BEYOND SUCCESSFUL ROBOT DEMONSTRATIONS: LEARNING FROM HUMAN INTERACTIONS AND MODELING FAILURE DYNAMICS

Anonymous authors
Paper under double-blind review

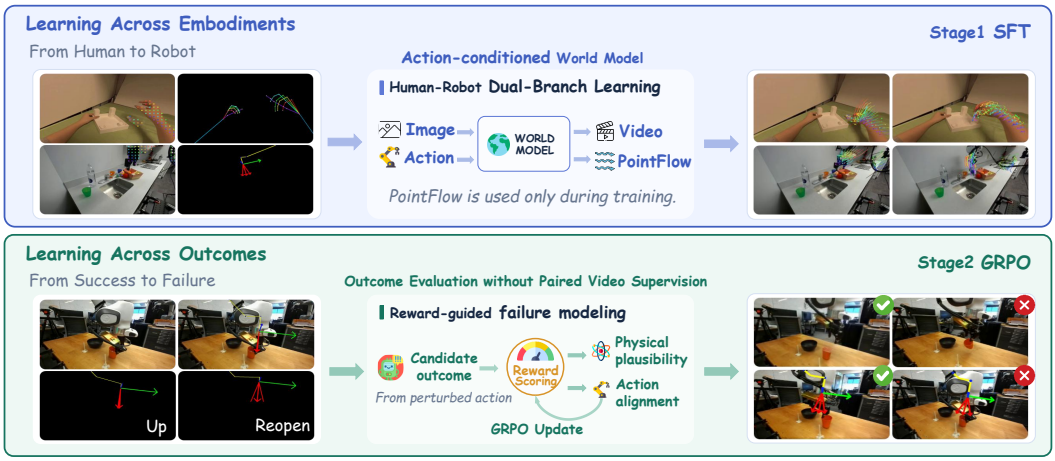


Figure 1: **Overview of HF-WM architecture.** HF-WM extends learning beyond successful robot demonstrations in two stages. Stage 1 uses a dual-branch architecture with stabilized pointflow supervision to learn interaction dynamics from egocentric human interactions (Sec. 3.2). Stage 2 applies GRPO to counterfactual action perturbations, guided by action alignment and physical plausibility rewards, to model failure dynamics without paired failure-video supervision (Sec. 3.3).

ABSTRACT

Action-conditioned world models offer a promising foundation for generalist robotic intelligence by predicting the consequences of actions. However, existing robotic datasets provide limited coverage of the diverse interaction dynamics required for generalizable world modeling. Large-scale egocentric human videos offer complementary interaction experience, yet the embodiment gap between human hands and robotic manipulators hinders effective cross-embodiment learning. Meanwhile, the predominance of successful demonstrations leaves failure dynamics substantially underrepresented, limiting the ability of world models to anticipate unsuccessful executions. To address these limitations, we introduce **HF-WM**, an action-conditioned world model that combines **H**uman interaction learning with **F**ailure-aware post-training. To bridge the embodiment gap, we develop a dual-branch framework that couples video generation with auxiliary pointflow prediction, using stabilized pointflow as a unified motion representation across human and robot interactions while suppressing egocentric camera motion. To improve failure modeling, we further perform Group Relative Policy Optimization (GRPO) post-training on perturbed actions derived from successful demonstrations, guided by action alignment and physical plausibility rewards, enabling the model to learn plausible failure dynamics without paired ground-truth videos. Extensive experiments demonstrate that HF-WM achieves state-of-the-art performance in action-conditioned video generation and exhibits strong utility for policy evaluation in downstream robotic tasks. The anonymous project page: `https://anonymous.4open.science/w/HF-WM/`

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

1 INTRODUCTION

With the rapid advancement of video generation models (Wan et al., 2025; Agarwal et al., 2026), world models have witnessed significant breakthroughs and are increasingly regarded as a promising pathway toward generalist robotics (Ye et al., 2026; Yuan et al., 2026). In particular, action-conditioned world models take robot action sequences as input and model the resulting state transitions to predict future states (Zhu et al., 2025). This predictive capability supports scalable data generation, policy evaluation, and reinforcement learning in generated environments (Sun et al., 2026; Ma et al., 2026; Tseng et al., 2026), thereby closing the loop between imagination, evaluation, and policy improvement. This paradigm has emerged as a key foundation for learning and improving robot action planning systems.

Despite these advances, action-conditioned world models still struggle to accurately capture interaction dynamics and execution outcomes, largely due to the limited scale and coverage of high-quality embodied data (Mei et al., 2026). To alleviate this limitation, recent studies have turned to large scale egocentric human object interaction datasets (Gao et al., 2026; Wu & Gao, 2026), leveraging human demonstrations to enhance model performance. **1)** However, a substantial domain gap persists between human and robotic manipulation, and egocentric recordings are often affected by camera jitter and unstable viewpoints, which can obscure interaction-relevant regions and hinder the learning of physically meaningful representations. Consequently, effectively exploiting human egocentric data for embodied world model learning remains an open challenge (Mei et al., 2026). **2)** In addition, existing embodied datasets are heavily biased toward successful demonstrations, with limited coverage of failure cases and off-target interactions (Yang et al., 2026a). Yet world models for embodied scenarios require not only modeling successful trajectories, but also capturing failure dynamics and counterfactual outcomes induced by erroneous or suboptimal actions (Ma et al., 2026). In the absence of failure demonstrations, how to effectively synthesize failure scenarios to improve physical reasoning remains an important open problem.

To this end, as shown in Fig. 1, we propose a set of key insights to address the aforementioned challenges. **1)** Firstly, we introduce a skeleton grounded action representation that bridges human and robotic action spaces through a shared spatial formulation (Wu & Gao, 2026), mitigating the cross embodiment gap. To better exploit egocentric human interaction data, we introduce pointflow supervision as a motion constraint that tracks point trajectories, capturing interaction driven dynamics (Huang et al., 2026; Lee et al., 2026). We further apply a stabilization procedure on pointflows to suppress camera induced motion, ensuring that point trajectories primarily reflect interaction driven dynamics rather than viewpoint variations. This allows the model to better focus on physically meaningful motion signals and improves the effective utilization of egocentric human interaction data. The primary challenge in introducing failure cases is the lack of real failure video supervision, which limits direct learning under failure scenarios. **2)** To address this, we propose a Group Relative Policy Optimization (GRPO) based reinforcement learning framework (Liu et al., 2026) for post-training, enabling better modeling of failure dynamics using reward signals without requiring ground-truth failure video supervision. Specifically, we construct pseudo failure actions by perturbing successful action trajectories, thereby inducing counterfactual action conditions. These perturbed actions are then used as conditions for video generation, where the model is optimized via reward-based feedback under GRPO. The reward function consists of two complementary terms: a skeleton-based action following score that measures the consistency between the generated motion and the given action condition (Chen et al., 2026), and a physical plausibility score that evaluates whether the predicted dynamics remain physically consistent. Both rewards are computable without paired ground-truth videos, enabling GRPO post-training on perturbed actions to strengthen failure-aware dynamics modeling without direct supervision of the corresponding outcomes.

In summary, our main contributions are:

- We present HF-WM, an action-conditioned world model that learns from egocentric human interactions and counterfactual action perturbations for embodied dynamics modeling.
- We introduce a dual-branch framework with auxiliary pointflow supervision, establishing a shared motion representation across human and robot interactions that enables embodied world models to learn more effectively from egocentric human data.
- We present a GRPO based post-training framework that enables failure aware optimization without paired failure video supervision.

- Extensive experiments demonstrate that HF-WM achieves state-of-the-art performance in action-conditioned video generation, and further validate its effectiveness as a policy evaluator for downstream robotic tasks.

2 Related Work

Controllable video generation. Controllable video generation as an important extension of video foundation models, aims to introduce diverse external control signals beyond text conditioning to enable fine-grained control over generated content. Recent work on controllable generation incorporates a variety of conditioning signals into existing video foundation models, including spatial structural information (e.g., human poses (Hao et al., 2026), depth maps (Zhu et al., 2026), and skeletons (Wu & Gao, 2026)), temporal dynamics (e.g., optical flow (Wang et al., 2026), motion trajectories (Zhu et al., 2025), and camera parameters (Tu et al., 2026)), as well as multimodal inputs such as reference images or audio. These methods typically inject external control signals into pretrained video diffusion models through dedicated condition encoders (Zhen et al., 2025), cross-attention mechanisms (YU et al., 2025), or ControlNet architectures (Yang et al., 2026b), thereby constraining and guiding the generation process.

Action-conditioned world models. Recent advances have positioned action-conditioned world models as a key interface between generative video modeling, action-conditioned dynamics prediction, and policy evaluation in embodied intelligence (Chen et al., 2026). IRASim (Zhu et al., 2025) incorporates robotic trajectories into the video generation process, demonstrating the feasibility of synthesizing future visual states under explicit action conditions. Ctrl-World (Guo et al., 2025) extends this paradigm to multi-view embodied generation, where three-view synthesis improves the spatial coverage of generated environments and yields practical gains in real-robot deployment. SC3-Eval (Tseng et al., 2026) further closes the loop by leveraging world models as offline policy evaluators, enabling policies to be assessed without costly physical rollouts. Despite recent progress, existing action-conditioned world models remain constrained by limited embodied data and failure coverage (Gao et al., 2026; Wu & Gao, 2026), motivating us to study how human object interaction data and counterfactual failure samples can improve the modeling of action consequences and physical dynamics.

3 Method

3.1 Preliminaries

Mixture-of-Transformers Foundation Model. We build on Cosmos 3 (Agarwal et al., 2026), an omnimodal world model that jointly models language, video, and actions using a Mixture-of-Transformers (MOT) architecture. It factorizes multimodal sequences into an autoregressive stream for text and ViT encoded visual tokens, and a diffusion stream for VAE encoded video latents and action tokens. Shared multimodal attention couples the two streams: autoregressive tokens use causal attention, whereas diffusion tokens attend bidirectionally to the autoregressive context and diffusion states. For visual generation, the Wan2.2 VAE (Wan et al., 2025) encodes videos into spatio-temporal latents. Given a clean latent $\mathbf{x}_0$, Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and a flow time $t \in [0, 1]$, the interpolated latent is

$$\mathbf{x}_t = (1 - t)\mathbf{x}_0 + t\boldsymbol{\epsilon}. \tag{1}$$

The generator is trained with the rectified-flow matching objective

$$\mathcal{L}_{\mathrm{RF}} = \mathbb{E}_{\mathbf{c},t,\mathbf{x}_0,\boldsymbol{\epsilon}} \left[\left\| v_\theta(\mathbf{x}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{x}_0) \right\|_2^2 \right], \tag{2}$$

where v_θ is the predicted velocity field and $\mathbf{c}$ denotes the conditioning context, including text, visual observations, and action tokens. In our setting, clean action tokens condition the denoising of future visual latents, enabling prediction of the resulting state transitions.

Flow-Based Group Relative Policy Optimization. We formulate reward-guided rectified-flow denoising as a sequential decision process. Given a conditioning input $\mathbf{c}$, the policy π_θ is induced by the velocity field $v_\theta(\mathbf{x}_t, t, \mathbf{c})$. Following Flow-GRPO (Liu et al., 2026), we replace the deterministic

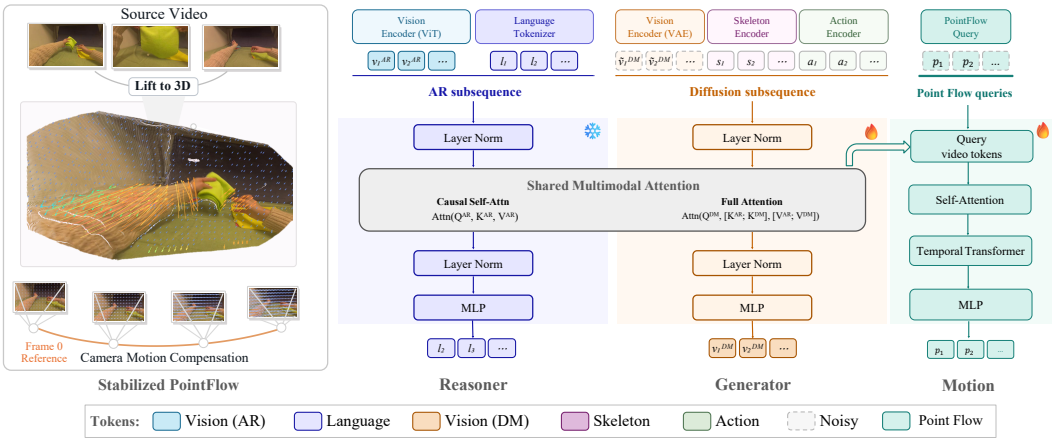


Figure 2: **Dual-branch architecture of HF-WM with stabilized pointflow.** The auxiliary pointflow branch predicts interaction motion from intermediate video representations, while pointflow stabilization suppresses egocentric camera motion so that supervision is better focused on interaction-induced dynamics. The pointflow branch is used only during training and introduces no additional inference overhead.

rectified-flow ODE with the marginal-preserving SDE

$$d\mathbf{x}_t = \left[v_\theta(\mathbf{x}_t, t, \mathbf{c}) + \frac{\sigma_t^2}{2t} \left(\mathbf{x}_t + (1-t) v_\theta(\mathbf{x}_t, t, \mathbf{c}) \right) \right] dt + \sigma_t \, d\mathbf{w}_t, \qquad t \in (0, 1], \tag{3}$$

where σ_t controls the exploration strength and $d\mathbf{w}_t$ is a Wiener increment.

After discretizing the denoising process into N steps, we sample K trajectories $\tau^k = (\mathbf{x}_N^k, \ldots, \mathbf{x}_0^k)$ from the old policy. At each step $n = 1, \ldots, N$, the transition ratio and its clipped version are

$$\rho_n^k(\theta) = \frac{\pi_\theta(\mathbf{x}_{n-1}^k \mid \mathbf{x}_n^k, \mathbf{c})}{\pi_{\theta_{\text{old}}}(\mathbf{x}_{n-1}^k \mid \mathbf{x}_n^k, \mathbf{c})}, \qquad \bar{\rho}_n^k(\theta) = \text{clip}\big(\rho_n^k(\theta), 1-\varepsilon, 1+\varepsilon\big). \tag{4}$$

The terminal reward is normalized within each group:

$$A^k = \frac{r(\mathbf{x}_0^k, \mathbf{c}) - \mu_r}{\sigma_r + \delta}, \qquad (\mu_r, \sigma_r) = \text{MeanStd}_{j=1}^K \left\{ r(\mathbf{x}_0^j, \mathbf{c}) \right\}, \tag{5}$$

where $\delta > 0$ is a small numerical stabilizer. The resulting GRPO objective is

$$\mathcal{J}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \frac{1}{N} \sum_{n=1}^{N} \left[\min\big(\rho_n^k(\theta) A^k, \bar{\rho}_n^k(\theta) A^k\big) - \beta D_{\text{KL}}\big(\pi_\theta(\cdot \mid \mathbf{x}_n^k, \mathbf{c}) \,\big\|\, \pi_{\text{ref}}(\cdot \mid \mathbf{x}_n^k, \mathbf{c})\big) \right]. \tag{6}$$

This formulation allows the generator to be optimized with pixel-level rewards from generated samples, without paired video ground-truth; group-relative feedback further improves generation quality and alignment.

3.2 Learning Across Embodiments: From Human to Robot

Introducing human egocentric interaction data into action-conditioned world models first requires addressing the representational mismatch between human hand actions and robot arm actions. These two action forms differ naturally in degrees of freedom, dimensionality, and motion structure, making it difficult to share a common action conditioning space directly. Thanks to the Cosmos 3 foundation model (Agarwal et al., 2026), which has been pretrained on both human interaction data and embodied manipulation data, the model already possesses certain cross embodiment interaction priors, thereby partially reducing the domain gap between human hands and robot arms. Building on this prior, we adopt the 2D skeleton representation from OSCAR (Wu & Gao, 2026), mapping human hand and robot arm actions into a shared spatial action space, thereby avoiding direct alignment

between heterogeneous action vectors. With this unified action representation, human egocentric interaction data can be incorporated as effective augmentation data for training action-conditioned world models. Furthermore, we develop a dual-branch learning framework that couples video generation with auxiliary pointflow prediction (Huang et al., 2026), where pointflow supervision guides the video branch to focus on motion regions that are truly relevant to physical interactions, enabling human interaction data to provide more effective supervision for embodied world model learning.

For the auxiliary pointflow branch, we introduce a stabilized pointflow representation to mitigate perturbations caused by egocentric camera motion in human object interaction data, thereby encouraging the model to focus on interaction induced dynamics. We first use CoWTracker (Lai et al., 2026) to obtain 2D point trajectories $\mathbf{p}_{i,t} = (u_{i,t}, v_{i,t})$, where i and t index points and frames, and (u, v) denote image coordinates. To disentangle camera motion from the tracked trajectories, we use monocular depth estimated with Depth Anything 3 (Lin et al., 2025) together with the camera parameters provided by the dataset. For each tracked point $\mathbf{p}_{i,0}$, we back-project its first frame location into 3D as

$$\mathbf{X}_{i,0} = D_0(\mathbf{p}_{i,0})\mathbf{K}_0^{-1}\tilde{\mathbf{p}}_{i,0}, \qquad \tilde{\mathbf{p}}_{i,0} = [u_{i,0}, v_{i,0}, 1]^\mathsf{T}, \tag{7}$$

where D_0 and $\mathbf{K}_0$ denote the depth map and camera intrinsics of the first frame, respectively. We then transform $\mathbf{X}_{i,0}$ to frame t using the corresponding camera extrinsics and reproject it onto the image plane:

$$\hat{\mathbf{p}}_{i,t}^{\mathrm{cam}} = \pi(\mathbf{K}_t\mathcal{T}_{0\to t}(\mathbf{X}_{i,0}))\,, \qquad \mathbf{f}_{i,t}^{\mathrm{stab}} = \mathbf{p}_{i,t} - \hat{\mathbf{p}}_{i,t}^{\mathrm{cam}}, \tag{8}$$

where $\mathcal{T}_{0\to t}$ maps 3D points from the camera coordinate system of frame 0 to that of frame t, π denotes perspective division, and $\mathbf{f}_{i,t}^{\mathrm{stab}}$ denotes the camera-motion-compensated displacement. By suppressing motion caused by viewpoint changes, the stabilized pointflow enables the model to better learn interaction dynamics from human demonstration data.

As shown in Fig. 2, our dual-branch architecture consists of a video generation branch and a lightweight auxiliary pointflow branch, which provides shared motion supervision across human and robot interactions. It reads intermediate video latents from layers 4, 6, 8, and 10, and predicts the flow matching velocity of noisy future pointflow using the first frame visual latent and initial point coordinates as additional inputs. Pointflow tokens are updated through self attention and cross attention to the intermediate video representations. Importantly, the pointflow branch attends only to the intermediate representations of the generated video, without direct access to the action sketch conditioning tokens. This design avoids over reliance on the action conditioning signal and encourages pointflow to capture motion from the generated visual dynamics, providing more effective supervision for the video backbone.

We train the pointflow branch with the following flow-matching objective:

$$\mathcal{L}_{\mathrm{point}} = \mathbb{E}_{\rho,\boldsymbol{\epsilon}}\left[\|g_\phi(\mathbf{S}_\rho, \rho, \mathbf{z}_0, \mathbf{p}_0; \mathbf{H}_\theta) - (\boldsymbol{\epsilon} - \mathbf{S})\|_2^2\right], \quad \mathbf{S}_\rho = (1-\rho)\mathbf{S} + \rho\boldsymbol{\epsilon}, \tag{9}$$

where $\mathbf{S}$ denotes the pointflow supervision target formed by stacking $\mathbf{f}_{i,t}^{\mathrm{stab}}$ over tracked points and future frames, $\mathbf{S}_\rho$ is its noisy interpolation at noise level $\rho \in [0, 1]$, and $\boldsymbol{\epsilon}$ is Gaussian noise. $\mathbf{z}_0$ and $\mathbf{p}_0$ denote the first-frame visual latent and point coordinates, respectively. $\mathbf{H}_\theta$ represents the intermediate video latents extracted from the four selected backbone layers, and g_ϕ denotes the pointflow head. The two branches are jointly optimized with the original video flow-matching objective:

$$\mathcal{L} = \mathcal{L}_{\mathrm{video}} + \lambda_p\mathcal{L}_{\mathrm{point}}, \tag{10}$$

where $\mathcal{L}_{\mathrm{video}}$ is the video flow-matching loss $\mathcal{L}_{\mathrm{RF}}$ defined in the preliminaries, and λ_p controls the contribution of pointflow supervision. The auxiliary pointflow branch is used only during training and does not alter the video generation pathway. Its gradients propagate through the intermediate video latents to the backbone, encouraging representations that better capture interaction-induced motion, while introducing no additional overhead at inference time.

3.3 Learning Across Outcomes: From Success to Failure

Robot manipulation datasets are dominated by successful demonstrations, leaving failure dynamics substantially underrepresented. Consequently, supervised fine-tuning on such data provides limited supervision for the consequences of abnormal gripper operations, mistimed actuation, and deviations

in end-effector pose. To overcome this limitation, we introduce a second-stage GRPO post-training framework that learns failure dynamics from counterfactual action perturbations.

We first classify each action segment according to its gripper-state transition pattern and select the perturbation accordingly. For trajectories with clear gripper transitions, we perturb either the gripper state and transition timing or the end-effector pose, covering different failure modes that may arise during interaction. For trajectories without clear gripper events, we primarily perturb the end-effector pose through rotation and translation, focusing optimization on action following and motion consistency while avoiding gripper changes inconsistent with the original interaction.

The counterfactual actions do not have corresponding ground-truth future videos, making direct optimization with supervised video losses infeasible. To address this, we apply GRPO by generating multiple candidate futures under the same counterfactual action condition and performing group-relative optimization with action alignment and physical plausibility rewards. The action alignment reward measures the consistency between generated motion and the conditioning action, while the physical plausibility reward evaluates the physical validity of contact, object motion, and robot structure.

Action Alignment Reward. We train a visual action alignment scorer with pairwise supervision. Each training comparison consists of two video–action samples, (V, A^+) and (V, A^-), sharing the same RGB video but containing different action skeletons. Comparisons include original actions against perturbed variants, as well as variants with different perturbation strengths, covering pose offsets, temporal shifts, and gripper-state changes. We overlay its action skeleton on the corresponding RGB frames, using the full-frame sequence as the global input and crops around the end effector as the local input. The global branch encodes overall robot pose and motion, while the local branch focuses on the wrist, gripper, and interaction region. Their features are fused and aggregated by a temporal Transformer. The two samples are evaluated independently by the same scoring network, producing predictions z^+ and z^-. We convert these predictions into a temperature-scaled preference probability and a signed preference gap:

$$p = \sigma\left(\frac{z^+ - z^-}{\tau}\right), \qquad r_\Delta = 2p - 1, \tag{11}$$

where p is the predicted probability of preferring the first sample over the second, σ is the sigmoid function, and τ is a temperature parameter. The core training objective supervises both the preference direction and the gap magnitude:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{rank}} + \lambda_{\text{gap}}\mathcal{L}_{\text{gap}}, \qquad \mathcal{L}_{\text{gap}} = \text{SmoothL1}(r_\Delta, g^*). \tag{12}$$

The ranking term encourages a higher score for the better-aligned sample, while the gap term matches the predicted preference gap to a target g^* specified by the perturbation type, relative severity, and visual discernibility. Subtle differences receive smaller target gaps, whereas pronounced differences in pose, timing, or gripper state receive larger ones. Gradients from the pairwise objective propagate through both predictions to update the shared scoring network.

During GRPO, the learned scorer independently evaluates candidate rollouts conditioned on the same action sequence. Their predictions are converted into pairwise preferences using Eq. 11, and the resulting preference gaps are used to construct the action alignment reward.

Physical Plausibility Reward. We adopt a hierarchical evaluation scheme (Chen et al., 2026) based on a shared physical checklist tailored to each counterfactual interaction. A vision-language model constructs the checklist from a multimodal prompt comprising the original demonstration video, task description, and a textual description of the action perturbation, grounding its questions in scene-specific object interactions, contact relations, and support conditions. By deriving the checklist from this context without access to candidate videos, we establish fixed evaluation criteria for all rollouts within each group. The evaluation assesses the physical plausibility of counterfactual interactions rather than task progress, without treating task failure as a physical violation. The checklist comprises two tiers: Tier 1 identifies critical physical violations, such as object interpenetration and unsupported floating; Tier 2 provides fine-grained assessments of physical consistency, robot structure, contact causality, and visual quality. The VLM evaluates each candidate against the checklist, with judgments aggregated across dimensions to obtain the physical reward. If a candidate fails any Tier 1 check or exhibits a severe violation, we cap its physical reward and omit its action alignment reward, preventing action consistency from offsetting physical violations. Prompt templates and scoring rules are provided in the Appendix A.2.



Figure 3: **Qualitative comparison of action-conditioned video generation.** Evaluated on both successful and failed execution samples, our method is compared against six models and achieves much better visual quality with precise action following.

Table 1: Quantitative comparison with baselines. **Best** / second-best results are bolded / underlined.

Method	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	$L2_{latent}$ ↓	IDM↓
Cosmos-Predict2.5 (Ali et al., 2025)	18.874	0.746	0.266	60.119	54.108	0.486	1.268
TesserAct (Zhen et al., 2025)	17.383	0.757	0.263	60.563	59.889	0.523	1.663
Ctrl-World (Guo et al., 2025)	21.683	0.770	0.305	36.916	64.665	0.418	0.949
Cosmos 3 (Agarwal et al., 2026)	20.233	0.766	0.178	42.987	36.780	0.365	1.108
OSCAR (Wu & Gao, 2026)	23.357	0.840	0.142	31.815	28.241	0.292	0.579
Kinema4D (Xu et al., 2026)	20.015	0.804	0.157	30.511	30.890	0.326	0.914
HF-WM (**Ours**)	**26.077**	**0.889**	**0.082**	**14.449**	**14.291**	**0.193**	**0.332**

4 Experiments

4.1 Experimental Setup

We fine-tune Cosmos 3 Nano in two stages on 8 NVIDIA A100 GPUs. The first stage performs supervised fine-tuning on robot and egocentric human interaction data with auxiliary stabilized point-flow supervision to improve motion modeling, while the second stage applies Group Relative Policy Optimization (GRPO) to enhance the modeling of failure outcomes. All training samples adopt the unified skeleton representation provided by OSCAR (Wu & Gao, 2026). The training set contains approximately 45K Franka trajectories from RH20T (Fang et al., 2023), InternData-A1 (Tian et al., 2025), and DROID (Khazatsky et al., 2024), together with 83K human interaction trajectories from EgoDex (Hoque et al., 2025) and EPIC-Kitchens (Damen et al., 2018). Further training details are provided in the Appendix B.

4.2 Comparison with Baselines

We evaluate all methods on 100 held-out Franka trajectories selected from DROID and 100 failed trajectories collected from real-world robot executions, with the reported results averaged over all evaluation trajectories. Based on how the conditioning signal is incorporated, the baselines are grouped into three categories: text-only methods (TesserAct (Zhen et al., 2025), Cosmos-Predict2.5 (Ali et al., 2025)), latent action methods (Ctrl-World (Guo et al., 2025) and Cos-

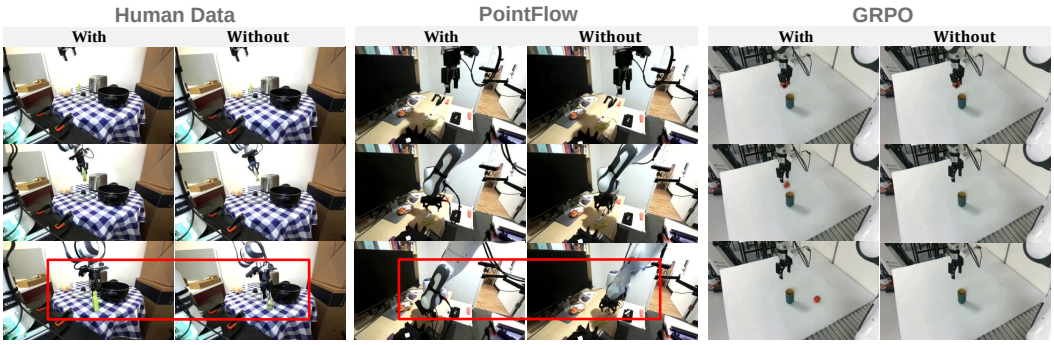


Figure 4: **Qualitative ablation of the key components in HF-WM.** Large-scale human interaction data broadens the diversity of object interactions, while unified pointflow supervision provides a shared motion representation that facilitates cross-embodiment transfer. Building on this foundation, GRPO post-training improves the modeling of failure dynamics across diverse execution outcomes.

Table 2: Ablation study on the effectiveness of human interaction data and pointflow constraints.

Variant	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	$L2_{latent}$ ↓	IDM↓
baseline(Cosmos3)	20.233	0.766	0.178	42.987	36.780	0.365	1.108
embodied data only	21.448	0.833	0.120	23.650	20.574	0.332	0.424
+ human	24.997	0.876	0.093	15.229	16.980	0.303	0.391
Δ gain ↑	3.549	0.043	0.027	8.421	3.594	0.029	0.033
embodied data + pointflow	21.573	0.835	0.117	22.993	20.011	0.329	0.406
+ human + pointflow	25.294	0.880	0.090	15.327	16.972	0.227	0.376
+ human + stabilized pointflow	**25.828**	**0.887**	**0.086**	**14.993**	**15.991**	**0.198**	**0.355**
Δ gain ↑	4.255	0.052	0.031	8.000	4.020	0.131	0.051

mos 3 (Agarwal et al., 2026)), and explicit geometry methods (OSCAR (Wu & Gao, 2026) and Kinema4D (Xu et al., 2026)). In addition to standard visual quality metrics, we fine-tune Cosmos 3 using its inverse dynamics modeling (IDM) capability to obtain a Franka-specific IDM. It predicts actions from generated videos, which are compared with the conditioning actions to assess action following fidelity.

As shown in Fig. 3 and Table 1, HF-WM achieves state-of-the-art performance across visual quality, action following, and dynamics modeling, demonstrating its strong overall capability for action-conditioned world modeling. Notably, Table 5 reveals a consistent performance gap between successful and failed executions across existing models, suggesting that models trained predominantly on successful demonstrations are less effective at capturing underrepresented failure dynamics. Such an imbalance is particularly limiting for world models, which are expected to faithfully predict action consequences across diverse execution outcomes. In contrast, HF-WM maintains strong performance on both successful and failed executions, demonstrating improved robustness in modeling failure dynamics. Across both tables, HF-WM achieves the best overall performance in visual quality, outcome modeling, and action following, demonstrating more accurate and robust modeling of action-conditioned dynamics across diverse execution outcomes.

4.3 Ablation Studies

Effectiveness of human data. In the first stage of fine-tuning, we evaluate the contributions of human interaction data and our dual-branch architecture with auxiliary pointflow supervision. Compared with using embodied data alone, incorporating human data yields substantial performance gains. Further aligning human and robot motions through a unified pointflow representation leads to additional improvements. For egocentric human videos, we additionally stabilize the pointflow to suppress camera-induced motion. As shown in Fig. 4 and Table 2, combining human data with stabilized pointflow yields larger Δ gains across most metrics than adding human data alone, measured against their respective embodied-data baselines. These results suggest that unifying human and robot motion representations while suppressing camera-induced motion facilitates cross-

Table 3: Quantitative ablation of GRPO for learning from failure cases.

Variant	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	$L2_{latent}$ ↓	IDM↓
w/o GRPO	25.828	0.887	0.086	14.993	15.991	0.198	0.355
failure-aware GRPO + action alignment	25.826	0.887	0.086	14.957	15.957	0.198	0.348
failure-aware GRPO + physical plausibility	25.574	0.862	0.098	15.307	15.454	0.209	0.381
failure-aware GRPO + full rewards	**26.077**	**0.889**	**0.082**	**14.449**	**14.291**	**0.193**	**0.332**

embodiment knowledge transfer, enabling embodied world models to learn more effectively from human manipulation data.

Effectiveness of failure-aware GRPO. We evaluate the contribution of GRPO and its reward components in the second stage. Fig. 4 illustrates improved failure modeling, while Table 3 quantifies the complementary effects of action alignment and physical plausibility. As shown in Fig. 10 and Fig. 11, using only the physical plausibility reward tends to produce nearly static scenes or physically plausible motion that contradicts the conditioning actions. Conversely, using only the action alignment reward preserves action-consistent robot motion but compromises temporal consistency in background regions and the physical plausibility of interaction dynamics. Combining the two rewards improves action adherence while maintaining physically coherent scene dynamics. Table 7 further shows that GRPO improves performance on both successful and failed executions. Additional ablation studies are provided in Appendix B.3.

4.4 Policy Evaluation

To assess HF-WM as a real-world policy evaluator, we select six generalist policies from the RoboArena leaderboard (Atreya et al., 2025) and evaluate each policy on the same 90 matched execution sessions, yielding 540 policy-session evaluations in total. For each session, HF-WM receives only the initial observation and the corresponding policy action sequence and autoregressively predicts the subsequent interaction. GPT-5 (Singh et al., 2025) evaluates task success from the generated videos, yielding a predicted success rate for each policy. The resulting policy estimates are further compared with real-world deployment performance. As shown in Fig. 5, the predicted success rates exhibit a strong linear correlation with their real-world counterparts, with Pearson $r = 0.983$. MMRV (Li et al., 2024) measures the preservation of relative policy performance, with $\mathrm{MMRV} = 0.002$. These results show that HF-WM accurately reflects both policy performance and relative rankings, demonstrating its potential as a robot policy evaluator. Additional metrics and statistical analyses are provided in the Appendix C.1.

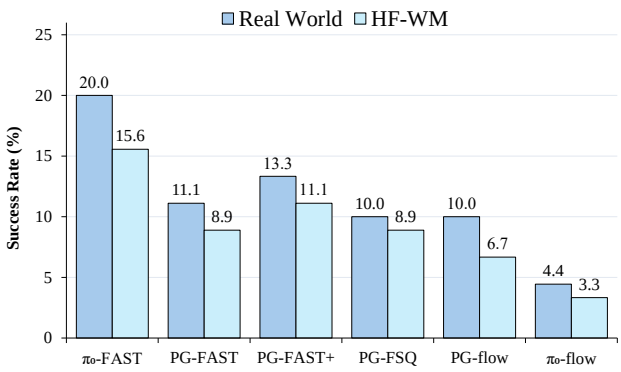


Figure 5: Real-world and HF-WM estimated success rates across six generalist robot policies.

5 Conclusion

We presented HF-WM, an action-conditioned world model that integrates egocentric human interaction learning with failure-aware post-training. Its dual-branch architecture employs stabilized pointflow supervision to establish a shared motion representation across human and robot interactions while reducing the effects of egocentric camera motion. GRPO post-training on perturbed actions improves failure dynamics modeling through complementary action alignment and physical plausibility rewards, without paired ground-truth videos. Experiments demonstrate state-of-the-art performance in action-conditioned video generation and accurate estimation of real-world policy performance. These results highlight the value of learning from human interactions and modeling failure dynamics in extending world models beyond successful robot demonstrations.

REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

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.

Pranav Atreya, Karl Pertsch, Tony Lee, Moo Jin Kim, Arhan Jain, Artur Kuramshin, Clemens Eppner, Cyrus Neary, Edward Hu, Fabio Ramos, et al. Roboarena: Distributed real-world evaluation of generalist robot policies. In *Proceedings of the Conference on Robot Learning (CoRL 2025)*, 2025.

Yuzhi Chen, Ronghan Chen, Dongjie Huo, Yandan Yang, Dekang Qi, Haoyun Liu, Tong Lin, Shuang Zeng, Junjin Xiao, Xinyuan Chang, et al. Abot-physworld: Interactive world foundation model for robotic manipulation with physics alignment. *arXiv preprint arXiv:2603.23376*, 2026.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

Hao-Shu Fang, Hongjie Fang, Zhenyu Tang, Jirong Liu, Chenxi Wang, Junbo Wang, Haoyi Zhu, and Cewu Lu. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. *arXiv preprint arXiv:2307.00595*, 2023.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Yanjiang Guo, Lucy Xiaoyang Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. *arXiv preprint arXiv:2510.10125*, 2025.

Jinkun Hao, Mingda Jia, Ruiyan Wang, Xihui Liu, Ran Yi, Lizhuang Ma, Jiangmiao Pang, and Xudong Xu. Egosim: Egocentric world simulator for embodied interaction generation. *arXiv preprint arXiv:2604.01001*, 2026.

Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025.

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

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. *arXiv preprint arXiv:2403.12945*, 2024.

Zihang Lai, Eldar Insafutdinov, Edgar Sucar, and Andrea Vedaldi. Cowtracker: Tracking by warping instead of correlation. *arXiv preprint arXiv:2602.04877*, 2026.

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

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024.

Anthony Liang, Yigit Korkmaz, Jiahui Zhang, Minyoung Hwang, Abrar Anwar, Sidhant Kaushik, Aditya Shah, Alex S. Huang, Luke Zettlemoyer, Dieter Fox, Yu Xiang, Anqi Li, Andreea Bobu, Abhishek Gupta, Stephen Tu, Erdem Biyik, and Jesse Zhang. Robometer: Scaling general-purpose robotic reward models via trajectory comparisons. In *Robotics: Science and Systems 2026*, 2026.

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.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *arXiv preprint arXiv:2306.03310*, 2023.

Jie Liu, Gongye Liu, Jiajun Liang, Yangguang Li, Jiaheng Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Wanli Ouyang. Flow-grpo: Training flow matching models via online rl. *Advances in neural information processing systems*, 38:40783–40818, 2026.

Chong Ma, Taiyi Su, Jian Zhu, Jianjun Zhang, Zitai Huang, Yi Xu, and Hanli Wang. Pil-world: A chunk-wise world model for vla policy-in-the-loop evaluation. *arXiv preprint arXiv:2606.05773*, 2026.

Zhiting Mei, Tenny Yin, Ola Shorinwa, Apurva Badithela, Zhonghe Zheng, Joseph Bruno, Madison Bland, Lihan Zha, Asher Hancock, Jaime Fernández Fisac, et al. Video generation models in robotics-applications, research challenges, future directions. *arXiv preprint arXiv:2601.07823*, 2026.

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.

Bowen Ping, Chengyou Jia, Minnan Luo, Hangwei Qian, and Ivor Tsang. Flow-factory: A unified framework for reinforcement learning in flow-matching models. *arXiv preprint arXiv:2602.12529*, 2026. URL https://arxiv.org/abs/2602.12529.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Zhenguo Sun, Yu Sun, Hande Huang, and Alois Knoll. ω-eva: Envision, verify, and act with latent interactive world models. *arXiv preprint arXiv:2606.09457*, 2026.

Yang Tian, Yuyin Yang, Yiman Xie, Zetao Cai, Xu Shi, Ning Gao, Hangxu Liu, Xuekun Jiang, Zherui Qiu, Feng Yuan, et al. Interndata-a1: Pioneering high-fidelity synthetic data for pre-training generalist policy. *arXiv preprint arXiv:2511.16651*, 2025.

Wei-Cheng Tseng, Gashon Hussein, Yuzhu Dong, Allen Z Ren, Lucy X Shi, XuDong Wang, Sergey Levine, Zhaoshuo Li, Jinwei Gu, Florian Shkurti, et al. Sc3-eval: Evaluating robot foundation models via self-consistent video generation. *arXiv preprint arXiv:2606.18610*, 2026.

Peiyan Tu, Hanxin Zhu, Jingwen Sun, Shaojie Ren, Cong Wang, Jiayi Luo, Xiaoqian Cheng, and Zhibo Chen. Embody4d: A generalist 4d world model for embodied ai. *arXiv preprint arXiv:2605.01799*, 2026.

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

Cong Wang, Hanxin Zhu, Xiao Tang, Jiayi Luo, Xin Jin, Long Chen, and Zhibo Chen. Orthophys: Physically plausible video generation with orthogonal-view geometry guidance, 2026.

Zhuoyuan Wu and Jun Gao. Oscar: Omni-embodiment action-conditioned world model for robotics. *arXiv e-prints*, pp. arXiv–2606, 2026.

Mutian Xu, Tianbao Zhang, Tianqi Liu, Zhaoxi Chen, Xiaoguang Han, and Ziwei Liu. Kinema4d: Kinematic 4d world modeling for spatiotemporal embodied simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4495–4500, 2026.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. Mirabench: Evaluating action-conditioned reliability in robotic world models. *arXiv preprint arXiv:2605.29360*, 2026a.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026b.

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.

Mark YU, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. *arXiv preprint arXiv:2503.05638*, 2025.

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

Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. TesserAct: Learning 4D embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

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

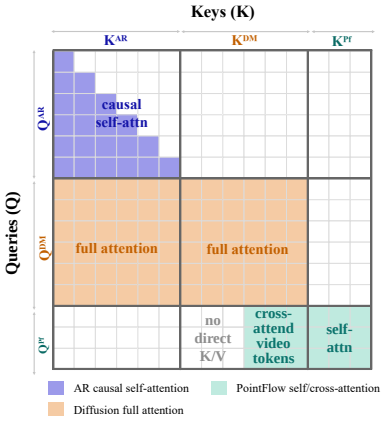
Hanxin Zhu, Cong Wang, Peiyan Tu, Jiayi Luo, Tianyu He, Xin Jin, and Zhibo Chen. Gta: Advancing image-to-3d world generation via geometry then appearance video diffusion. *arXiv preprint arXiv:2605.12957*, 2026.

APPENDIX FOR HF-WM

A IMPLEMENTATION DETAILS

A.1 LEARNING ACROSS EMBODIMENTS: FROM HUMAN TO ROBOT

Attention mask. As illustrated in Fig. 6, we retain the dual-stream attention mechanism of Cosmos 3 (Agarwal et al., 2026). The backbone consists of an autoregressive (AR) stream and a diffusion-model (DM) stream. AR queries attend causally within the AR stream, whereas DM queries attend bidirectionally to the AR stream and to all tokens in the DM stream, including video, skeleton, and action tokens. To explicitly constrain the motion of the generated video, we introduce pointflow (PF) queries as an auxiliary visual-motion readout branch outside the backbone token sequence. PF queries self-attend only within the PF branch and cross-attend exclusively to intermediate video features from the DM stream. Although video, skeleton, and action tokens are all part of the DM stream, skeleton and action keys and values are explicitly masked from PF cross-attention; AR, language, and other non-video features are also inaccessible. This design prevents the PF branch from taking a shortcut through the conditioning signals and forces the pointflow objective to be grounded in the generated video representations, improving motion consistency and geometric fidelity. The allowed attention pattern is:



Query group	Accessible keys and values	Masking constraint
AR queries	AR keys and values	Causal attention within AR stream
DM queries	AR keys/values and video, skeleton, action	Bidirectional attention over AR and DM streams
PF cross-attention	Intermediate DM video features only	Skeleton/action/language; AR and non-video keys/values masked
PF self-attention	PF keys and values	Self-attention restricted to the PF branch

Figure 6: HF-WM architecture and attention visibility. Left: attention mask. Right: accessible keys/values and masking constraints for AR, DM, and PF queries.

Stabilized PointFlow Processing. The main text describes our camera-motion compensation for pointflow; here, we provide additional preprocessing details. For each 81-frame video, we track $1008 = 24 \times 42$ points on a regular grid and establish their 3D reference positions from the first frame using depths predicted by Depth Anything 3 (Lin et al., 2025). Depth values are bilinearly sampled at the initial point locations, and the points are reprojected to subsequent frames using the ground-truth camera poses to estimate camera-induced motion. This motion is then subtracted from the original CoWTracker trajectories to obtain stabilized point trajectories. To align pointflow supervision with the temporal resolution of the video latents, we construct relative pointflow with respect to the first frame and temporally subsample the 81-frame trajectories with a stride of 4, yielding 20 supervision time steps.

As shown in Fig. 7, static background points in the raw pointflow exhibit coherent drift under egocentric camera motion. After camera-motion compensation, this global background motion is substantially reduced, while the relative motion of the hand and manipulated objects is preserved. This indicates that the stabilization primarily removes viewpoint-induced motion rather than indiscriminately smoothing the point trajectories.

A.2 LEARNING ACROSS OUTCOMES: FROM SUCCESS TO FAILURE

We construct fixed-length GRPO training clips anchored to gripper-state transitions. For each episode, we identify the frame with the largest absolute change in gripper state relative to the pre-

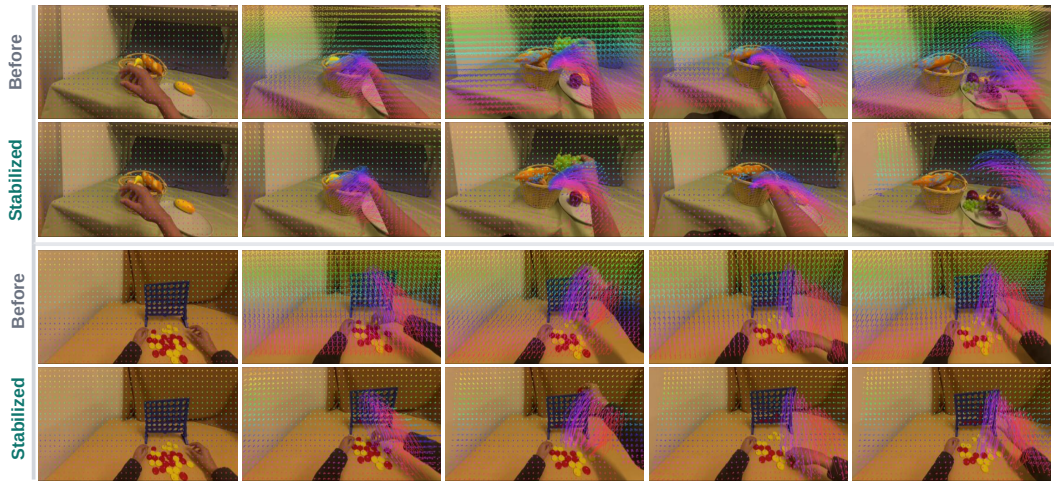


Figure 7: Visualization of pointflow stabilization on egocentric interactions. Camera-motion compensation suppresses coherent background drift while preserving hand-object interaction motion.

ceding frame. Each clip contains 41 frames sampled with a stride of 2, spanning 81 source frames. We set the window start 12 source frames before the detected transition and clip it to the valid range to keep the full window within the episode. This event-anchored sampling captures pre-transition context and subsequent interaction dynamics without requiring the transition to lie at the window midpoint. We then classify each sampled window by its gripper-state transition pattern to guide subsequent action perturbations. We apply a length-3 median filter to suppress transient fluctuations and identify stable states using dual-threshold hysteresis: values at or below 0.25 indicate a closed state, values at or above 0.75 indicate an open state, and intermediate values do not trigger state transitions. Starting from the first stable state, we track the initial transition direction and any subsequent reversal to distinguish four patterns: closure without reopening (`close_hold`), closure followed by reopening (`close_reopen`), opening without reclosing (`open_hold`), and opening followed by reclosing (`open_close`). Windows without valid transitions are classified by their overall gripper-value range as continuously open (`remain_open`), continuously closed (`remain_closed`), or indeterminate (`partial_hold`). This seven-way classification determines eligibility for gripper perturbations and guides the choice between gripper and geometric perturbations. After gripper-event classification, we sample training windows according to predefined class quotas, allocating 75% to the four classes with explicit gripper-state transitions and 25% to the remaining three classes. For each window, we select the unperturbed route (`none`) with probability 50%, retaining the original actions and corresponding video; otherwise, we enter the counterfactual perturbation branch. This balanced sampling aims to prevent overemphasis on perturbed executions while preserving the modeling of original executions as the model learns failure-related dynamics.

The counterfactual branch selects perturbations according to the event class. For windows with explicit state transitions, gripper and geometric perturbations are sampled with probabilities of 70% and 30%, respectively. This yields within-class routing probabilities of 50% for no perturbation, 35% for gripper perturbations, and 15% for geometric perturbations. The specific gripper operation depends on the event pattern, as summarized in Table 4. Within the `open_hold` gripper branch, reclosing (`reclose`) and suppression of gripper-state changes (`suppress`) are selected with probabilities of 25% and 75%, respectively. For windows without valid state transitions, the counterfactual branch applies only geometric perturbations. Geometric perturbations are sampled uniformly across one rotation route and four translation routes. Each geometric route therefore has a within-class selection probability of 3% for classes with explicit state transitions and 10% for those without valid transitions.

During route execution, gripper perturbations modify only the gripper-state sequence, preserving the arm joint trajectory and end-effector poses without requiring inverse kinematics (IK). Geometric perturbations must satisfy both visual and kinematic constraints. For each geometric candidate, we compute end-effector poses from the original joint states and construct a target trajectory using local $\mathrm{SE}(3)$ perturbations. A cosine ramp smoothly increases the translational and rotational offsets

Table 4: Gripper event patterns, target sampling shares, and perturbation routing probabilities. Target shares are relative to all training samples; routing probabilities are conditional on the event class and sum to 100% per row.

Event class	Gripper event pattern	Target share	None	Gripper perturbation	Geometric total
close_hold	Open → Closed (hold)	18.75%	50%	reopen: 35%	15%
close_reopen	Open → Closed → Open	18.75%	50%	force_closed: 35%	15%
open_hold	Closed → Open (hold)	18.75%	50%	reclose: 8.75% suppress: 26.25%	15%
open_close	Closed → Open → Closed	18.75%	50%	force_open: 35%	15%
remain_open	Open (hold)	12%	50%	Not applied	50%
partial_hold	Indeterminate	8%	50%	Not applied	50%
remain_closed	Closed (hold)	5%	50%	Not applied	50%

from zero to their target magnitudes, mitigating abrupt trajectory changes. Before solving IK, we verify sufficient gripper visibility in both the source skeleton and the unperturbed baseline. We solve IK via bounded nonlinear least squares under the URDF-specified joint limits, minimizing normalized end-effector position and orientation errors. Each subsequent solve is warm-started from the preceding solution. A candidate passes kinematic validation only if all joint values are finite and within limits, position and orientation residuals fall below their respective thresholds, and joint changes between consecutive states satisfy continuity constraints. For kinematically feasible candidates, we apply forward kinematics to the resulting joint trajectory and re-render the skeleton and gripper using the camera intrinsics and extrinsics. We verify that the end effector remains within the valid image region and the gripper remains visible. Only candidates satisfying both visual and kinematic constraints are accepted. If IK or action reconstruction fails, or re-rendering reveals insufficient gripper visibility or anomalous displacement, we restore the original inputs and try another eligible route.

A.2.1 Action Alignment Reward

As illustrated in Fig. 8, our action alignment reward model uses a global-local dual-branch architecture to assess consistency between generated videos and conditioning actions at complementary visual scales. Both branches process 41-frame sequences aligned with GRPO rollouts. The global branch jointly encodes RGB frames and their action-skeleton overlays at 224×224 resolution to capture overall robot motion. The local branch extracts 80×80 end-effector crops from the overlays and resizes them to 160×160 to capture fine-grained motion. Independent convolutional encoders extract per-frame visual features from each branch, which are fused and projected into unified frame-level representations. These representations are augmented with temporal positional encodings and processed by a three-layer temporal Transformer to jointly model global motion trajectories and local end-effector dynamics. We construct training samples spanning different alignment levels through controlled action perturbations. For each real interaction clip, we keep the original RGB video fixed, perturb the action sequence with varying types and magnitudes, and re-render the corresponding skeletons using the robot's URDF and forward kinematics. Each action candidate is paired with the same video, focusing supervision on alignment differences between conditioning actions and observed motion.

Perturbations cover arm pose, wrist rotation, temporal shifts, and gripper states. Gripper perturbations include forced opening or closure, state inversion, delayed transitions, and reopening after closure. Pose and temporal perturbations span multiple magnitudes, and combined perturbations extend coverage from minor deviations to severe misalignment. Since few-step denoising in GRPO

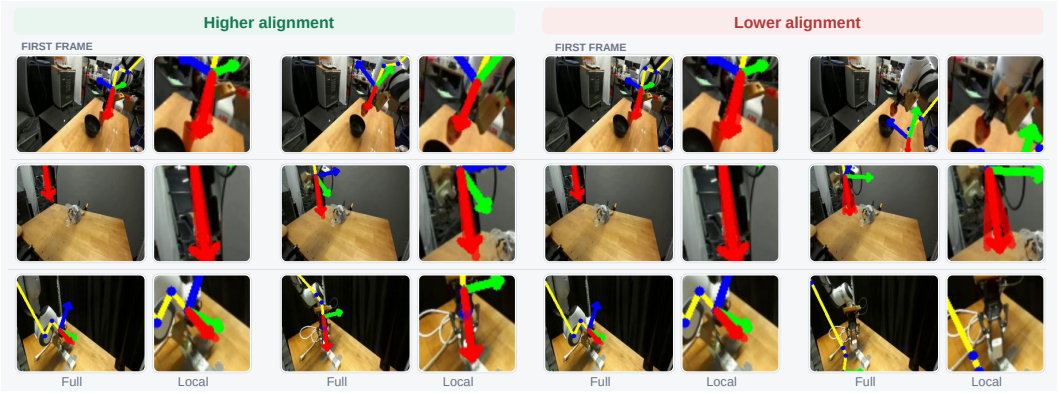


Figure 8: **Qualitative examples for the action alignment reward.** Each row compares video-action pairs with higher (left) and lower (right) alignment. Action-skeleton overlays in full-frame views and end-effector crops highlight differences in consistency between the specified actions and observed robot motion.

may degrade local image quality, we additionally apply visual corruptions, such as blur and noise, to end-effector regions while preserving the actions and skeleton trajectories. These augmentations aim to reduce the influence of appearance quality on alignment assessment. Training samples are organized into preference pairs according to relative alignment, including comparisons between original and perturbed actions as well as between perturbed actions with different degrees of deviation. The preferred candidate is more consistent with the observed video but need not be fully correct. Thus, supervision distinguishes not only correct from incorrect action conditions but also mild from severe misalignment.

Each preference pair is supervised with a continuous target score gap, determined by predefined rules based on perturbation type and magnitude. Gripper deviations are quantified by differences in openness from the original trajectory over non-conditioning frames. For pose, rotation, and temporal perturbations, target gaps are determined by joint offsets, rotation angles, and temporal shifts, respectively. Identical or visually indistinguishable inputs receive a zero target gap to avoid amplifying unobservable action differences into reward differences. This graded supervision captures both relative ordering and the extent of alignment differences, rather than binary correctness alone.

A.2.2 Physical Plausibility Reward

We extract interaction events, contact relations and support conditions from the original RGB video, annotating the corresponding evidence frame intervals. This establishes an observational baseline independent of perturbation routing and candidate outcomes for subsequent counterfactual assessment. Combining these observations with perturbation metadata from the preceding classification and routing stage, we construct counterfactual event descriptions based on interaction type, gripper dependence, and scene dynamics. These descriptions specify event relations to preserve, permissible state changes, and uncertainty where evidence is insufficient, defining a set of physically admissible future outcomes.

Specifically, translational or rotational perturbations should generally preserve the original event graph of contact, grasping, release, and support. In contact-free scenes, perturbations to gripper opening and closing should affect only the gripper itself. For gripper-dependent grasps, state changes such as slippage or falling require visible evidence of contact or support loss in the candidate video, rather than being inferred solely from action intent or gripper signals.

Building on these observations and counterfactual event descriptions, we assess physical plausibility through tiered binary question answering (Chen et al., 2026), using Gemini 3.7 Flash for both question generation and candidate evaluation. Without access to candidate videos, the question generator constructs and freezes a checklist shared by all candidates in a group. Tier-1 comprises three critical checks targeting severe physical or structural violations, such as object interpenetration, unsupported floating, and major inconsistencies in robot geometry. Tier-2 comprises five fine-grained checks spanning four dimensions: **physical consistency**, assessing object persistence, motion continu-

GT Describer gemini-3.1-flash

You are the GT factual describer for a robot manipulation video. Observe only the unperturbed original RGB window and the synchronized gripper signal. Your task is to write a frame-grounded, neutral, factual record of the events that are
actually visible in this window. Do not score anything, select a route, design a perturbation, or predict a candidate video. Do not write that a candidate should succeed or fail, and do not write any reward. Do not output candidate-specific conclusions.

MANDATORY CAUSAL RULES:
First watch the complete video from F00 through the final frame. Then record the following in temporal order:

- 1. The scene, camera, table or support surface, and all relevant objects;
- 2. The gripper fingers, wrist connection, opening/closing direction, and the frame interval in which opening or closing occurs;
- 3. Gripper approach, visible contact, grasp, lift, hold, push, release, drop, or stillness;
- 4. The support state and motion state of each target before, during, and after each event;
- 5. The final visible state, including occlusion, cropping, blur, and portions that can not be confirmed.

Every event must include a continuous Fxx-Fyy evidence interval. Record only pixel-visible: start_state -> event -> end_state

Output strict JSON:

```
{
 "scene": "...",
 "objects": [{"id": "obj_1", "appearance": "...", "initial_support": "..."}],
 "gripper_observation": {"shape": "...", "event": "...", "interval": "Fxx-Fyy"},
 "event_timeline": [{"time": "Fxx-Fyy", "subject": "...", "start_state": "...", "event": "...", "end_state": "...", "evidence_frames": ["Fxx", "Fyy"]}],
 "contact_observations": [{"time": "Fxx-Fyy", "objects": "...", "evidence": "..."}],
 "support_and_motion": [{"object": "...", "before": "...", "during": "...", "after": "..."}],
 "final_state": "...",
 "uncertainties": ["..."],
 "description_source": "gt_video_understanding_api"
}
```

Future-Event Contract Generator gemini-3.7-flash

You are the GT event classifier and counterfactual event-contract generator.
INPUTS: (a) ORIGINAL_GT_DESCRIPTION; (b) NUMERICAL_GRIPPER_FACTS; (c) upstream-generated APPLIED_PERTURBATION metadata.
The perturbation has already been completed. Do not resample, delete, modify, or change the perturbation. Do not change any perturbation proportions.

INTERACTION CLASSIFICATION
Classify the interaction using only continuous frame evidence from ORIGINAL_GT_DESCRIPTION.
interaction_class ∈ {no_interaction, incidental_contact, contact_non_grasp, grasp_attempt_failed, grasp_success, release_drop, push, pull, articulated, uncertain}
interaction_dependency ∈ {none, gripper_dependent, other_contact, unknown}
dynamicity ∈ {static, weak, moderate, strong, uncertain}

Combine APPLIED_PERTURBATION with the GT facts to construct future_event_contract containing:
- invariant_events: Original events and support transitions that must remain invariant even if the perturbation changes screen coordinates.
- predicted_event_hypotheses: The most likely next events given the specific perturbation direction and GT contact dependency. These are hypotheses, not mandatory outcomes for every candidate.
- allowed_outcomes: Candidate events permitted after the perturbation, together with their evidence windows.
- forbidden_inventions: Grasps, collisions, attachments, drops, or object motions that cannot be created by this perturbation alone.
- focus_dimensions: Dimensions that should receive priority during question generation.
- uncertainty: When the original evidence is insufficient, conservatively report uncertain rather than inventing an event.

MANDATORY CAUSAL RULES:
(1) geometry, including translate or rotate: Preserve the original contact, grasp, and release event graph and support transitions. A change in screen coordinates, end-effector pose, or orientation alone does not prove that the contact topology changed. A natural drop that already occurred in the original video remains an allowed inherited event.
(2) aperture + (no_interaction or dependency = none): Allow only finger opening/closing and local gripper appearance changes. Do not add a new grasp, collision, attachment, object motion, or drop.
(3) aperture + gripper_dependent + (grasp_success or release_drop): The original target's contact-to-hold-to-lift/maintain-to-release chain may branch into continued success, slip, or drop only after visible contact and support-loss evidence. Unrelated objects must retain their original events.
- reopen or force_open: Prefer the hypothesis that gripper support may be lost, followed by continuous slipping or dropping. If reliable contact remains visible, continued holding is also allowed.
- force_closed, suppress, or close_hold: Prefer the hypothesis that the original release is suppressed and the target remains supported. Slip or drop is allowed only after visible loss of contact.
- reclose: Allow contact to be restored only if the target remains aligned with both fingers. Never invent a regrasp.
(4) other_contact, including push, pull, articulated, and contact_non_grasp: Preserve the original contact, collision, and support response. Do not rewrite these events as grasping.
(5) static or weak: Treat gripper finger count, wrist-to-finger connection, rigid-body shape, opening continuity, and local clarity as primary observation targets. A stationary object is not itself a failure.

Output strict JSON:

```
{"interaction_class": "...", "interaction_dependency": "...", "dynamicity": "...", "classification_evidence": ["Fxx-Fyy: ..."], "future_event_contract": {"invariant_events": ["..."], "predicted_event_hypotheses": ["..."], "allowed_outcomes": ["..."], "forbidden_inventions": ["..."], "focus_dimensions": ["..."], "uncertainty": ["..."]}}
```

Figure 9: Prompt template for GT Describer and Future-Event Contract Generator.

ity, and plausible dynamics under the observed support conditions; **robot structure**, examining the geometric integrity and temporal consistency of the arm and gripper; **contact causality**, assessing whether object motion is consistent with visible contact, grasping, release, and support transitions; and **visual quality**, checking for artifacts that obscure object states or interaction events. These checks are tailored to scene-specific objects and counterfactual interaction conditions,

Table 5: Quantitative comparison with baselines on successful and failed executions. Each entry reports Success / Failure. **Bold** / underline denote the best / second-best results.

Method (Success/Failure)	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	$\mathbf{L2}_{\text{latent}}$ ↓	IDM↓
Cosmos-Predict2.5 (Ali et al., 2025)	19.869/17.878	0.755/0.736	0.268/0.263	42.001/78.237	47.525/60.691	0.449/0.523	1.191/1.344
TesserAct (Zhen et al., 2025)	18.192/16.574	0.762/0.751	0.245/0.280	59.452/61.673	59.053/60.724	0.517/0.528	1.522/1.804
Ctrl-World (Guo et al., 2025)	21.746/21.620	0.788/0.752	0.301/0.309	32.247/41.585	61.721/67.608	0.406/0.429	0.887/1.011
Cosmos 3 (Agarwal et al., 2026)	21.172/19.293	0.815/0.716	0.142/0.213	38.512/47.461	34.205/39.354	0.349/0.381	1.101/1.114
OSCAR (Wu & Gao, 2026)	23.674/23.040	0.860/0.820	0.118/0.165	29.638/33.992	21.297/35.185	0.243/0.340	0.445/0.712
Kinema4D (Xu et al., 2026)	20.630/19.400	0.834/0.773	0.117/0.197	27.155/33.866	23.495/38.285	0.270/0.381	0.904/0.923
HF-WM (**Ours**)	**26.321/25.833**	**0.892/0.886**	**0.066/0.098**	**13.350/15.547**	**14.310/14.271**	**0.176/0.210**	**0.328/0.336**

rather than requiring the original demonstration outcome to be reproduced. Task failure alone is therefore not treated as a physical violation. In a separate model call, the evaluator answers each question with `yes` (pass) or `no` (fail) based on evidence from consecutive frames of the candidate video. The physical plausibility score is the fraction of the eight checks passed. Any Tier-1 failure triggers hard rejection, regardless of the Tier-2 score, preventing favorable assessments on other dimensions from compensating for a severe physical or structural violation.

B More Experimental Details

B.1 Training Details

In the first stage of supervised fine-tuning, we adopt and empirically validate the human data injection strategy from OSCAR (Wu & Gao, 2026). We first warm up the model on robot only embodied data and then jointly train with egocentric human interaction data, which yields larger gains than mixing the two data sources from the start. We additionally find that text conditioning degrades action following and physical consistency in action-conditioned world models, as generated dynamics may follow textual semantics rather than the causal effects of the specified actions. To mitigate this interference, we use fixed text prompts throughout training and evaluation: "A person performs a hand manipulation task." for human interaction samples and "A robot performs a manipulation task." for robot interaction samples. This keeps the textual condition consistent and encourages the model to focus on action-driven dynamics.

During GRPO training, we follow the efficiency-oriented setting adopted in FlowFactory (Ping et al., 2026), where effective optimization can be achieved with shorter clips, lower spatial resolution, and fewer denoising steps. Accordingly, for each sample, we generate four candidate rollouts at a resolution of 256×256 using 12 denoising steps, with SDE noise of strength 0.8 injected at the first step. The first observation frame is provided as the visual condition, from which each candidate predicts 40 future frames. To preserve sufficient interaction dynamics within this compact temporal window, we select an 81-frame action segment around each gripper state transition and uniformly subsample it to 40 action steps aligned with the predicted frames. This design retains the key motion and contact changes around gripper interactions while keeping GRPO training computationally efficient. We combine the action alignment and physical plausibility rewards with weights of 0.4 and 0.6, respectively.

B.2 Comparison with Baselines

As shown in Table 5, existing methods exhibit a clear performance drop on failed executions, with particularly pronounced degradation in IDM. This suggests that, under failure-inducing actions, these models remain biased toward the distribution of successful trajectories and tend to generate future states that resemble successful outcomes, thereby reducing fidelity to the conditioned actions. In contrast, HF-WM maintains closely comparable performance across successful and failed executions, indicating more consistent modeling of action consequences across different execution outcomes. (We use the Wan2.2 VAE to obtain the $\mathbf{L2}_{\text{latent}}$)

B.3 Ablation Studies

B.3.1 Ablation on PointFlow Feature Layers

To study the effect of where the auxiliary pointflow branch is attached, we compare intermediate video features extracted from different backbone depths as inputs to pointflow supervision. Specif-

Table 6: Ablation study on the feature extraction layers of the pointflow branch.(SFT)

Variant	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	$L2_{latent}$ ↓	IDM↓
mid [12,14,16,18]	25.756	0.886	0.087	15.163	16.031	0.201	0.377
back [20,22,24,26]	25.689	0.885	0.087	15.326	16.361	0.202	0.393
uniform [8,14,20,26]	25.761	0.886	0.086	15.047	16.211	0.201	0.357
front [4,6,8,10]	**25.828**	**0.887**	**0.086**	**14.993**	**15.991**	**0.198**	**0.355**

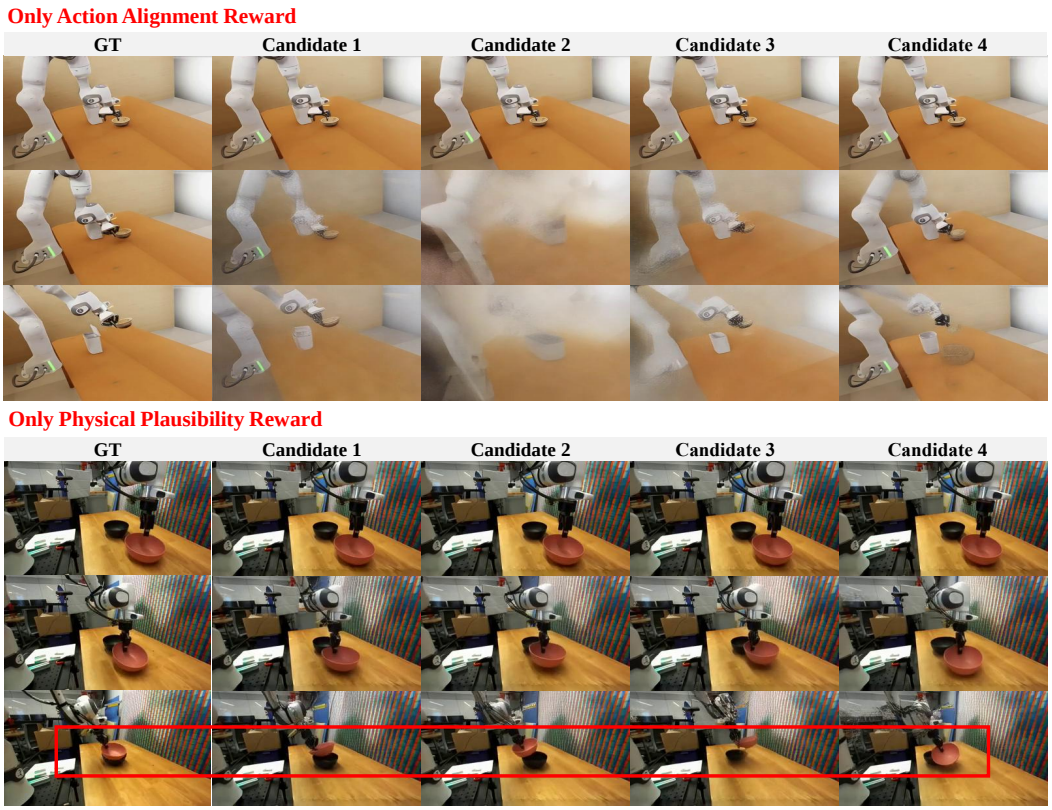


Figure 10: **Intermediate rollout examples during GRPO training with individual reward components.** Each panel shows corresponding frames from a ground-truth video and four candidate rollouts. Top: With only the action alignment reward, some candidates exhibit physically implausible interaction dynamics and degraded background quality. Bottom: With only the physical plausibility reward, the generated motion appears physically plausible but deviates from the conditioning actions. (**Note.** Variation across candidates arises from stochastic exploration during GRPO training. Their relatively lower visual fidelity is a consequence of the reduced denoising steps adopted for training efficiency.)

ically, we evaluate four configurations: front layers [4, 6, 8, 10], middle layers [12, 14, 16, 18], back layers [20, 22, 24, 26], and uniformly distributed layers [8, 14, 20, 26]. As shown in Table 6, the front-layer configuration performs best, followed by uniform sampling, while the middle- and back-layer variants show progressively lower performance. This trend suggests that pointflow supervision benefits more from shallower representations that retain fine-grained spatial structure and motion cues. Uniform sampling still captures complementary information across feature hierarchies, but deeper representations provide less direct motion information. We therefore use [4, 6, 8, 10] for the pointflow branch in our final model.

Table 7: Ablation study of GRPO on successful and failed outcomes.

Variant (Success/Failure)	PSNR↑	SSIM↑	LPIPS↓	FVD↓	FID↓	$L2_{latent}$ ↓	IDM↓
w/o GRPO	26.165/25.490	0.891/0.883	0.068/0.104	13.691/16.295	14.989/16.993	0.182/0.214	0.334/0.376
GRPO+action alignment	26.169/25.482	0.891/0.882	0.068/0.104	13.549/16.365	14.983/16.930	0.182/0.213	0.330/0.365
GRPO+physical plausibility	25.870/25.278	0.887/0.837	0.071/0.125	14.674/15.940	14.951/15.957	0.190/0.228	0.363/0.398
GRPO w/o action perturbations	26.319/25.434	**0.893**/0.878	**0.066**/0.106	13.381/16.143	14.321/15.736	0.179/0.221	**0.328**/0.374
GRPO (full)	**26.321/25.833**	0.892/**0.886**	**0.066/0.098**	**13.350/15.547**	**14.310/14.271**	**0.176/0.210**	**0.328/0.336**

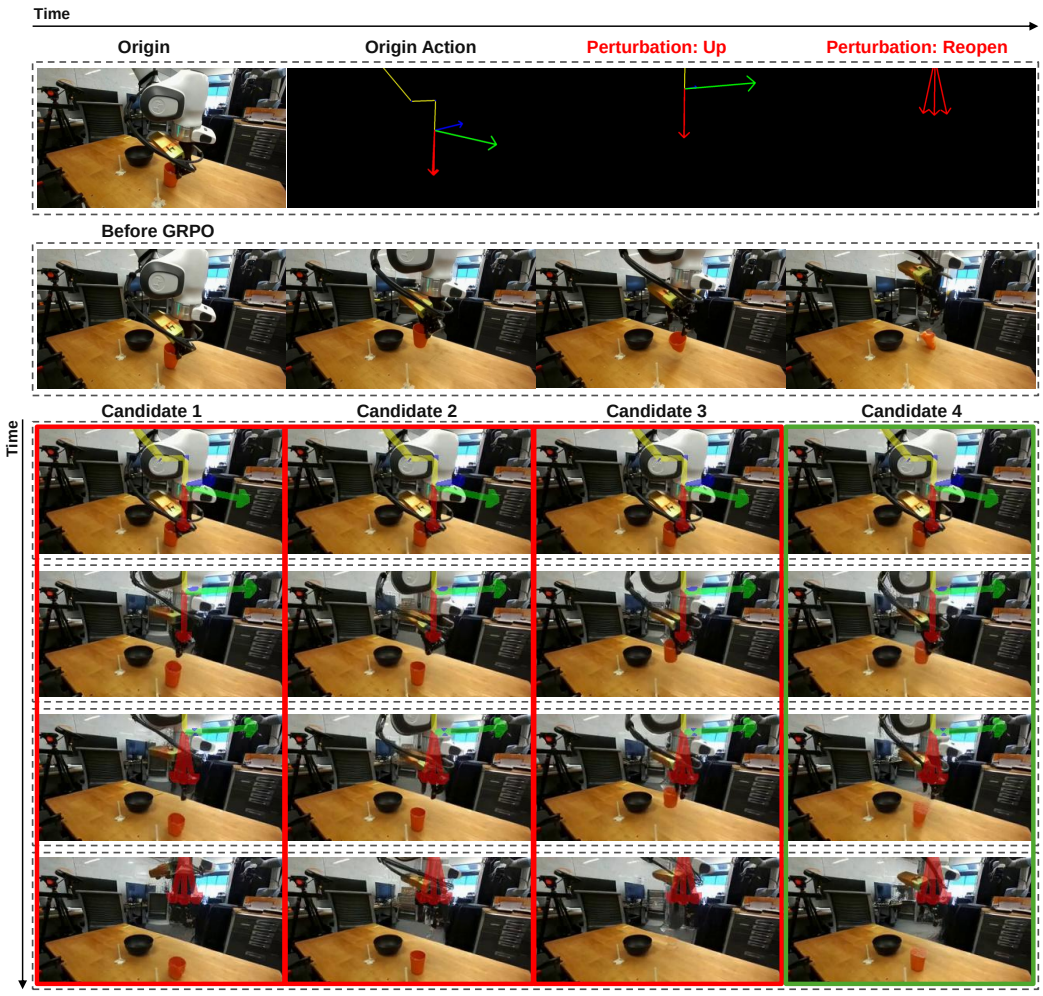


Figure 11: **Qualitative case of GRPO post-training under action perturbations.** Starting from the original action skeleton, we introduce two counterfactual perturbations, *Up* and *Reopen*. Before GRPO, the model fails to faithfully reflect the perturbed actions and produces physically implausible outcomes, such as the cup remaining unnaturally suspended instead of falling after release. After GRPO, the generated candidates better follow the perturbed actions and exhibit more physically plausible object dynamics.

B.3.2 Ablation on GRPO Rewards

Table 7 complements the aggregate evaluation in the main text by reporting each reward configuration separately on successful and failed executions. Either reward alone improves some metrics but does not consistently improve both generation quality and action following. In contrast, combining the two rewards improves performance on both subsets(Fig. 10). As shown in Fig. 11, GRPO yields more physically plausible predictions of object dynamics and failure outcomes under perturbed actions. These results suggest that reward-guided post-training helps address the limited coverage of failure dynamics in successful demonstrations.

C Downstream Task Performance Evaluation

C.1 Additional Analysis of Policy Evaluation

We complement Sec. 4.4 with an analysis of success-rate estimation and policy ranking consistency. Table 8 summarizes the results for $K = 6$ policies evaluated on $N = 90$ matched execution sessions. For policy i, the real-world and predicted success rates are $s_i = n_i/N$ and $\hat{s}_i = \hat{n}_i/N$, where n_i counts successful real-world executions and $\hat{n}_i$ counts successful outcomes identified by GPT-5 in the HF-WM-generated videos.

Pearson correlation measures the linear association between predicted and real-world success rates, while the mean absolute error SISR_Δ quantifies their numerical discrepancy:

$$\begin{aligned} r &= \frac{\sum_{i=1}^{K}(s_i-\mu_s)(\hat{s}_i-\mu_{\hat{s}})}{\sqrt{\sum_{i=1}^{K}(s_i-\mu_s)^2}\sqrt{\sum_{i=1}^{K}(\hat{s}_i-\mu_{\hat{s}})^2}} \approx 0.9826, \\ \mathrm{SISR}_\Delta &= \frac{100}{K}\sum_{i=1}^{K}|\hat{s}_i - s_i| \approx 2.41\,\mathrm{pp}, \end{aligned} \tag{13}$$

where $\mu_s = K^{-1}\sum_{i=1}^{K} s_i$ and $\mu_{\hat{s}} = K^{-1}\sum_{i=1}^{K}\hat{s}_i$ are the mean real-world and predicted success rates across policies, respectively, and pp denotes percentage points. These results show that the predicted success rates closely track real-world performance, with a mean absolute deviation of 2.41 percentage points.

To assess whether these estimates preserve relative policy performance, we rank policies by their real-world and predicted success rates in descending order. Let R_i and $\hat{R}_i$ denote the resulting ranks, with ties assigned their average rank. Spearman's ρ measures agreement between the rankings, whereas Kendall's τ_b measures pairwise ordering consistency with tie correction:

$$\begin{aligned} \rho &= \mathrm{Corr}\Big((R_i)_{i=1}^{K}, (\hat{R}_i)_{i=1}^{K}\Big) \approx 0.9559, \\ \tau_b &= \frac{C-D}{\sqrt{(C+D+T_s)(C+D+T_{\hat{s}})}} \approx 0.9286. \end{aligned} \tag{14}$$

Here, Corr denotes Pearson correlation; C and D count concordant and discordant policy pairs; and T_s and $T_{\hat{s}}$ count pairs tied only in the real-world or predicted success rates, respectively. Both rank correlations indicate strong agreement with the real-world policy ordering.

To account for the magnitude of the performance differences involved in ranking discrepancies, we further report the mean maximum rank violation (MMRV), which accounts for real-world performance gaps when measuring ranking errors. Following the official SIMPLER implementation (Li et al., 2024), we compute

$$\begin{aligned} v_{ij} &= |s_i - s_j|\,\mathbf{1}[(\hat{s}_i > \hat{s}_j) \neq (s_i > s_j)], \\ \mathrm{MMRV} &= \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i} v_{ij} \approx 0.002, \end{aligned} \tag{15}$$

where $\mathbf{1}[\cdot]$ denotes the indicator function and v_{ij} measures a pairwise ordering violation weighted by the real-world success-rate gap. MMRV averages the largest violation for each policy. In our evaluation, the nonzero value arises from a predicted tie between PG-FAST and PG-FSQ, whose real-world success counts differ by one. Together with the high rank correlations, this result indicates that HF-WM closely preserves relative policy performance. Additional policy evaluation results are provided in Figs. 16 and 17.

C.2 Test-Time Scaling for Policy Inference

We further evaluate the test-time scaling capability of HF-WM in downstream closed-loop manipulation. Experiments are conducted on all 90 tasks in LIBERO-90 using a shared $\pi_{0.5}$ policy as the base policy. To adapt HF-WM to the LIBERO domain (Liu et al., 2023), we first fine-tune it on the LIBERO training data. Following the training setup of SC3-Eval (Tseng et al., 2026), we use 24

Table 8: HF-WM policy evaluation on 90 matched sessions.

Pearson $r \uparrow$	Spearman $\rho \uparrow$	Kendall $\tau_b \uparrow$	$\text{SISR}_\Delta \downarrow$ (pp)	MMRV $\downarrow$
0.9826	0.9559	0.9286	2.41	0.002

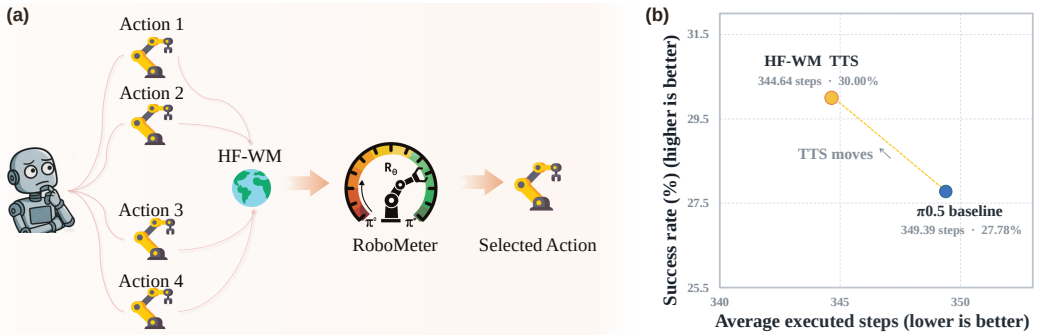


Figure 12: Test-time scaling with HF-WM. (a) Candidate action selection with HF-WM and RoboMeter. (b) Improved success rate and reduced execution steps on LIBERO-90.

frame video clips during fine-tuning to better preserve the long-horizon temporal prior of Cosmos 3 (Agarwal et al., 2026), while evaluating only the first 16 future steps at test time.

As shown in Fig. 12, at each decision step, we sample four candidate action chunks from the same $\pi_{0.5}$ (Physical Intelligence et al., 2025) policy and use HF-WM to predict their corresponding short-horizon futures. RoboMeter (Liang et al., 2026) scores the predicted outcomes, and the highest scoring candidate is selected for closed-loop execution. Both settings use the same $\pi_{0.5}$ policy and closed-loop execution protocol; the only difference is whether HF-WM prediction and RoboMeter scoring are used for test-time candidate selection.

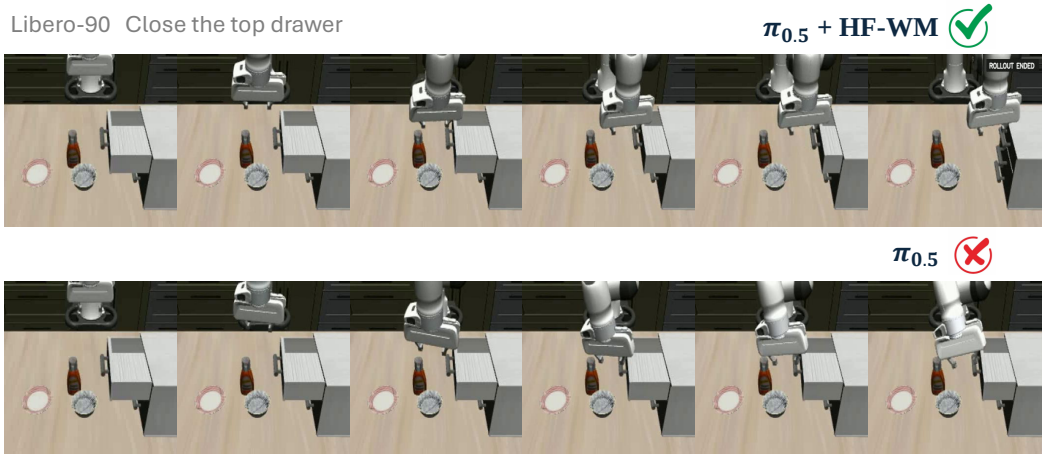


Figure 13: An example episode from the LIBERO-90 benchmark. HF-WM with RoboMeter routing reduces action steps and improves success rate.

HF-WM-based test-time selection improves the closed-loop success rate from 27.8% to 30.0%, while reducing the average episode length across the 90 tasks from 349.39 to 344.64 steps. These results show that HF-WM enables RoboMeter to identify more promising action candidates at test time, translating policy sampling diversity into measurable closed-loop gains with more efficient execution and higher task success.

C.3 Additional Visualization Results

We provide additional HF-WM rollouts with corresponding pointflow visualizations in Fig. 14 and qualitative comparisons with representative baselines in Fig. 15.

AI use statement

In this paper, large language models (LLMs) were used solely for grammar correction and readability improvement. They were not employed in any other part of the research process. In addition, Gemini was used to score the physical plausibility of generated videos during GRPO, while GPT-5 was used to assign task-success labels for policy evaluation, as described in Secs. A.2.2 and 4.4, respectively.

Ethics Statement

This work does not involve human subjects, animal experiments, or the use of personally identifiable or sensitive information. The research presents no known risks of harm, bias, discrimination, or misuse, and raises no concerns regarding privacy, security, legal compliance, or research integrity. The authors have no conflicts of interest or external sponsorship that could influence the research outcomes.

Reproducibility Statement

We have taken steps to ensure the reproducibility of our results. All experimental details and implementation choices are described in the main text and appendix. The complete source code will be made publicly available upon publication to facilitate reproduction and further research.

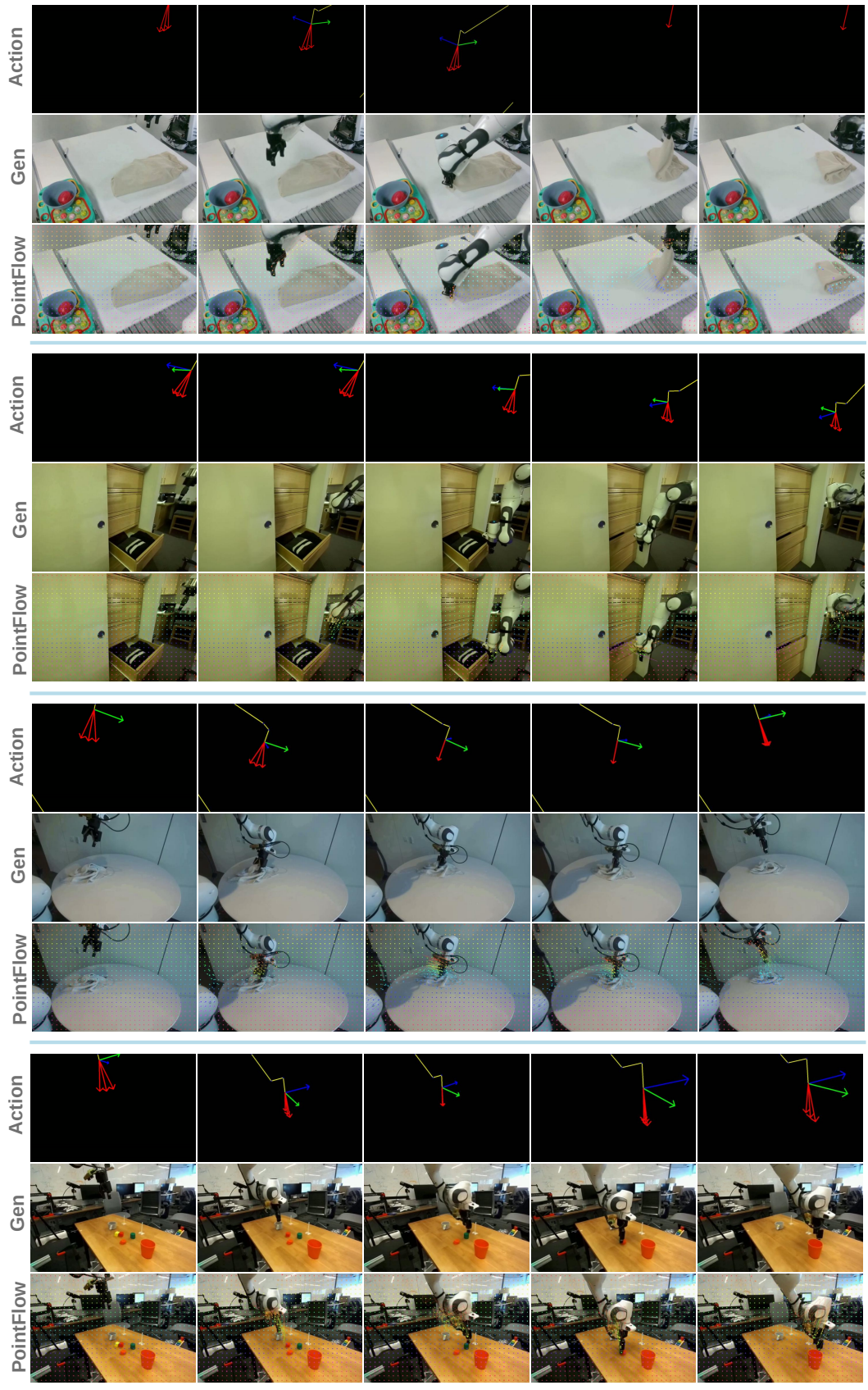


Figure 14: Additional qualitative results for HF-WM.



Figure 15: Additional qualitative comparison results

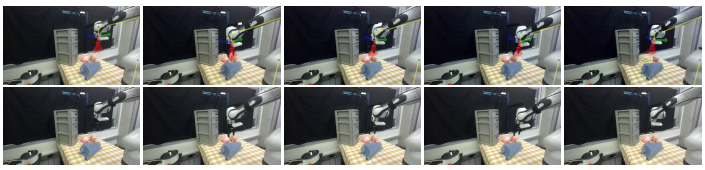
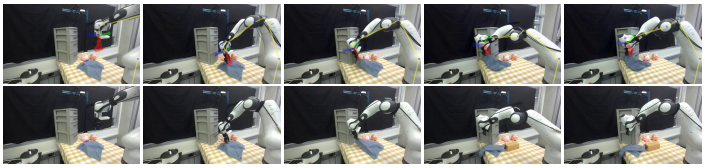
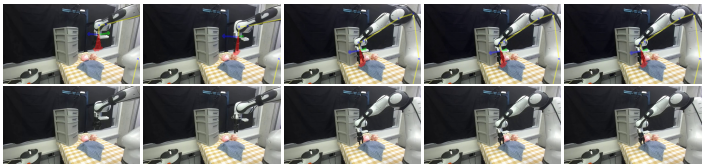
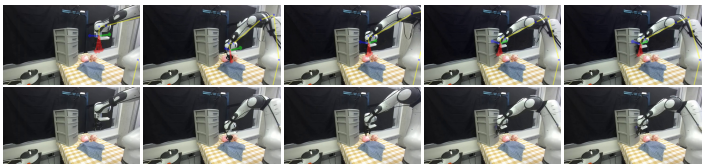
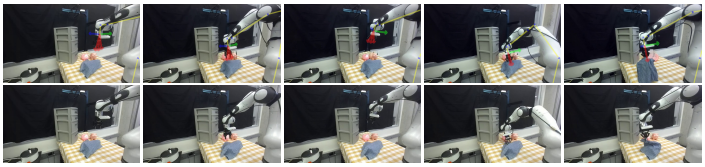
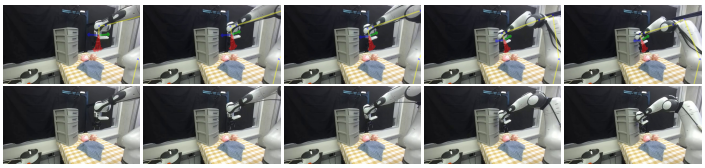
Policy	real	WM	time →
π_0-flow	✗	✗	
π_0-FAST	✓	✗	
PG-flow	✗	✗	
PG-FSQ	✗	✗	
PG-FAST	✗	✗	
PG-FAST+	✗	✗	

Figure 16: Uncover the wooden block. Each strip shows the world-model rollout (top) and the real-robot video (bottom) at six time steps. The *real* column is the RoboArena human success label; the WM column is the world-model verdict (Sec. 4.4). A red cell marks disagreement. On this task 1 of 6 policies succeeds on the real robot.

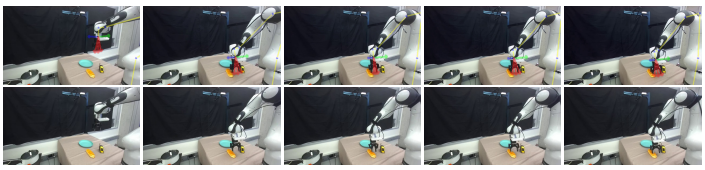
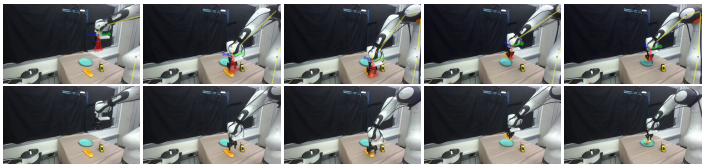
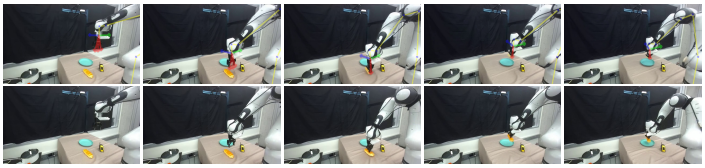
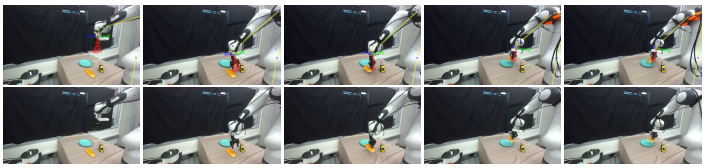
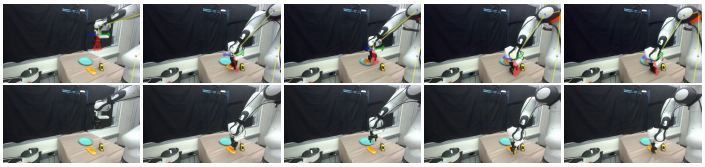
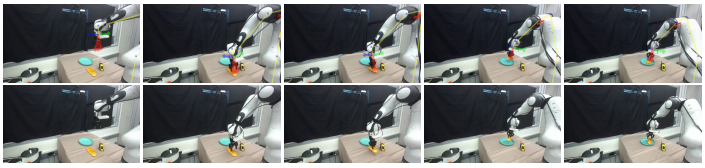
Policy	real	WM	time →
π_0-flow	✗	✗	
π_0-FAST	✓	✓	
PG-flow	✓	✓	
PG-FSQ	✓	✓	
PG-FAST	✓	✗	
PG-FAST+	✓	✓	

Figure 17: Put the food on the plate. Each strip shows the world-model rollout (top) and the real-robot video (bottom) at six time steps. The *real* column is the RoboArena human success label; the WM column is the world-model verdict (Sec. 4.4). A red cell marks disagreement. On this task 5 of 6 policies succeed on the real robot.