, and optical flow) representation is its inherent geometric interpretability. By combining the co-generated depth and optical flow, we can reconstruct a temporally consistent 3D scene and derive *metric scene flow*.

Geometric Unprojection. Given the generated depth map D_t at frame t, each pixel $\mathbf{p}_t = [u, v, 1]^\top$ in homogeneous coordinates is unprojected into the 3D camera space as:

$$\mathbf{P}_t = D_t(u, v) \cdot \mathbf{K}^{-1}\mathbf{p}_t, \tag{9}$$

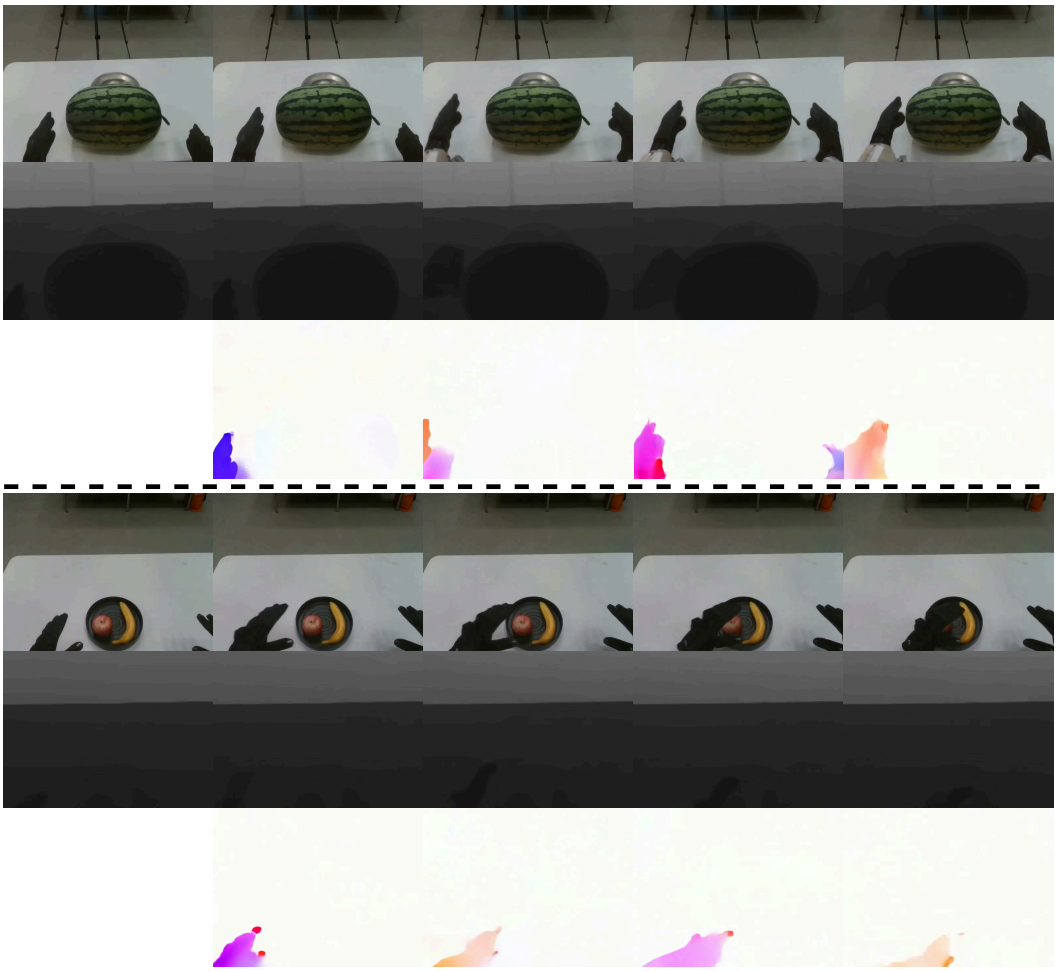


Figure 7: Extended qualitative results. Each row displays the generated RGB, depth, and optical flow sequences. The results highlight World-4D's ability to produce spatially and temporally coherent 4D predictions across various manipulation scenarios.

where $\mathbf{K}$ is the camera intrinsic matrix. This process lifts the 2D projective sequence into a metric 3D point cloud $\mathcal{C}_t = \{\mathbf{P}_t^i\}_{i=1}^{H\times W}$.

Metric Scene Flow Derivation. To capture the underlying 4D dynamics, we leverage the co-generated dense optical flow $\mathbf{f}_{opt} = [\Delta u, \Delta v]^\top$ to establish temporal correspondences. A 3D point $\mathbf{P}_t$ is tracked to its position at $t+1$ by:

$$\mathbf{P}_{t+1} = D_{t+1}(u+\Delta u, v+\Delta v)\cdot \mathbf{K}^{-1}(\mathbf{p}_t + [\Delta u, \Delta v, 0]^\top). \tag{10}$$

The *3D scene flow* is then defined as $\mathbf{f}_{3D} = \mathbf{P}_{t+1} - \mathbf{P}_t$, representing the per-point metric displacement. This explicit 4D mapping ensures that the generated trajectories are not merely visual hallucinations but correspond to physically plausible 3D movements.

Refinement and Visualization. To suppress artifacts at depth discontinuities, we apply a depth-gradient-based edge filter, masking out pixels where $\|\nabla D\| > \tau$. The resulting refined 3D trajectories are projected into a canonical bird's-eye view (BEV). In our qualitative analysis (see the "3D Flow" panel in Fig. 1), these trajectories are rendered as colored trails over a depth-ordered point cloud backdrop, providing an intuitive verification of the model's spatial-temporal coherence.

G World-4D-200M Dataset

The pipeline for our multimodal annotation is illustrated in Fig. 8.

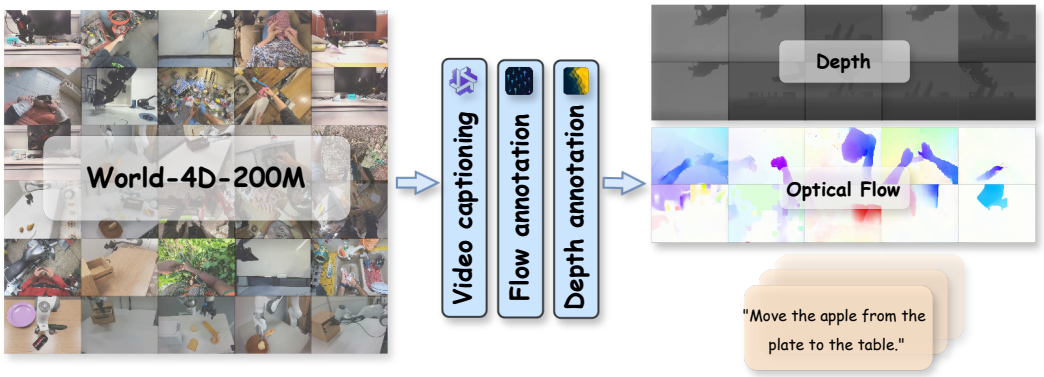


Figure 8: **Data Curation Pipeline.** The video data is collected from diverse sources and partitioned into short clips during data preprocessing. Each clip undergoes a multi-modal annotation process: (1) **Video Captioning**: Qwen3-VL (Bai et al., 2025) generates detailed natural language descriptions of the video content; (2) **Optical Flow Estimation**: DPFlow (Morimitsu et al., 2025) computes dense per-frame motion fields, which are encoded as color-coded flow videos for training; (3) **Depth Estimation**: Depth Anything 3 (Lin et al., 2025) produces monocular depth predictions.

Video Captioning. We use Qwen3-VL (Bai et al., 2025) to generate captions for the video data (Damen et al., 2020; Wang et al., 2024; Wu et al., 2024; Liu et al., 2024; Jiang et al., 2025b). Specifically, we leverage the model's strong video-language understanding capabilities to produce detailed, structured descriptions of each video clip. The videos are first sampled at a frame rate of 1 FPS and split into segments of 5 seconds. For each segment, we provide the following prompt to the model:

```
Please describe this video in detail. Include the
following aspects:
1. The main subject and action in the video.
2. The environment and background.
3. Any objects and their interactions.
4. The overall scene context and atmosphere.
Provide a concise but comprehensive caption in one
paragraph.
```

The generation is performed with a maximum output length of 512 tokens and a temperature of 0.7 to balance creativity and coherence. The generated captions are then collected and stored in JSON format for downstream tasks.

Optical Flow Annotation. We employ DPFlow (Morimitsu et al., 2025), a state-of-the-art optical flow estimation model. For each video, frame pairs are processed sequentially at native resolution. The estimated flow fields are encoded into color-coded RGB videos using the Middlebury color-wheel convention (see Appendix K.1 for details) and stored at 25 FPS. This color-coded representation is used as the training target for the flow branch, enabling it to reuse the pretrained RGB video VAE. At evaluation time, the generated color-coded flow videos are decoded back to flow vectors before computing endpoint error metrics (Appendix B.3).

Depth Annotation. We employ Depth Anything 3 (Lin et al., 2025), specifically the `DA3NESTED-GIANT-LARGE-1.1` checkpoint, which provides dense per-frame depth predictions along with camera pose estimation. Each video is sampled at 30 FPS and resized such that its shorter side is 392 pixels.

The statistics and composition of World-4D-200M are visualized in Fig. 9. To convert the estimated depth maps into viewable depth videos, we load the compressed depth arrays and upsample each frame to the original video resolution using bilinear interpolation. Depth values are clipped to a global range of $[0.0, 5.0]$ meters and quantized to 8-bit grayscale via $I = \lfloor d/d_{\max} \times 255 \rfloor$. The resulting frames are saved as RGB videos.

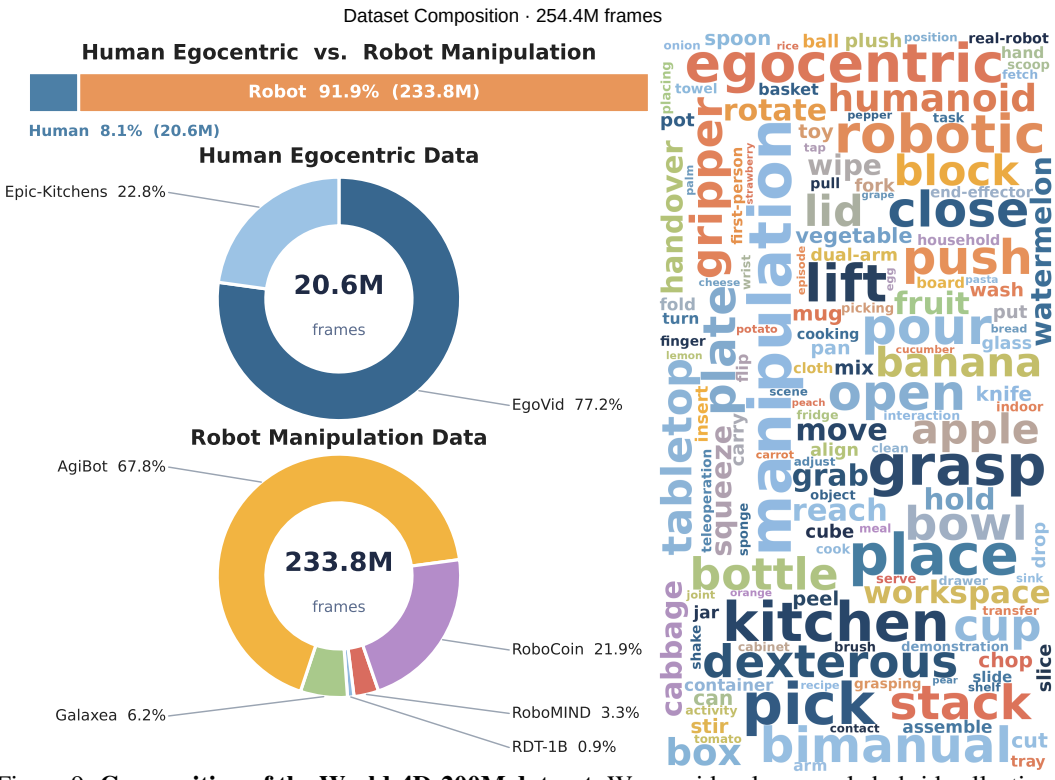


Figure 9: **Composition of the World-4D-200M dataset.** We provide a large-scale hybrid collection of 254.4M frames, balancing human egocentric videos with diverse robotic manipulation data. This diversity ensures that the world model learns both general object interaction priors and robot-specific execution traces.

Table 3: Inference latency breakdown on NVIDIA RTX 5090 (FP8).

Phase	Latency (ms)	Percentage
Depth Estimation (DA3 Lin et al. (2025))	85	7.7%
VAE Encoding & Latent Prep	18	1.6%
World-4D	990	89.5%
Feature Reshape & Concat	1	0.1%
Flow Former	4	0.4%
Action Flow Matching Head	8	0.7%
Total Evaluation (Forward)	**1,106**	**100%**

H Inference Latency and Real-time Control

To evaluate the feasibility of World-4D-Policy in real-world scenarios, we conduct a detailed timing analysis of our inference pipeline. The model is deployed on a workstation equipped with an NVIDIA RTX 5090 GPU, leveraging FP8 quantization and FlashAttention 3 (FA3) kernels to accelerate the transformer-based 4D generation.

Note that the 4D visual features are extracted from the frozen World-4D in a single forward pass ($N = 1$). The subsequent 4-step ODE sampling for action generation occurs only within the lightweight World-4D-Policy head.

Latency Breakdown. The overall control frequency is determined by the total cycle time of the World-4D forward pass and the action generation head. Given a sequence of $K = 10$ actions generated per forward pass, a control frequency of approximately 9 Hz is achieved with a cycle time of $\sim$1.1 s. Tab. 3 provides a granular breakdown of the time spent in each phase.

The primary bottleneck is the tri-branch Transformer, which accounts for 89.5% of the total latency.

Control Frequency. It is important to distinguish between the *planning frequency* (the rate at which the 4D world model refreshes its mental state) and the *effective control frequency*. Although a single forward pass takes $\sim$1.1 s (yielding a planning frequency of $\approx$0.9 Hz), the World-4D-Policy employs **action chunking** by predicting $K = 10$ future steps in a single inference. As these 10 actions are executed sequentially while the next planning cycle is computed in parallel, the system achieves an effective control frequency of $\approx$9 Hz.

Closed-loop Robustness. While 9 Hz is lower than traditional low-level PID controllers (typically $>$500 Hz), World-4D-Policy maintains high robustness through two mechanisms:

Instead of a single action, the policy outputs a sequence of $K = 10$ future actions. During the $\sim$1.1 s inference cycle, the robot executes the previously planned action chunk. The $K = 10$ actions correspond to an effective action rate of approximately 9 Hz, while low-level interpolation is performed at 500 Hz.

Unlike 2D policies that suffer from visual aliasing or depth ambiguity, our policy's internal latents are grounded in 3D scene flow. As shown in our ablation, the inclusion of explicit kinetic cues allows the policy to predict object movements. This anticipation compensates for the slight sensing-to-actuation lag, as the model is not just reacting to the current frame but is conditioned on a predicted 4D trajectory.

In real-world tests, we observe that even when objects are slightly bumped during the 1s execution window, the next 9 Hz update effectively re-plans the trajectory because the World-4D latents encompass a spatial volume rather than just a pixel point, providing a wider capture range for recovery.

I Robot Tasks

To demonstrate the generalization ability of our pipeline across diverse manipulation challenges, we evaluate our method on six distinct tasks that span varying levels of bimanual coordination, contact richness, and precision:

- **(1)** ***Dual Picking***: The robot uses its left arm to pick an apple and its right arm to pick a banana from a plate, sequentially placing both objects onto the tabletop.
- **(2)** ***Block Pushing***: A sequential pushing task where the left arm pushes a large block from the left zone to the center, after which the right arm takes over to push it to the designated right target zone.
- **(3)** ***Hand-over***: An intra-robot transfer task where the left gripper picks up a cabbage and hands it over to the right gripper, which then completes the placement.
- **(4)** ***Bimanual Lifting***: A heavy-load coordination task where both arms must synchronously lift a watermelon plush from the table and accurately place it into a plate.
- **(5)** ***Lid Placement***: A precision-oriented task requiring the robot to pick up a lid and accurately align it to cover a cardboard box.
- **(6)** ***Bowl Stacking***: There are two small bowls on the table; the task involves picking up one bowl and carefully stacking it on top of the other.

As shown in Fig. 10, these tasks collectively challenge the model's proficiency in dual-arm synergy, temporal sequencing, and long-horizon interaction. Fig. 11 further presents World-4D's synchronized RGB, depth, and optical-flow predictions on these real-robot tasks. Tab. 4 summarizes the per-task demonstration budget: the full corpus is used to train World-4D, exposing it to a rich diversity of physical interactions, while a fixed budget of 200 episodes per task is allocated for training World-4D-Policy. For each task, we evaluate the model's generalization by applying significant randomization to the initial environment state. This includes varying the 6-DoF poses of task-relevant objects (*e.g.*, fruits) in terms of both workspace coordinates and axial rotations.

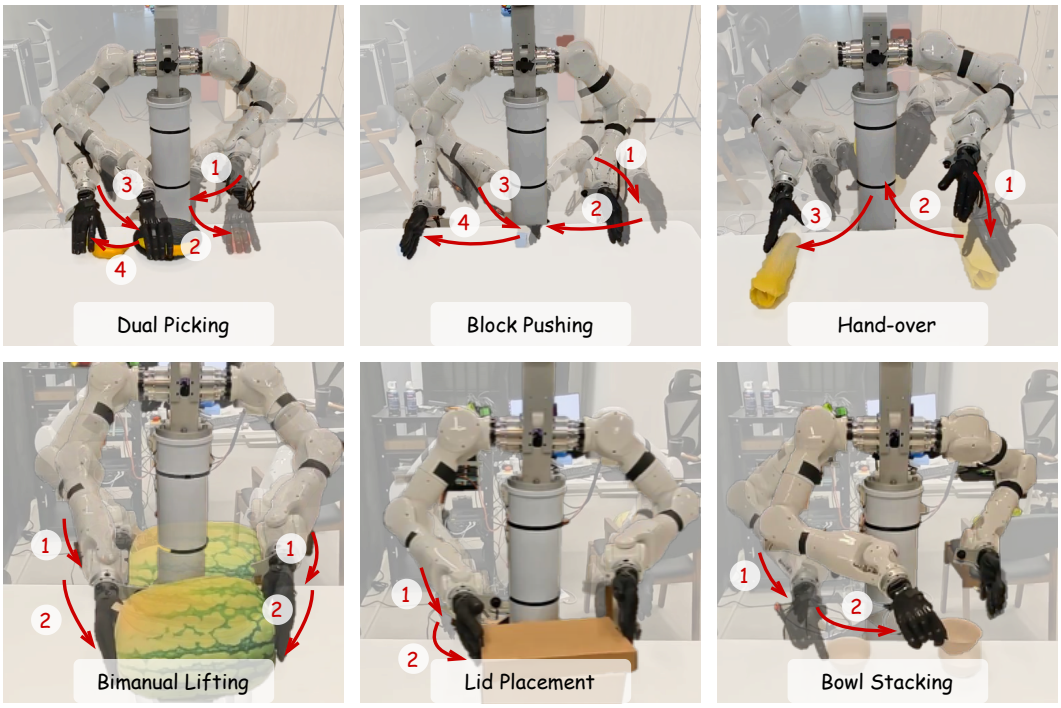


Figure 10: **Real-world Manipulation Benchmark.** We establish a comprehensive evaluation suite comprising six diverse tasks to assess the model's performance in open-world manipulation, providing a rigorous testbed for our 4D world model.

Table 4: Statistics of the Real-World Dataset used for training.

Task Name	Description	World-4D	World-4D-Policy
Dual Picking	Pick apple and banana from plate to table.	500	200
Block Pushing	Push block from left to center, then to the right.	500	200
Hand-over	Hand-over cabbage from left to right hand and place.	300	200
Bimanual Lifting	Lift a watermelon plush and place it into a plate.	500	200
Lid Placement	Pick the lid and place it precisely on top of the box.	300	200
Bowl Stacking	Stack one bowl onto another.	300	200
Total		**2,400**	**1,200**

J Implementation Details

To ensure semantic alignment while minimizing overhead, we share the text cross-attention Key/-Value projections across all three branches, as the linguistic task description provides a modality-agnostic semantic signal. We insert Joint Cross-Modal Attention (JA) modules every three transformer blocks across all 30 layers (at layers 0, 3, 6, ..., 27), yielding 10 modules in total.

We now give the full JA formulation condensed in Sec. 3.2. Before cross-modal mixing, each branch $m \in \{\text{rgb}, \text{depth}, \text{flow}\}$ receives a learnable modality embedding $\boldsymbol{e}^m \in \mathbb{R}^{1\times1\times d}$ (zero-initialized so the module starts as a pure residual) and is normalized by a per-modality LayerNorm LN^m that aligns numerical scales across branches:

$$\tilde{\boldsymbol{z}}_l^m = \text{LN}^m\big(\boldsymbol{z}_l^m + \boldsymbol{e}^m\big). \tag{11}$$

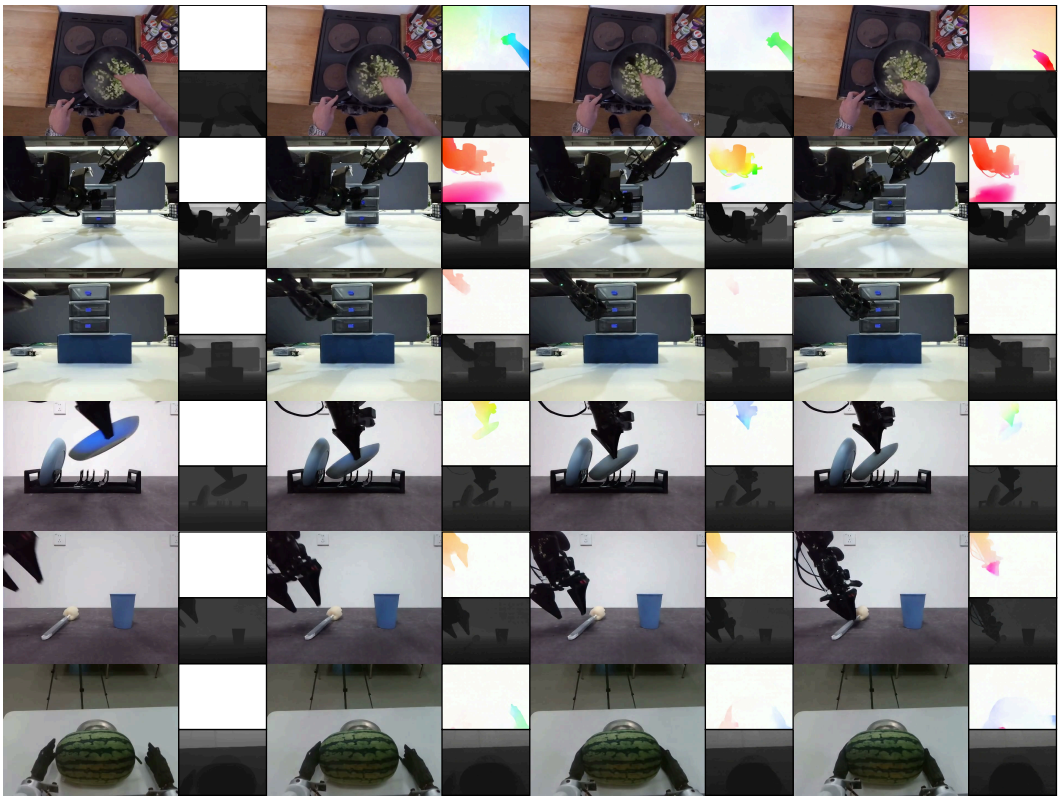


Figure 11: **Qualitative 4D predictions on our real-robot tasks.** Given a single RGB-D observation and the task instruction, World-4D generates synchronized RGB, depth, and optical-flow sequences on the six evaluated manipulation tasks, providing the predictive 4D representations consumed by World-4D-Policy.

Each branch then produces *one* query and *one* shared key/value pair reused by all other branches' queries, with RMSNorm applied to queries and keys for stability:

$$\boldsymbol{Q}_l^m = \mathrm{RMSNorm}_q\big(\mathrm{QProj}_l^m(\tilde{\boldsymbol{z}}_l^m)\big), \quad [\boldsymbol{K}_l^m, \boldsymbol{V}_l^m] = \mathrm{KVProj}_l^m(\tilde{\boldsymbol{z}}_l^m), \quad \boldsymbol{K}_l^m \leftarrow \mathrm{RMSNorm}_k(\boldsymbol{K}_l^m). \tag{12}$$

The cross-modal update then follows the gated residual form of the main text:

$$\hat{\boldsymbol{z}}_l^m = \boldsymbol{z}_l^m + \tanh(g_l^m) \cdot \mathrm{CrossBranchAttn}(\boldsymbol{Q}_l^m, \boldsymbol{K}_l^{\mathrm{cross}}, \boldsymbol{V}_l^{\mathrm{cross}}), \tag{13}$$

where g_l^m is a learnable scalar gate initialized to 1. To ensure a smooth transition from independent branch training, we initialize the output projection to zero while keeping $g_l^m = 1$. JA employs 3D RoPE and a frame-wise mask to restrict attention to tokens within the same temporal frame.

J.1 Phased Training Strategy

We adopt a three-stage curriculum to bridge modality distribution gaps. Tab. 5 summarizes the stage-wise configuration. Each stage is initialized from the *model-only* checkpoint of the previous stage (optimizer/scheduler reset) to ensure stability. We utilize the AdamW optimizer (β_1=0.9, β_2=0.95, weight decay 1×10^{-4}) with cosine scheduling and linear warm-up. We track an Exponential Moving Average (EMA, decay 0.9999) with shadow weights on CPU for inference.

Stage 1: Modality Adaptation. Branches are trained independently under modality-specific flow-matching supervision to repurpose RGB priors for geometric and kinetic modeling without gradient interference. We use a learning rate of 2×10^{-5} with 500 warm-up steps and flow weight λ_{flow}=0.5.

Stage 2: Frozen-Backbone Joint Attention. We freeze the backbone and instantiate the 10 JA modules. To preserve established representations, we employ **zero-initialization** for the output projections of the JA modules. The learning rate is set to 5×10^{-5} with 200 warm-up steps.

Table 5: Stage-wise training configuration. Effective batch size is per-GPU batch (1) $\times$ gradient accumulation (2-4) $\times$ N_{GPU}.

	Stage 1	**Stage 2**	**Stage 3**
Fusion mode	none	joint (frozen bb.)	joint (full SFT)
Trainable params	all branches	JA + mod. embed.	all parameters
Learning rate	2×10^{-5}	5×10^{-5}	1×10^{-5}
LR warm-up steps	500	200	500
λ_{flow}	0.5	1.0	1.0
Branch Dropout	—	0.2	0.1

Stage 3: Full-Parameter Joint SFT. We unfreeze the entire model for joint fine-tuning on the World-4D-200M. We employ the learning rate as 1×10^{-5} for all trainable parameters.

Branch Dropout. In Stages 2 and 3 we apply Branch Dropout to enhance cross-modal robustness: with probability p_{drop} we randomly select one of $\{\text{depth}, \text{flow}\}$ at each training step and replace its noisy latent (frames $[1{:}]$) with pure Gaussian noise, forcing the JA modules to reconstruct the missing modality from the visible ones. The RGB branch is never dropped, since it serves as the appearance anchor: destroying it would leave the joint module with no consistent reference. We use p_{drop}=0.2 in Stage 2 and reduce it to p_{drop}=0.1 in Stage 3.

J.2 Resource and Optimization

All stages train at $81\times480\times640$ (yielding $T = 21$ latent frames under the causal VAE's $4\times$ temporal compression, i.e., $T_{\text{latent}} = (T_{\text{pixel}}-1)/4+1$) with `bf16` mixed precision and gradient checkpointing. Stages 2 and 3 leverage DeepSpeed ZeRO-2 with optimizer offload to manage memory for additional JA parameters.

J.3 World-4D-Policy

We utilize World-4D as a frozen 4D vision encoder. The model operates at 480×640 resolution, producing $T = 21$ latent frames (after VAE temporal compression with a ratio of 4). At each inference step, we perform single-step feature extraction at normalized flow-matching time $\tau = 0.5$, capturing intermediate hidden states from block 15 of the transformer. By concatenating the 3072-dimensional features from the RGB, depth, and optical flow branches, we obtain a comprehensive 4D representation F_p.

At each decision step, we take a single RGB observation frame as input and estimate its depth map using a monocular depth estimator. The RGB-D frame and estimated depth map are then fed into the frozen World-4D encoder. The condition frame is center-cropped to 480×640 and normalized to $[-1, 1]$. The extracted spatiotemporal features are reshaped and fed into Flow Former. This compresses the high-dimensional World-4D features into a fixed-size representation suitable for policy decoding.

We employ a flow matching policy head with 4-step Euler ODE sampling at inference time. Despite the multi-step sampling in the action space, the policy maintains high efficiency because the heavy 4D backbone features are only computed once. The policy predicts action chunks of length 10, where each action is 54-dimensional. During deployment, the model executes 10 actions open-loop before re-querying the visual backbone, yielding an effective control frequency of $\sim$9 Hz.

We train with AdamW optimizer using a learning rate of 1×10^{-4}, $\beta = (0.9, 0.9)$, and weight decay 0.05. We employ a tri-stage learning rate schedule: 2% linear warmup, 8% constant hold, and 90% cosine decay to $10^{-6}\times$ the peak learning rate. The World-4D backbone is kept frozen throughout training; only the Flow Former and flow matching policy head are optimized. Training uses mixed precision with a batch size of 1 per GPU for 100 epochs.

Table 6: Results on the LARY benchmark (Nie et al., 2026). Representation probing results for low-level control regression (MSE, lower is better).

Modality	Avg.	CALVIN	VLABench	RoboCOIN	AgiBot-B
RGB	0.28	0.35	0.08	0.18	0.52
RGB-D	0.26	0.30	0.07	0.15	0.50
RGB-DF	**0.22**	**0.25**	**0.06**	**0.10**	**0.47**

K Representation Probing

To examine whether RGB, depth, and optical flow (RGB-DF) is closer to the robot action space than RGB or RGB-D, we conduct a representation probing experiment following the low-level control regression protocol of LARYBench Nie et al. (2026), on its four regression datasets—CALVIN, VLABench, RoboCOIN, and AgiBot-Beta—which span simulated and real-robot manipulation as well as single-arm and bimanual embodiments. Each dataset uses the benchmark's native temporal stride (5 for CALVIN and VLABench, 10 for RoboCOIN, 45 for AgiBot-Beta): a sample is a frame pair $(\mathbf{I}_t, \mathbf{I}_{t+s})$ together with the ground-truth action chunk over the interval, in the benchmark's native action space (7-dim per-step end-effector control for CALVIN and VLABench, 12-dim bimanual for RoboCOIN, and 16-dim bimanual end-effector poses with grippers for AgiBot-Beta).

The three modalities are RGB $[\mathbf{I}_t, \mathbf{I}_{t+s}]$ (6 channels), RGB-D $[\mathbf{I}_t, \mathbf{D}_t, \mathbf{I}_{t+s}, \mathbf{D}_{t+s}]$ (8 channels), and RGB-DF, which further appends the dense optical flow $\mathbf{F}_{t\to t+s}$ estimated by DPFlow (Morimitsu et al., 2025) (10 channels). Inputs are resized to 224×224 and fed into an ImageNet-pretrained ResNet-18 whose first convolution is widened to the corresponding channel count; the backbone is frozen and only the input convolution and the MLP action expert are trained with MSE loss on per-dimension z-normalized targets (statistics from the training split). We report the best validation MSE over 30 epochs; under this normalization a mean predictor scores 1.00. All modalities are probed on identical frame pairs and splits, so differences reflect only the action-relevant information carried by the added channels.

K.1 Optical-flow Colorization and Decoding

Optical flow is a dense 2D vector field and does not have a natural RGB appearance. In our pipeline, color-coded flow videos serve as the training representation for the flow branch, enabling it to reuse the pretrained RGB video VAE and standard video codecs. This representation is also convenient for visualization and inspection. However, color space is not a metric space for optical-flow accuracy. Therefore, all quantitative flow metrics in this work are computed in flow-vector space after decoding the color-coded representations back to flow vectors.

For each pixel p, the flow vector is

$$\mathbf{f}_p = (u_p, v_p), \tag{14}$$

where u_p and v_p denote horizontal and vertical displacements. Since flow vectors do not have a natural RGB appearance, we use a Middlebury-style color-wheel visualization to convert flow fields into RGB videos.

Why color-coded flow videos? The color-coded representation is used for two practical reasons. First, it provides an intuitive visualization: hue encodes flow direction, while brightness or saturation encodes flow magnitude. Second, it allows flow sequences to be stored and processed using standard RGB video pipelines, which is convenient for video codecs, visualization, and pretrained video models. However, this representation is not a metric space for optical-flow accuracy. Therefore, quantitative flow evaluation should be performed in flow-vector space rather than in RGB color space.

Forward colorization. Given a flow vector $\mathbf{f}_p = (u_p, v_p)$, we compute its direction and magnitude:

$$\theta_p = \operatorname{atan2}(v_p, u_p), \qquad r_p = \sqrt{u_p^2 + v_p^2}. \tag{15}$$

The direction θ_p is mapped to hue, and the normalized magnitude is mapped to value or saturation. A common formulation is

$$H_p = \frac{\theta_p + \pi}{2\pi}, \qquad S_p = 1, \qquad V_p = \min\left(\frac{r_p}{r_{\max}}, 1\right), \tag{16}$$

where $r_{\max}$ is a normalization constant, such as the maximum flow magnitude in the clip or a robust percentile, e.g., the 95th percentile. The RGB color is then obtained by HSV-to-RGB conversion:

$$I_p = \mathrm{HSV2RGB}(H_p, S_p, V_p). \tag{17}$$

Invalid or occluded pixels are typically displayed as black.

Inverse decoding. In principle, if the colorization parameters are known, a color-coded flow image can be approximately decoded back to flow vectors. Given an RGB pixel, we first convert it to HSV. Assuming hue encodes direction and value encodes normalized magnitude, the flow direction and magnitude can be recovered as

$$\theta_p = 2\pi H_p - \pi, \qquad r_p = V_p\, r_{\max}. \tag{18}$$

The flow vector is then

$$u_p = r_p \cos\theta_p, \qquad v_p = r_p \sin\theta_p. \tag{19}$$

In practice, this decoding is approximate because video compression, color quantization, clipping, and unknown normalization can introduce errors. In practice, this decoding is approximate because video compression, color quantization, clipping, and unknown normalization can introduce errors. We therefore use the color-coded representation consistently throughout training and generation, and decode back to flow vectors only at evaluation time for metric computation.

Metric implication. Color-space distances between color-coded flow visualizations do not correspond to pixel displacement error and are therefore not used as evaluation metrics. All endpoint-error metrics in this work are computed in flow-vector space using the definitions in Appendix B.3.