WHO ACTS, WHERE, AND HOW: UNIFIED WORLD MODEL WITH HETEROGENEOUS CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce a unified video world model for compositional control over viewpoint, embodiment, and action. Joint learning is hindered by inconsistent control representations and sparse paired supervision. We align partially annotated controls from 18 datasets spanning robot demonstrations, human egocentric recordings, and camera-motion videos through reference-camera geometry, a unified action template, and a shared action encoder. This provides consistent physical semantics for learning from complementary sources. Visual context supplies embodiment cues without requiring explicit identifiers at inference. A hybrid Mixture-of-Transformers and Mixture-of-Experts architecture couples pretrained vision-language conditioning with noise-specialized generation. We pretrain with supervised image-to-video learning and cross-segment self-supervised video prediction, then adapt the high-noise expert using limited cross-view pairs for video re-shooting. The low-noise expert remains frozen, preserving refinement learned from heterogeneous pretraining. Experiments show strong visual fidelity and control accuracy in human and robot action-conditioned generation, camera-controlled generation, and video re-shooting. Qualitative examples further illustrate compositional action–camera control and human-to-robot transfer.

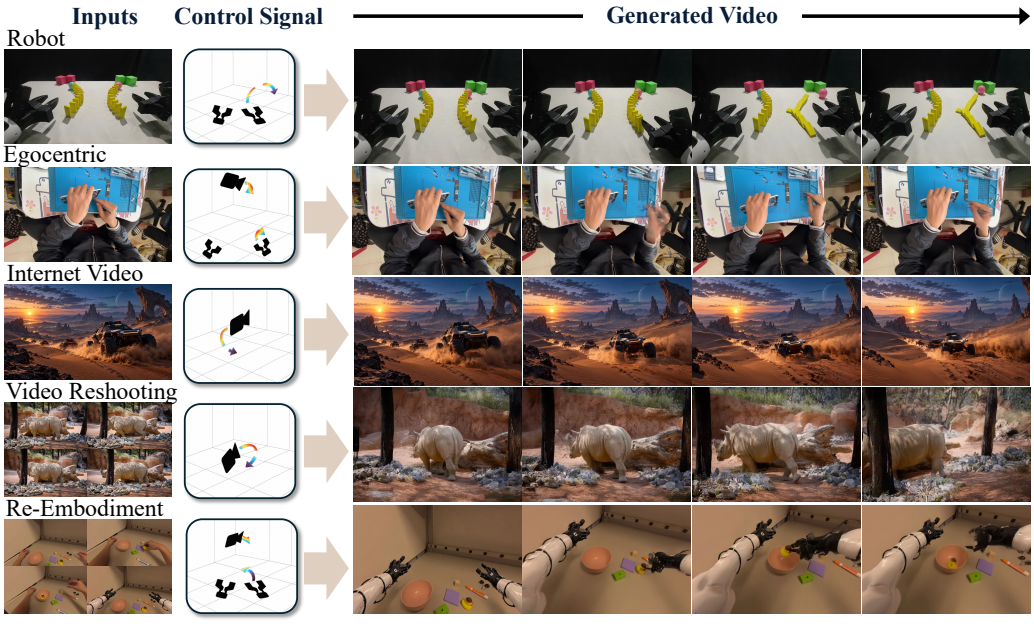

Figure 1: Compositional control over action, viewpoint, and embodiment in a unified video world model. For image-to-video (I2V) generation, we demonstrate action control, camera control, and action–camera control across varied scenarios. For video-to-video (V2V) generation, we demonstrate camera-controlled video re-shooting and embodiment transformation from human to robot.

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

Interactive video world models extend video generation beyond passive content creation toward interactive simulation, with applications spanning cinematography, advertising, autonomous driving, and embodied artificial intelligence (Yang et al., 2023; Hu et al., 2023; Bruce et al., 2024). One line of research focuses on controlling camera trajectories while maintaining the visual quality and plausibility of generated videos (Wang et al., 2024c; He et al., 2024; Ren et al., 2025). A complementary line of work studies action-conditioned world models, which predict the visual consequences for a predefined acting body under a fixed viewpoint, conditioned on its action trajectory (Finn et al., 2016; Ebert et al., 2018; Zhu et al., 2025). Despite their complementary roles in interactive simulation, these two directions are still largely studied separately.

We argue that these two directions should be studied jointly. An interaction's visual outcome depends on who acts, from where it is observed, and how it unfolds—namely, embodiment, viewpoint, and action. This motivates an interactive world model with independently specified, composable controls over these factors. Such a model could capture realistic appearances and complex contact dynamics, including deformable-object interactions that challenge conventional simulators, enabling scalable embodied evaluation and closed-loop learning (Ebert et al., 2018; Guo et al., 2026).

Achieving such compositional control, however, poses a fundamental data challenge. Training a model with flexible control over embodiment, viewpoint, and action requires broad coverage of their combinations, yet collecting and annotating such data in the real world is prohibitively expensive (O'Neill et al., 2024; Khazatsky et al., 2024). Human egocentric videos and a curated subset of large-scale Internet videos characterized by substantial camera motion offer a scalable source of such experience. Specifically, we therefore aggregate training data from 18 datasets spanning three complementary sources of experience. Internet videos, together with the real-world, synthetic and game-engine videos, provide diverse scenes, appearances, and camera motions (Ju et al., 2024; Li et al., 2025d; Zhou et al., 2026); human egocentric videos capture rich human–environment interactions (Grauman et al., 2022; 2024; Hoque et al., 2026); and robot demonstrations provide explicit action trajectories (Khazatsky et al., 2024; Bu et al., 2025). Although each source offers only partial supervision, together they broaden the coverage of visual and physical experience while reducing the need for dedicated data collection.

Realizing these complementary benefits requires addressing substantial heterogeneity across sources. These datasets differ in temporal sampling rates, motion distributions, embodiment morphology, action parameterizations, coordinate systems, and camera conventions (O'Neill et al., 2024; Doshi et al., 2024). Beyond the obvious heterogeneity across robot platforms, human data are themselves highly heterogeneous. Different individuals vary substantially in body shape, proportions, limb lengths, appearance, and kinematics, which in turn induce large variations in egocentric viewpoints and motion patterns. Robot data introduce another layer of heterogeneity, as different platforms may encode actions using joint positions, joint deltas, end-effector poses, Cartesian commands, velocities, or gripper states (Wang et al., 2024a). In particular, geometric inconsistencies and embodiment-specific action encodings can assign inconsistent numerical representations to physically corresponding camera and effector motions (Zheng et al., 2025b; Wang et al., 2024b; Yuan et al., 2026). Naively pooling these data can induce cross-domain interference and limit the benefits of increased scale. Prior efforts toward cross-embodiment learning have made important progress through shared architectures, embodiment-aware conditioning, unified action tokenization, and learned action representations (Team et al., 2024; Kim et al., 2024; Wang et al., 2024a), but geometric inconsistencies can persist. We therefore align the control signals within a shared physical representation, reducing the need for the model to reconcile dataset-specific control conventions and facilitating knowledge transfer across heterogeneous datasets.

We introduce a control alignment framework with three components. a) a unified geometric parameterization expresses camera and action trajectories relative to a single canonical 3D camera-frame coordinate system defined by a selected reference video frame, while a unified action template accommodates different embodiment configurations. b) an action encoder shared across datasets and embodiments replaces embodiment-specific encoders. Together, the geometric parameterization and shared encoder establish consistent spatial semantics and promote consistent encoding of corresponding motions across embodiments. c) and an in-context mechanism serves as an adaptive embodiment identifier. We use the initial visual observation or a reference video as the adaptive context, which implicitly specifies the target embodiment, allowing a fully shared model predic-

tions to adapt to embodiment-specific kinematics without embodiment-specific action encoders or identifiers.

Building on the aligned control representation, we employ a hybrid architecture that combines a Mixture of Transformers (MoT) (Liang et al., 2025; Agarwal et al., 2026) with a Mixture of Experts (MoE) (Balaji et al., 2022; Wan et al., 2025). The MoT component uses unidirectional attention to reduce interference from the generative objective with the pretrained perception and reasoning representations of the vision-language model (VLM) backbone. The MoE component introduces separate generation experts for different denoising regimes, motivated by the different roles of diffusion stages: high-noise denoising primarily establishes coarse spatiotemporal structure and motion, whereas later stages progressively refine appearance and visual details. Most training videos depict an interaction performed by a single embodiment and observed along a single camera trajectory, leaving paired supervision across embodiments or viewpoints scarce. We therefore adopt a two-stage training strategy. First, large-scale supervised and cross-segment self-supervised pretraining learns from the full heterogeneous video collection without requiring paired observations across embodiments or viewpoints. Second, a limited set of paired videos is used to adapt only the high-noise expert, concentrating the available paired supervision on the denoising regime primarily responsible for global structure and motion while leaving the remaining generation pathway largely unchanged. The resulting model supports flexible conditioning on embodiment, viewpoint, and action from a reference image or video, enabling source-to-target generation tasks such as video re-shooting (Bai et al., 2025; Paliwal et al., 2026) and the transformation of human interaction videos into robot-native videos (Chen et al., 2026b).

We evaluate our model against state-of-the-art baselines on robot and human action-conditioned generation using Agibot-World-Beta (Bu et al., 2025) and EgoDex (Hoque et al., 2026), camera-controlled image-to-video generation using a multi-dataset suite including RealCam-Vid (Zheng et al., 2025a) and OmniWorld (Zhou et al., 2026), and video re-shooting using MultiCamVideo (Bai et al., 2025). Our model achieves the best performance on most evaluated metrics across these tasks, demonstrating its ability to combine high-fidelity visual prediction with accurate action and camera control within a unified framework.

Our contributions are threefold:

- We introduce an interactive video world model with compositional control over viewpoint, action, and visually specified embodiment, without requiring explicit embodiment identifiers at inference.
- We align heterogeneous controls through a unified geometric parameterization, a common action template, and a shared action encoder, enabling joint learning across datasets and embodiments.
- We develop a two-stage training strategy that combines supervised and cross-segment self-supervised pretraining with selective high-noise adaptation on limited paired videos, retaining pretrained low-noise refinement for video re-shooting.

2 METHOD

Our method learns compositional control over viewpoint, embodiment, and action from heterogeneous videos with partial supervision. We formulate a common prediction task (Sec. 2.1), align control semantics across sources (Sec. 2.2), and train a shared MoT–MoE model through single-video pretraining and paired adaptation (Sec. 2.3).

2.1 PROBLEM FORMULATION

We model the distribution of a target video conditioned on visual context, a text instruction, and available geometric controls:

$$p_\theta\big(V^{\mathrm{tgt}} \mid V^{\mathrm{ctx}}, y, S_0, C_{1:T}, A_{1:T}\big)\,, \tag{1}$$

where V^{ctx} denotes visual context, y is the instruction, and S_0 specifies the target's camera and effector state at $t = 0$. The sequences $C_{1:T}$ and $A_{1:T}$ describe the desired camera motion and embodied actions over horizon T. Visual context supplies appearance and embodiment cues, with video context additionally conveying motion and interaction history.

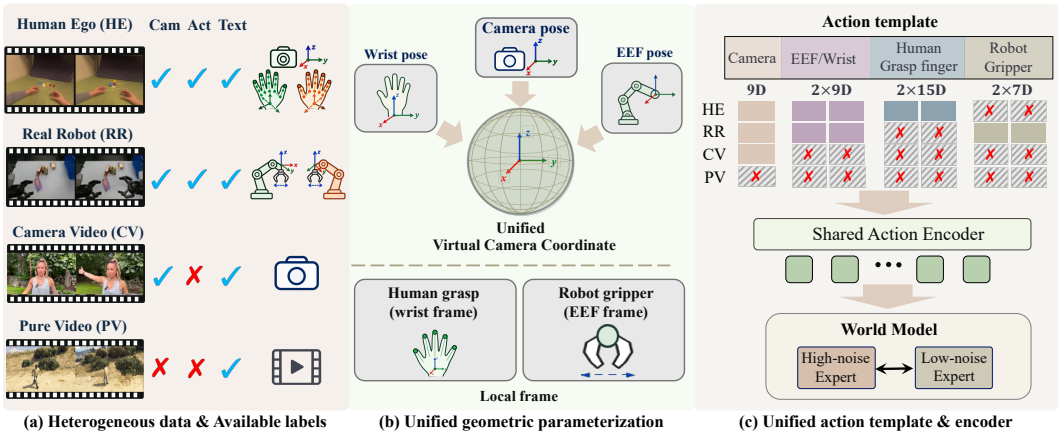

Figure 2: Control alignment across heterogeneous sources. (a) Data sources expose different subsets of camera, action, and text annotations. (b) Rigid poses use reference-camera coordinates, while fingertip and gripper configurations use local effector frames. (c) A fixed action template represents available controls and missing fields, and a shared action encoder maps it to control tokens.

2.2 Control Alignment across Heterogeneous Sources

To turn complementary but heterogeneous annotations into shared supervision, we align both their physical meaning and their encoding. Fig. 2 summarizes the available labels, unified geometric parameterization, and shared action template and encoder.

Heterogeneous data and available labels. Our 18 datasets provide complementary subsets of camera, action, and text annotations (Fig. 2(a)). Human egocentric recordings and robot demonstrations provide interaction trajectories, camera-annotated videos contribute viewpoint variation, and videos without geometric annotations broaden visual coverage. Even observed controls are not directly comparable: robot actions may encode joint states, Cartesian poses, increments, or velocities, while human motion may be reconstructed in moving camera coordinates. We recover camera and effector trajectories using the available kinematics, calibration, and motion reconstruction, interpreting each annotation under its source conventions. The recovered trajectories provide a common basis for geometric alignment, while unavailable control signals remain explicitly marked as missing.

Unified geometric parameterization. The initial state S_0 expresses the target's initial rigid poses in the source first-frame camera coordinates. Future controls encode motion relative to S_0. After harmonizing units, axes, and temporal sampling, we express camera and effector trajectories in a virtual camera frame anchored at the first target frame (Fig. 2(b)). For target camera- and effector-to-world transforms $\mathbf{T}_t^{\text{cam}}, \mathbf{T}_t^{\text{eff}} \in \mathrm{SE}(3)$, the aligned poses are:

$$\bar{\mathbf{T}}_t^{\text{cam}} = (\mathbf{T}_0^{\text{cam}})^{-1}\mathbf{T}_t^{\text{cam}}, \qquad \bar{\mathbf{T}}_t^{\text{eff}} = (\mathbf{T}_0^{\text{cam}})^{-1}\mathbf{T}_t^{\text{eff}}. \tag{2}$$

The camera trajectory encodes motion from the initial target camera. For each effector, let $\mathbf{R}_t$ and $\mathbf{x}_t$ denote the rotation and translation of $\bar{\mathbf{T}}_t^{\text{eff}}$. We encode motion relative to its initial state as:

$$\Delta\mathbf{x}_t = \mathbf{x}_t - \mathbf{x}_0, \qquad \Delta\mathbf{R}_t = \mathbf{R}_t\mathbf{R}_0^\top, \quad t > 0. \tag{3}$$

Translation and rotation changes are thus expressed along the same reference-camera axes across embodiments. Rotations use a continuous six-dimensional representation (Zhou et al., 2019).

We represent fingertip and gripper contact positions in their local wrist/end-effector frames, separating articulation from rigid motion. Their initial configuration is included in S_0, and subsequent controls describe changes relative to this configuration. Gripper opening is represented as an absolute distance. This decomposition gives camera and effector motion consistent spatial semantics while retaining embodiment-specific articulation.

Unified action template and encoder. A common coordinate system aligns motion geometry, but embodiments still differ in articulation and sources provide different subsets of control annotations.

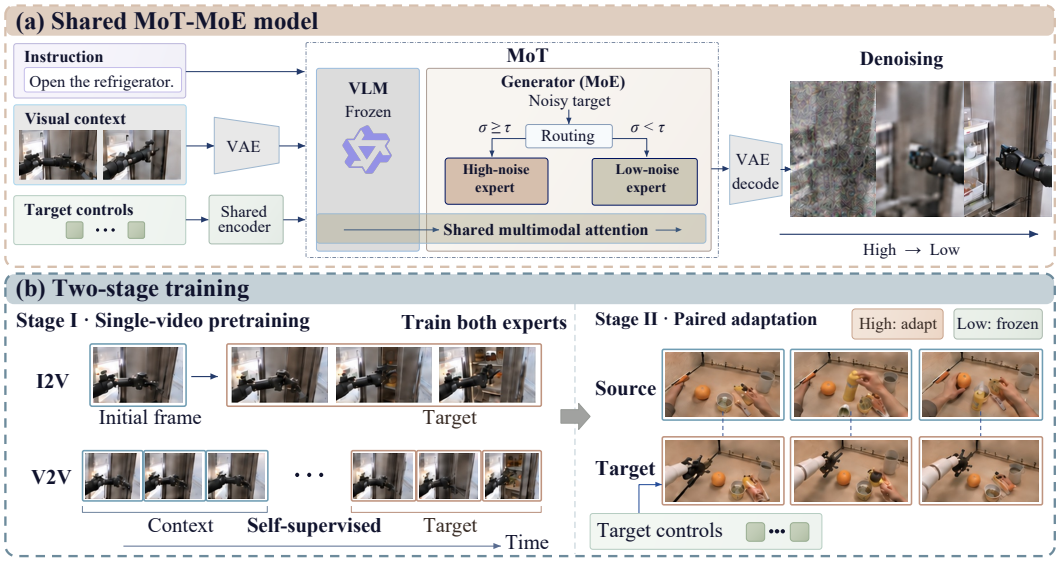

Figure 3: Shared model and two-stage training. (a) A frozen VLM conditions noise-routed generation experts through shared multimodal attention. Sampling proceeds from high to low noise. (b) Single-video I2V/V2V pretraining trains both experts. Paired adaptation trains the high-noise expert for source-to-target video re-shooting while keeping the low-noise expert frozen.

We address both through a unified action template $\mathcal{U}_t$, which assigns camera and effector motions to fields with fixed physical meanings while accommodating embodiment-specific local configurations (Fig. 2(c)). Unavailable fields use learned null embeddings, allowing partially annotated videos to share the same control interface without conflating missing information with stationary motion.

We encode the aligned template with a single action encoder shared across datasets and embodiments, allowing complementary supervision to shape a common control representation:

$$\mathbf{h}_t = E_{\text{shared}}\big(\mathcal{U}_t; \mathbf{e}^{\text{emb}}\big) + \mathbf{e}_t^{\text{role}}. \tag{4}$$

Here, $\mathbf{e}_t^{\text{role}}$ distinguishes the initial state from subsequent controls. To support large-scale learning from videos whose initial views only partially reveal embodiment morphology, we learn an auxiliary embedding $\mathbf{e}^{\text{emb}}$ for each training embodiment type. During training, we randomly replace this cue with a shared learned unknown embedding $\mathbf{e}_{\text{unk}}$, encouraging prediction from visual context and aligned controls. At inference, we always set $\mathbf{e}^{\text{emb}} = \mathbf{e}_{\text{unk}}$, so the initial observation serves as an adaptive embodiment identifier without requiring new embeddings for unseen embodiments.

2.3 Hybrid Architecture and Two-Stage Training

Control alignment makes heterogeneous videos compatible with a shared model, but individual recordings do not provide the correspondences needed to directly supervise source-to-target transformations. We therefore pretrain on single videos and adapt on limited synchronized pairs. The MoE architecture supports this strategy by separating the high-noise expert that learns paired transformations from the low-noise expert that retains pretrained refinement capabilities (Fig. 3).

Shared MoT–MoE model. The architecture couples pretrained multimodal conditioning with generation at different noise levels (Fig. 3(a)). A Mixture of Transformers (MoT) connects a frozen vision-language model (VLM) to the video generation experts through shared multimodal attention. Generation tokens attend to VLM features, while VLM tokens do not attend to noisy target tokens. This makes pretrained representations available to generation while keeping the VLM's conditioning features independent of the denoising state. A Mixture of Experts (MoE) routes generation to separate high- and low-noise experts (Balaji et al., 2022; Wan et al., 2025):

$$\mathbf{v}_\theta(\mathbf{z}_\sigma, \sigma \mid \mathcal{C}) = \begin{cases} \mathbf{v}_{\theta_{\text{hi}}}(\mathbf{z}_\sigma, \sigma \mid \mathcal{C}), & \sigma \geq \tau, \\ \mathbf{v}_{\theta_{\text{lo}}}(\mathbf{z}_\sigma, \sigma \mid \mathcal{C}), & \sigma < \tau. \end{cases} \tag{5}$$

Here, θ_{hi} and θ_{lo} parameterize the two velocity prediction networks, $\sigma \in [0, 1]$ is the noise level, and τ is the routing threshold. The conditions $\mathcal{C}$ comprise the instruction y, VAE-encoded visual context V^{ctx}, and the shared control tokens $\mathbf{h}_{0:T}$. The visual latents $\mathbf{z}_\sigma$ contain clean conditioning frames and noisy prediction targets, as defined below.

High-noise denoising relies on the supplied conditions to establish global scene structure and motion, whereas low-noise denoising refines the emerging video's appearance and details. We use this specialization as a mechanism for selective adaptation. Limited paired supervision trains the high-noise expert to infer the target video from source context and target controls, while the low-noise expert retains the refinement learned from large-scale pretraining.

Two-stage training. **Stage I: Single-video pretraining.** We train both experts using supervised I2V generation and cross-segment self-supervised V2V prediction (Fig. 3(b)). Supervised I2V learns the visual consequences of annotated camera motion and actions, conditioned on an initial image. Self-supervised V2V uses an earlier segment to predict a later segment of the same video, with temporal gaps allowed. The recording itself supplies the context–target relation, enabling learning from motion and interaction history without paired recordings across viewpoints or embodiments. The two tasks use available text and controls, with null embeddings for missing controls. Each expert trains on the same task and data mixture within its assigned noise range. Nonvisual conditions are randomly dropped for classifier-free guidance.

We use a unified flow-matching objective across both training stages, with corruption and supervision restricted to the prediction target (Lipman et al., 2023). Let $\mathbf{z}_0$ contain the clean visual latents of the context and target frames. A binary mask $\mathbf{M}$ of the same shape selects the elements to generate and is zero on the visual context. For $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the target velocity and conditional flow path are:

$$\mathbf{v}^\star = \mathbf{M} \odot (\boldsymbol{\epsilon} - \mathbf{z}_0), \qquad \mathbf{z}_\sigma = \mathbf{z}_0 + \sigma \mathbf{v}^\star. \tag{6}$$

Here, $\odot$ denotes elementwise multiplication. The mask restricts noise corruption to the prediction target, keeping the visual conditioning frames clean at every noise level. For expert $e \in \{\text{hi}, \text{lo}\}$, let q_e denote its noise sampling distribution, restricted to $[\tau, 1]$ for the high-noise expert and $[0, \tau)$ for the low-noise expert. We optimize:

$$\mathcal{L}_e(\theta_e; \mathcal{D}) = \mathbb{E}_{\substack{(\mathbf{z}_0, \mathcal{C}, \mathbf{M}) \sim \mathcal{D} \\ \sigma \sim q_e,\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})}} \left[\frac{\| \mathbf{M} \odot (\mathbf{v}_{\theta_e}(\mathbf{z}_\sigma, \sigma \mid \mathcal{C}) - \mathbf{v}^\star) \|_2^2}{\|\mathbf{M}\|_1} \right]. \tag{7}$$

The loss averages velocity errors over generated elements only, with $\|\mathbf{M}\|_1$ counting those elements. The conditions $\mathcal{C}$ and target mask $\mathbf{M}$ accommodate the different I2V and V2V settings within the same objective, while q_e matches each expert to its denoising regime. For the heterogeneous pretraining dataset $\mathcal{D}_{\text{pre}}$, Stage I minimizes $\sum_{e \in \{\text{hi,lo}\}} \mathcal{L}_e(\theta_e; \mathcal{D}_{\text{pre}})$, updating both experts.

Stage II: Paired adaptation. A small paired dataset $\mathcal{D}_{\text{pair}}$ provides synchronized source–target videos for video re-shooting. We reuse the V2V interface with the source video as clean visual context and the entire target video as the prediction target, conditioned on the aligned target controls. Accordingly, $\mathbf{M}$ selects all target video latents and excludes the source context.

We adapt the high-noise expert by minimizing $\mathcal{L}_{\text{hi}}(\theta_{\text{hi}}; \mathcal{D}_{\text{pair}})$, while keeping θ_{lo} and the VLM parameters frozen. At inference, the adapted high-noise expert initiates generation, the pretrained low-noise expert refines the prediction below τ, and the VAE decoder produces the re-shot video.

3 Experiments

We evaluate the model's action and camera control capabilities in both image-to-video (I2V) and video-to-video (V2V) settings. In I2V, we assess action-conditioned prediction for robot and human manipulation, as well as generation conditioned on target camera trajectories. In V2V, we evaluate camera-controlled video re-shooting. We further illustrate compositional action–camera control and human-to-robot transformation through qualitative examples and ablate the contributions of cross-segment self-supervision, the shared action encoder, and noise-specific experts.

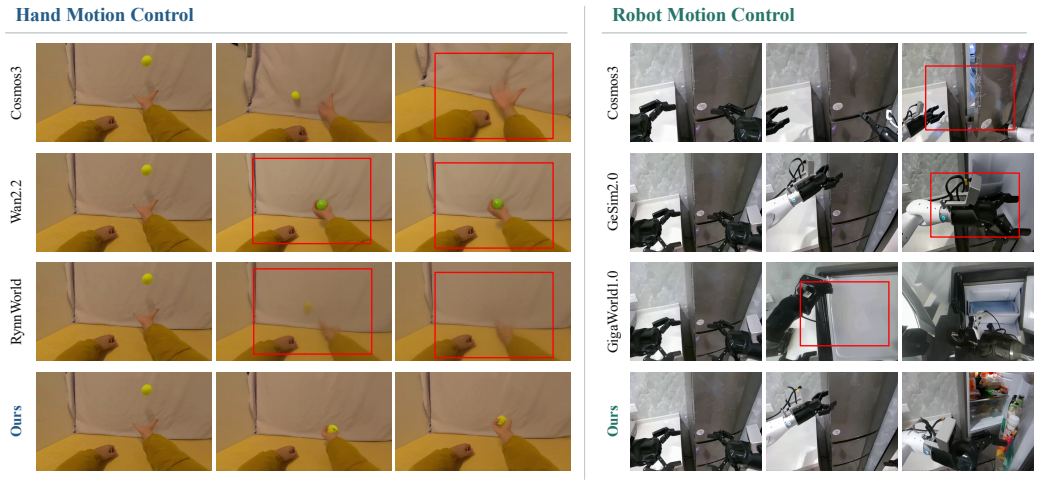

Figure 4: Qualitative comparisons of action-conditioned video generation on EgoDex (left) and AgiBot (right). Our model better maintains hand structure and coherent hand–object interactions during human manipulation, and preserves object appearance and spatial consistency during robot manipulation. Red boxes highlight inconsistencies in baseline results.

3.1 Experimental Setup

Implementation and evaluation data. We initialize the video backbone from Cosmos3-Nano 16B (Agarwal et al., 2026), with Qwen3-VL-8B-Instruct as the text backbone and a frozen Wan2.2 VAE, and introduce separate parameter banks for high- and low-noise denoising. In Stage I, we train the generation tower and action encoder on heterogeneous action-, camera-, and unannotated videos while freezing the VAE and text-understanding pathway. In Stage II, we adapt the high-noise expert using synchronized cross-view pairs from MultiCamVideo and Ego-Exo4D, while keeping the low-noise expert fixed. We evaluate four held-out test suites: 451 robot action-conditioned samples from Agibot-World-Beta (Bu et al., 2025), 285 human action-conditioned samples from EgoDex (Hoque et al., 2026), 300 camera-controlled I2V samples from RealCam-Vid (Zheng et al., 2025a), Omni-World (Zhou et al., 2026), and four additional video datasets, and 150 MultiCamVideo (Bai et al., 2025) re-shooting scenarios, each evaluated under three camera trajectories. All evaluation episodes are excluded from both training stages.

Evaluation protocol. We compare against publicly released baselines using the conditioning modalities supported by each method, with ground-truth action, skeleton, or camera controls provided where applicable. For action-conditioned generation, we adapt the WorldArena protocol (Shang et al., 2026) to evaluate visual quality, content consistency, motion quality, physics adherence, 3D accuracy, and controllability. For camera-controlled generation, we report RotErr, TransErr, and CamMC following CamI2V (Zheng et al., 2024), together with ATE, PSNR, SSIM, CLIP, FID (Heusel et al., 2017), FVD (Unterthiner et al., 2018), image quality, and aesthetic quality.

3.2 Action-Conditioned Generation

Egocentric human actions. The comparison on EgoDex (Tab. 1) includes Rynn-World (Zhao et al., 2026) trained on EgoDex dataset, alongside Wan2.2 (Wan et al., 2025) and Cosmos 3 (Agarwal et al., 2026), two general models based on much broader data. Our model achieves the strongest interaction-quality score and further leads in trajectory accuracy and instruction following, suggesting that the predicted futures are better aligned with both hand actions and language instructions. These gains over both domain-specific and general models indicate that the shared representation transfers effectively to human-centric action conditioning.

Robot actions. On robot manipulation tasks (Tab. 1), GigaWorld-1 (Team et al., 2026b) and GESim 2.0 (Qiu et al., 2026) are strong baselines that were trained on Agibot-World-Beta, with GESim2.0 further distilled for the Genie-01 embodiment. Against these baselines, our model achieves the strongest visual quality and content consistency, as well as the best controllability score. Moreover, it surpasses the specialized GESim2.0 on trajectory accuracy and depth accuracy, indicating that the

Table 1: Action-conditioned generation on robot and human manipulation tasks.

Method	Visual Quality				Content Consistency		Motion Quality		Physics Adherence		3D Accuracy	Controllability	
	PSNR↑	SSIM↑	Image quality↑	Aesthetic quality↑	Background consistency↑	Subject consistency↑	Photometric smoothness↑	Motion smoothness↑	Interaction quality↑	Trajectory accuracy↑	Depth accuracy↑	Instruction following↑	Semantic alignment↑
Ego-centric													
Wan2.2	17.32	0.83	67.9	**44.6**	15.2	14.5	23.3	42.2	54.7	17.1	95.2	43.3	81.8
Rynn-World	22.50	0.88	54.9	37.0	32.9	32.2	44.9	36.5	79.3	61.5	96.3	62.3	82.4
Cosmos3	18.30	0.79	59.8	38.7	**75.5**	66.3	48.2	39	82.3	23.6	93.2	69.0	82.4
Ours	**23.73**	**0.88**	**68.3**	39.9	68.7	**70.1**	**52.3**	**44.9**	**92.3**	**62.2**	**97.7**	**76.1**	**84.1**
Robotics													
GigaWorld-1	17.84	0.63	33.7	19.1	15.3	14.9	19.9	**63.8**	44.7	48.6	44.2	57.4	43.2
GESim2.0	19.81	0.77	72.1	44.7	20.3	20.5	37.7	47.9	**71.5**	88.4	98.3	**75.5**	90.0
Ours	**21.69**	**0.82**	**74.2**	**47.7**	**38.5**	**38.4**	**47.9**	45.1	69.5	**88.9**	**98.7**	75.0	**92.4**

Table 2: Camera-controlled image-to-video generation.

Method	Camera Control				Semantic and Distribution			Visual Quality			
	RotErr↓	TransErr↓	CamMC↓	ATE↓	CLIP↓	FID↓	FVD↓	PSNR↑	SSIM↑	Image quality↑	Aesthetic quality↑
Wan2.2	0.082	0.180	0.239	0.207	28.81	24.49	890.45	13.74	0.435	66.89	50.22
SCoPE	0.125	0.185	0.283	0.217	30.43	20.50	831.76	14.00	0.432	67.23	54.67
Cosmos 3	0.106	0.293	0.356	0.337	29.66	19.01	774.72	14.83	0.455	67.43	52.35
DreamX	0.091	0.183	0.238	0.251	**28.47**	26.49	961.64	13.49	0.422	64.38	50.00
HY-WorldPlay	0.065	0.182	0.220	0.217	30.59	18.59	707.08	14.42	0.422	65.67	**54.97**
Ours	**0.051**	**0.177**	**0.205**	**0.207**	29.15	**15.63**	**551.28**	**15.89**	**0.486**	**67.45**	52.51

unified model can follow robot actions while preserving the underlying scene geometry. The result suggests that our model provides a stronger balance between action execution and visual prediction.

Compositional action-camera control. We qualitatively illustrate compositional control by varying camera trajectories under a fixed action trajectory, and varying action trajectories under a fixed camera trajectory. Qualitative comparisons of video generation results are provided in App.Fig. 5 .

3.3 Camera-Controlled Generation

For camera-controlled I2V generation, our model combines lower camera-motion errors with improved reconstruction and distributional quality (Tab. 2). We compare with the camera-control model SCoPE (Yin et al., 2026) and general-purpose models Wan2.2 (Wan et al., 2025), Cosmos 3 Agarwal et al. (2026), DreamX-World (Team et al., 2026a), and HY-WorldPlay (Sun et al., 2025). Our model leads in camera control and distributional quality and remains competitive elsewhere, demonstrating strong overall control and generation quality, although HY-WorldPlay leads in aesthetics. Qualitative results appear in App.Fig. 6.

For video re-shooting, our model improves target-camera tracking and target-video reconstruction over ReCamMaster (Bai et al., 2025), TrajectoryCrafter (Yu et al., 2025b), and Vista4D (Lin et al., 2026) (Tab. 3). It leads all four camera-control metrics, following the target trajectory more accurately while retaining the reference video as conditioning. It also leads all reported visual and distributional metrics, including the lowest FID/FVD and highest PSNR/SSIM, though ReCamMaster performs slightly better on CLIP. Qualitative results appear in App.Fig. 7. ReCamMaster performs slightly better on the CLIP metric, while our model leads all four camera-control metrics and the remaining reported quality metrics. Video generation results are provided in App.Fig. 7.

Human-to-robot transformation. App.Fig. 8 presents qualitative examples of transforming human interaction videos into robot videos.

3.4 Ablation Studies

We assess the effects of noise-specific experts, shared action encoding, and cross-segment self-supervision on generation quality and control accuracy. All variants share preprocessing and reference-camera alignment, so the encoder ablation isolates cross-embodiment parameter sharing under aligned controls.

Effect of Noise-Specific Experts. Adding NS-Expert to the base model improves all four camera-control metrics, reducing RotErr from 0.115 to 0.057 and FVD from 672.67 to 536.47 (Tab. 4). For robot- and human-conditioned generation, it improves PSNR, SSIM, and trajectory accuracy, with

Table 3: Camera-controlled video re-shooting.

Method	Camera Control				Semantic and Distribution			Visual Quality			
	RotErr↓	TransErr↓	CamMC↓	ATE↓	CLIP↓	FID↓	FVD↓	PSNR↑	SSIM↑	Image quality↑	Aesthetic quality↑
ReCamMaster	0.112	0.711	0.739	0.837	**23.58**	85.19	550.65	13.19	0.449	59.35	54.63
TrajectoryCrafter	0.202	0.635	0.740	0.734	23.69	48.01	384.27	15.37	0.497	58.05	54.74
Vista4D	0.127	0.658	0.698	0.751	28.45	101.05	699.12	12.87	0.300	63.43	54.92
Ours	**0.089**	**0.469**	**0.505**	**0.532**	23.61	**23.29**	**164.03**	**18.99**	**0.534**	**64.26**	**55.01**

Table 4: Ablation study of the noise-specific expert (NS-Expert) and the unified shared action encoder (UniEncoder) on camera-controlled generation.

Model	Camera Control				Semantic and Distribution			Visual Quality			
	RotErr↓	TransErr↓	CamMC↓	ATE↓	CLIP↓	FID↓	FVD↓	PSNR↑	SSIM↑	Image quality↑	Aesthetic quality↑
Base	0.115	0.288	0.363	0.334	29.40	18.56	672.67	14.99	0.457	66.61	52.13
+ NS-Expert	0.057	0.209	0.231	0.229	**29.06**	16.13	**536.47**	15.63	0.479	**67.66**	52.31
+ NS-Expert & UniEncoder	**0.051**	**0.177**	**0.205**	**0.207**	29.15	**15.63**	551.28	**15.89**	**0.486**	67.45	**52.51**

the latter increasing from 40.98 to 60.36 in the human domain (Tab. 5). These results show better reconstruction and motion control, although some perceptual metrics remain unchanged or decline.

Effect of Shared Action Encoding. Adding UniEncoder on top of NS-Expert further improves trajectory accuracy, depth accuracy, instruction following, and semantic alignment in both robot and human domains (Tab. 5). These gains come with small decreases in subject consistency in both domains and interaction quality in the robot domain. For camera-conditioned generation, UniEncoder improves RotErr, TransErr, FID, and FVD, while slightly increasing CamMC and ATE (Tab. 4).

Table 5: Ablation of noise-specific experts (NS-Expert) and shared action encoding (UniEncoder) on robot and human action-conditioned generation.

Model	Visual Quality				Content Consistency		Physics Adherence		3D Accuracy	Controllability	
	PSNR↑	SSIM↑	Image quality↑	Aesthetic quality↑	Background consistency↑	Subject consistency↑	Interaction quality↑	Trajectory accuracy↑	Depth accuracy↑	Instruction following↑	Semantic alignment↑
Robotics											
Base	17.53	0.76	71.4	**47.77**	37.64	37.97	67.2	83.16	97.14	73.45	90.57
+ NS-Expert	21.19	0.81	73.37	47.25	38.18	**38.66**	**70.32**	87.4	98.3	73.44	92.15
+ NS-Expert & UniEncoder	**21.69**	**0.82**	**74.2**	47.7	**38.5**	38.4	69.5	**88.9**	**98.7**	**75.0**	**92.4**
Ego-centric											
Base	21.49	0.85	68.04	39.7	**68.9**	**70.48**	88.46	40.98	97.56	73.13	82.73
+ NS-Expert	23.66	0.88	68.04	39.86	68.7	70.13	92.22	60.36	97.69	74.36	84.04
+ NS-Expert & UniEncoder	**23.73**	**0.88**	**68.3**	**39.9**	68.7	70.1	**92.3**	**62.2**	**97.7**	**76.1**	**84.1**

Table 6: Ablation study of video self-supervision for camera-controlled video re-shooting.

Model	Camera Control				Semantic and Distribution			Visual Quality			
	RotErr↓	TransErr↓	CamMC↓	ATE↓	CLIP↓	FID↓	FVD↓	PSNR↑	SSIM↑	Image quality↑	Aesthetic quality↑
Base	0.095	0.470	**0.495**	0.536	**23.42**	24.45	174.51	18.68	0.523	**64.72**	54.65
+ Self-Supervision	**0.089**	**0.469**	0.505	**0.532**	23.61	**23.29**	**164.03**	**18.99**	**0.534**	64.26	**55.01**

Effect of Cross-Segment Self-Supervision. For video re-shooting, cross-segment self-supervision primarily improves key video-quality and camera control metrics, such as PSNR, TransErr and RotErr(Tab. 6). These consistent gains indicate that video self-supervision enables the model to learn accurate visual and spatial representations from single-view videos, even when paired multi-view data are scarce, thereby improving its video re-shooting capability.

4 Conclusion

We presented a unified video world model for camera-controlled generation and embodied action-conditioned prediction. Reference-camera alignment, shared action encoding, and visual embodiment conditioning enable joint training on heterogeneous video sources with partially available controls. Cross-segment pretraining and selective high-noise adaptation further support camera-controlled video re-shooting using limited paired observations. Experiments demonstrate strong visual fidelity and control accuracy across human and robot action-conditioned generation and camera-controlled I2V and V2V tasks. These results support control alignment as a useful approach to joint camera- and action-conditioned video modeling.

AI USE STATEMENT

In this work, we used generative AI tools for language editing and writing assistance. All AI-assisted content was reviewed and revised by the authors. The authors take full responsibility for the final content of the paper, including all text, claims, experiments, and conclusions produced with the aid of generative AI tools.

REPRODUCIBILITY STATEMENT

Model architecture, training settings, and evaluation protocols are described in the main paper. Dataset sources and preprocessing procedures are detailed in Appendix B.

REFERENCES

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.

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.

Build AI. Egocentric-100k, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22875–22889. IEEE, 2025a.

Sherwin Bahmani, Ivan Skorokhodov, Aliaksandr Siarohin, Willi Menapace, Guocheng Qian, Michael Vasilkovsky, Hsin-Ying Lee, Chaoyang Wang, Jiaxu Zou, Andrea Tagliasacchi, et al. Vd3d: Taming large video diffusion transformers for 3d camera control. In *International Conference on Learning Representations*, volume 2025, pp. 66712–66737, 2025b.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. eDiff-I: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.

Erik Bauer, Elvis Nava, and Robert K Katzschmann. Latent action diffusion for cross-embodiment manipulation. *arXiv preprint arXiv:2506.14608*, 2025.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Xindong He, Xu Huang, et al. Agibot world colosseo: A large-scale manipulation platform for scalable and intelligent embodied systems. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2025.

Chenjie Cao, Jingkai Zhou, Shikai Li, Jingyun Liang, Chaohui Yu, Fan Wang, Xiangyang Xue, and Yanwei Fu. Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

Boyu Chen, Yi Chen, Lu Qiu, Jerry Bai, Yuying Ge, and Yixiao Ge. Unit: Toward a unified physical language for human-to-humanoid policy learning and world modeling. *arXiv preprint arXiv:2604.19734*, 2026a.

Cheng Chen, Jerry Bai, Jiacheng Wei, Boyu Chen, Xiaoji Zheng, Fan Wu, Minghao Yang, Tianrun Chen, Ruibo Li, Xiaoyu Yue, et al. Anyworld: Factorized egocentric world models for cross-embodiment generalization. *arXiv preprint arXiv:2608.29242*, 2026b.

Lawrence Yunliang Chen, Chenfeng Xu, Karthik Dharmarajan, Muhammad Zubair Irshad, Richard Cheng, Kurt Keutzer, Masayoshi Tomizuka, Quan Vuong, and Ken Goldberg. Rovi-aug: Robot and viewpoint augmentation for cross-embodiment robot learning. *arXiv preprint arXiv:2409.03403*, 2024.

AgiBot World Colosseum contributors. Agibot world colosseum. GitHub repository, 2024.

Prithwish Dan, Kushal Kedia, Angela Chao, Edward Weiyi Duan, Maximus Adrian Pace, Wei-Chiu Ma, and Sanjiban Choudhury. X-sim: Cross-embodiment learning via real-to-sim-to-real. *arXiv preprint arXiv:2505.07096*, 2025.

Ria Doshi, Homer Walke, Oier Mees, Sudeep Dasari, and Sergey Levine. Scaling cross-embodied learning: One policy for manipulation, navigation, locomotion and aviation. *arXiv preprint arXiv:2408.11812*, 2024.

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

Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. *Advances in neural information processing systems*, 29, 2016.

Chongkai Gao, Haozhuo Zhang, Zhixuan Xu, Cai Zhehao, and Lin Shao. Flip: Flow-centric generative planning as general-purpose manipulation world model. In *International Conference on Learning Representations*, volume 2025, pp. 21927–21948, 2025.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19383–19400, 2024.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Yanjiang Guo, Lucy Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 6121–6138, 2026.

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 *International conference on machine learning*, pp. 2555–2565. PMLR, 2019b.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Hao He, Ceyuan Yang, Shanchuan Lin, Yinghao Xu, Meng Wei, Liangke Gui, Qi Zhao, Gordon Wetzstein, Lu Jiang, and Hongsheng Li. Cameractrl ii: Dynamic scene exploration via camera-controlled video diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426. IEEE, 2025.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

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

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, 2026.

Chen Hou and Zhibo Chen. Training-free camera control for video generation. *arXiv preprint arXiv:2406.10126*, 2024.

Chengkai Hou, Kun Wu, Jiaming Liu, Zhengping Che, Di Wu, Fei Liao, Guangrun Li, Jingyang He, Qiuxuan Feng, Zhao Jin, Chenyang Gu, Zhuoyang Liu, Nuowei Han, Xiangju Mi, Yaoxu Lv, Yankai Fu, Gaole Dai, Langzhe Gu, Tao Li, Yuheng Zhang, Yixue Zhang, Xinhua Wang, Shichao Fan, Meng Li, Zhen Zhao, Ning Liu, Zhiyuan Xu, Pei Ren, Junjie Ji, Haonan Liu, Kuan Cheng, Shanghang Zhang, and Jian Tang. Robomind 2.0: A multimodal, bimanual mobile manipulation dataset for generalizable embodied intelligence, 2025.

Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

Tao Hu, Haoyang Peng, Xiao Liu, and Yuewen Ma. Ex-4d: Extreme viewpoint 4d video synthesis via depth watertight mesh. *arXiv preprint arXiv:2506.05554*, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, Junfeng Ni, Hongyu Pan, Zhongxu Sun, Fei Yu, Zengye Ge, Mengmeng Du, Nianfei Fan, Mingchao Sun, Yu Liu, Yongchang, Yanqing Zhu, Jiahang Wang, Ning Ying, Yuze Xuan, Di Yang, Zhicheng Liu, Zhe Gao, Tingbing Xu, Jiacheng Sui, Wenjin Yang, Junnan Lai, Shufeng Liu, Yuan Liu, Zheng Zhou, Yingliang Peng, Dawei Cao, Kaifeng Sheng, Yuxiang Cai, Fei Lu, Mu Xu, and Ning Guo. ABot-World-0: Infinite interactive world rollout on a single desktop gpu, 2026.

Wonjoon Jin, Qi Dai, Chong Luo, Seung-Hwan Baek, and Sunghyun Cho. Flovd: Optical flow meets video diffusion model for enhanced camera-controlled video synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2040–2049. IEEE, 2025.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Simar Kareer, Karl Pertsch, James Darpinian, Judy Hoffman, Danfei Xu, Sergey Levine, Chelsea Finn, and Suraj Nair. Emergence of human to robot transfer in vision-language-action models. *arXiv preprint arXiv:2512.22414*, 2025.

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. In *Proceedings of Robotics: Science and Systems*, 2024.

Byungjun Kim, Taeksoo Kim, Hyunsoo Cha, and Hanbyul Joo. Robot-factored world models via robot rendering. *arXiv preprint arXiv:2607.22535*, 2026.

Hanjung Kim, Jaehyun Kang, Hyolim Kang, Meedeum Cho, Seon Joo Kim, and Youngwoon Lee. Uniskill: Imitating human videos via cross-embodiment skill representations. *arXiv preprint arXiv:2505.08787*, 2025.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

LejuRobotics. Let:full-size humanoid robot real-world dataset. AtomGit dataset, 2025.

Dayou Li, Lulin Liu, Bangya Liu, Shijie Zhou, Jiu Feng, Ziqi Lu, Minghui Zheng, Chenyu You, and Zhiwen Fan. Egocentric world model for photorealistic hand-object interaction synthesis. *arXiv preprint arXiv:2603.13615*, 2026.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025a.

Qixiu Li, Yu Deng, Yaobo Liang, Lin Luo, Lei Zhou, Chengtang Yao, Lingqi Zeng, Zhiyuan Feng, Huizhi Liang, Sicheng Xu, Yizhong Zhang, Xi Chen, Hao Chen, Lily Sun, Dong Chen, Jiaolong Yang, and Baining Guo. Scalable vision-language-action model pretraining for robotic manipulation with real-life human activity videos. *arXiv preprint arXiv:2510.21571*, 2025b.

Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, Chuanyun Deng, Yepan Xiong, Min Chen, et al. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 28785–28796. IEEE, 2025c.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. In *Advances in Neural Information Processing Systems*, 2025d.

Weixin Liang, Lili Yu, Liang Luo, Srini Iyer, Ning Dong, Chunting Zhou, Gargi Ghosh, Mike Lewis, Wen-tau Yih, Luke Zettlemoyer, and Xi Victoria Lin. Mixture-of-Transformers: A sparse and scalable architecture for multi-modal foundation models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856.

Kuan Heng Lin, Zhizheng Liu, Pablo Salamanca, Yash Kant, Ryan Burgert, Yuancheng Xu, Koichi Namekata, Yiwei Zhao, Bolei Zhou, Micah Goldblum, Paul Debevec, and Ning Yu. Vista4D: Video reshooting with 4d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32671–32682, June 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Jiwen Liu, Shujuan Li, Zhixue Fang, Xiaohan Li, Yan Zhou, Zijie Meng, Zhimin Zhang, Yawen Luo, Guoxin Zhang, Yu-Shen Liu, et al. Omnidirector: General multi-shot camera cloning without cross-paired data. *arXiv preprint arXiv:2606.13432*, 2026a.

Jiwen Liu, Shujuan Li, Xiaohan Li, Zijie Meng, Xinyue Liu, Yulong Xu, Yan Zhou, and Guoxin Zhang. Tars: Timestep-aware data scaling for 3d-free video re-shooting. *arXiv preprint arXiv:2607.28261*, 2026b.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024.

Avinash Paliwal, Adithya Iyer, Shivin Yadav, Muhammad Afridi, and Midhun Harikumar. Reshoot-anything: A self-supervised model for in-the-wild video reshooting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11596–11606, 2026.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, et al. Genie 2: A large-scale foundation world model. Google DeepMind technical blog, 2024.

Ryan Punamiya, Simar Kareer, Zeyi Liu, Josh Citron, Ri-Zhao Qiu, Xiongyi Cai, Alexey Gavryushin, Jiaqi Chen, Davide Liconti, Lawrence Y. Zhu, Patcharapong Aphiwetsa, Baoyu Li, Aniketh Cheluva, Pranav Kuppili, Yangcen Liu, Dhruv Patel, Aidan Gao, Hye-Young Chung, Ryan Co, Renee Zbizika, Jeff Liu, Xiaomeng Xu, Haoyu Xiong, Geng Chen, Sebastiano Oliani, Chenyu Yang, Xi Wang, James Fort, Richard A. Newcombe, Josh Gao, Jason Chong, Garrett Matsuda, Aseem Doriwala, Marc Pollefeys, Robert K. Katzschmann, Xiaolong Wang, Shuran Song, Judy Hoffman, and Danfei Xu. Egoverse: An egocentric human dataset for robot learning from around the world. In *Robotics: Science and Systems*, 2026.

Boxiang Qiu, Liliang Chen, Yue Liao, Nan Wang, Lintao Wang, Jiayi Luo, Wenzhi Zhao, Shengcong Chen, Di Chen, Ye Li, Chen Gao, Shuicheng Yan, Si Liu, Maoqing Yao, and Guanghui Ren. Ge-sim 2.0: A roadmap towards comprehensive closed-loop video world simulators for robotic manipulation. *arXiv preprint arXiv:2605.27491*, 2026.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *arXiv preprint arXiv:2410.12822*, 2024.

Ropedia. Xperience-10m: A large-scale egocentric multimodal dataset with structured 3d/4d annotations, 2026. Dataset.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *arXiv preprint arXiv:2602.08971*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world model. *arXiv preprint*, 2025.

AgiBot World Team. Agibot world 2026. Hugging Face dataset, 2026.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026a.

Galaxea Team. Galaxea g0: Open-world dataset and dual-system vla model. *arXiv preprint arXiv:2509.00576v1*, 2025.

GigaWorld Team, Angyuan Ma, Boyuan Wang, Bohan Li, Chaojun Ni, Guo Li, Guan Huang, Guosheng Zhao, Hao Li, Hengtao Li, et al. Gigaworld-1: A roadmap to build world models for robot policy evaluation. *arXiv preprint arXiv:2607.02642*, 2026b.

Octo Model Team, Dibya Ghosh, Homer Walke, Karl Pertsch, Kevin Black, Oier Mees, Sudeep Dasari, Joey Hejna, Tobias Kreiman, Charles Xu, et al. Octo: An open-source generalist robot policy. *arXiv preprint arXiv:2405.12213*, 2024.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 976–985, 2026.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

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.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025.

Lirui Wang, Xinlei Chen, Jialiang Zhao, and Kaiming He. Scaling proprioceptive-visual learning with heterogeneous pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124420–124450, 2024a.

Tianyu Wang, Dwait Bhatt, Xiaolong Wang, and Nikolay Atanasov. Cross-embodiment robot manipulation skill transfer using latent space alignment. *arXiv preprint arXiv:2406.01968*, 2024b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024c.

Dejia Xu, Weili Nie, Chao Liu, Sifei Liu, Jan Kautz, Zhangyang Wang, and Arash Vahdat. Camco: Camera-controllable 3d-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

Sherry Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Minghao Yin, Jiahao Lu, Wenbo Hu, Wang Zhao, Ying Shan, and Kai Han. Scope: Sightline-coordinate positional encoding for video diffusion transformers. *arXiv preprint arXiv:2606.27345*, 2026.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactorly: Creating new games with generative interactive videos. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599. IEEE, 2025a.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025b.

Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

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

Haoqi Yuan, Zhixuan Liang, Anzhe Chen, Ye Wang, Haoyang Li, Pei Lin, Yiyang Huang, Zixing Lei, Tong Zhang, Jiazhao Zhang, et al. Qwen-robotmanip technical report: Alignment unlocks scale for robotic manipulation foundation models. *arXiv preprint arXiv:2606.17846*, 2026.

Gu Zhang, Qicheng Xu, Haozhe Zhang, Jianhan Ma, Long He, Yiming Bao, Zeyu Ping, Zhecheng Yuan, Chenhao Lu, Chengbo Yuan, et al. Unidex: A robot foundation suite for universal dexterous hand control from egocentric human videos. *arXiv preprint arXiv:2603.22264*, 2026a.

Jie Zhang, Xiaoyue Chen, Anzhe Chen, Dayiheng Liu, Deqing Li, Gengze Zhou, Hale Yin, Haoqi Yuan, Haoyang Li, Jiahao Li, et al. Qwen-robotworld technical report: Unifying embodied world modeling through language-conditioned video generation. *arXiv preprint arXiv:2606.17030*, 2026b.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

Haoyu Zhao, Xingyue Zhao, Hangyu Li, Biao Gong, Kehan Li, Siteng Huang, Xin Li, Deli Zhao, and Zhongyu Li. Rynnworld-teleop: An action-conditioned world model for digital teleoperation. *arXiv preprint arXiv:2607.06558*, 2026.

Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *arXiv preprint arXiv:2410.15957*, 2024.

Guangcong Zheng, Teng Li, Xianpan Zhou, and Xi Li. Realcam-vid: High-resolution video dataset with dynamic scenes and metric-scale camera movements. *arXiv preprint arXiv:2504.08212*, 2025a.

Jinliang Zheng, Jianxiong Li, Dongxiu Liu, Yinan Zheng, Zhihao Wang, Zhonghong Ou, Yu Liu, Jingjing Liu, Ya-Qin Zhang, and Xianyuan Zhan. Universal actions for enhanced embodied foundation models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22508–22519. IEEE, 2025b.

Jensen Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12405–12414. IEEE, 2025a.

Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination. *arXiv preprint arXiv:2404.12377*, 2024.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *ACM Transactions on Graphics*, 37(4), 2018.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025b.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. In *International Conference on Learning Representations*, 2026.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

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

APPENDIX

A Training and Implementation Details **17**
A.1 Model Initialization and Training 17
A.2 Temporal and Control Conditioning 17
A.3 Inference and Evaluation Settings 18

B Data Preparation **18**
B.1 Robot Videos 18
B.2 Human Egocentric Videos 18
B.3 Internet Videos 19

C Related Work **19**
C.1 Action-Conditioned World Models 19
C.2 Camera Control Video Generation 19
C.3 Embodied Foundation Models and Cross-Embodiment Learning 20

D More Results **20**

A TRAINING AND IMPLEMENTATION DETAILS

A.1 MODEL INITIALIZATION AND TRAINING

We initialize the video backbone from Cosmos3-Nano 16B and use Qwen3-VL-8B-Instruct for pre-trained multimodal conditioning. The Wan2.2 VAE encodes and decodes video latents. The VLM and VAE remain frozen throughout training, while separate generation parameter banks handle high- and low-noise denoising. During Stage I, both experts and the shared action encoder learn from the heterogeneous video mixture. Each expert receives the same task and source mixture within its assigned noise range, allowing all sources to contribute to both denoising regimes.

Stage II uses synchronized cross-view pairs from MultiCamVideo and Ego-Exo4D. We update the high-noise expert and camera-conditioning interface while retaining the Stage I low-noise expert. The source video remains clean context, and the entire target video is predicted under the target controls. No target-view image is supplied in this re-shooting setting. Both stages use the masked flow-matching objective in Eq. 7, with supervision restricted to generated latent elements.

A.2 TEMPORAL AND CONTROL CONDITIONING

I2V examples use an initial image as visual context. For cross-segment V2V pretraining, context and target are sampled in temporal order from the same recording, allowing a gap between them. Available controls are aligned with the target segment. The self-supervision comes from this within-video context–target relation, while available text and geometric annotations provide additional conditioning. At paired adaptation, synchronized source–target observations replace the temporal pairs without changing the control interface.

The initial state S_0 relates the target's initial rigid poses to the source first-frame camera, while subsequent controls describe motion relative to the target's initial state. Camera and effector poses follow the reference-camera conventions in Sec. 2.2. Fingertip and gripper contact positions instead use local wrist/end-effector frames, and gripper opening remains an absolute distance. Missing fields retain learned null embeddings, distinguishing unavailable annotations from observed stationary motion.

Embodiment conditioning is separate from control availability. During training, the learned embedding for a training embodiment is randomly replaced by the shared unknown embedding, encouraging prediction from visual context and aligned controls. At inference, all examples use $\mathbf{e}_{\mathrm{unk}}$. This removes the need to select or introduce an embodiment identifier at test time. Nonvisual conditioning is also randomly dropped during training to support classifier-free guidance.

A.3 Inference and Evaluation Settings

Generation starts with the high-noise expert and switches to the low-noise expert below the routing threshold τ. For video re-shooting, the adapted high-noise expert is followed by the pretrained low-noise expert. Visual context remains clean throughout denoising. The qualitative human-to-robot examples in Fig. 8 use a human demonstration video together with a rendered initial robot image, which supplies target appearance and embodiment cues. This additional image condition is specific to those examples.

All evaluation episodes are excluded from both training stages. Baselines receive the conditioning modalities supported by their released interfaces, including action, skeleton, or camera controls where applicable. In the encoder ablation, preprocessing and reference-camera alignment are held fixed, isolating the effect of cross-embodiment parameter sharing. The compositional-control examples hold actions fixed when varying camera trajectories and hold camera trajectories fixed when varying actions.

B Data Preparation

Our training corpus comprises three categories: robot manipulation videos, human egocentric interaction videos, and internet videos. We combine publicly available datasets with internal datasets, selecting subsets through quality filtering, post-processing, and sampling to obtain the desired training mixture. These sources differ in their embodiments, geometric conventions, and availability of control annotations. We therefore apply category-specific processing to align the available action and camera signals within a shared 3D coordinate system anchored to a selected reference camera frame.

B.1 Robot Videos

We use AgiBotWorld2026 (Team, 2026), AgiBotWorldBeta (contributors, 2024), the Galaxea Open-World Dataset (Team, 2025), InternData-A1 (Tian et al., 2026), the LET Dataset (LejuRobotics, 2025), and RoboMIND2.0 (Hou et al., 2025). These datasets cover real-world and simulated manipulation across diverse robot platforms. They provide explicit physical signals, such as end-effector poses, joint states, and gripper commands, but differ in embodiment morphology, action parameterization, and coordinate conventions.

We convert the available robot action signals into a unified end-effector representation. Using camera extrinsics, we transform poses expressed in robot-base, torso, or world coordinates into the shared reference-camera coordinate system, giving corresponding spatial motions a consistent geometric interpretation across platforms. For recordings with substantial camera motion, such as those collected by mobile or humanoid robots, we also extract and retain camera trajectories as explicit viewpoint controls.

B.2 Human Egocentric Videos

We use EgoDex (Hoque et al., 2025), EgoVerse (Punamiya et al., 2026), VITRA-1M (Li et al., 2025b), Xperience-10M (Ropedia, 2026), and Egocentric-100K (AI, 2025). These datasets capture diverse manipulation behaviors and rich hand–object interactions from first-person viewpoints. Their annotations vary in availability and detail, while differences in human morphology, capture devices, and recording conditions introduce additional variation in viewpoints and motion patterns.

We retain available camera and hand-motion annotations and recover missing signals from visual observations using SLAM-based reconstruction and hand-tracking models. We represent human actions using a compact physical specification comprising wrist motion and five fingertip trajectories for each hand, rather than full-body representations or dense hand meshes. We express these signals and the camera trajectories in the shared reference-camera coordinate system, aligning their spatial semantics with robot actions while retaining fine-grained information about hand motion.

B.3 INTERNET VIDEOS

We use RealEstate10K (Zhou et al., 2018), DL3DV-10K (Ling et al., 2024), RealCam-Vid (Zheng et al., 2025a), Sekai-Game (Li et al., 2025d), SpatialVID (Wang et al., 2025), ABot-World (Jiang et al., 2026), and OmniWorld-Game (Zhou et al., 2025b). This category includes real-world footage and synthetic or game-engine videos, providing diverse scenes, appearances, and viewpoint variations. Unlike robot demonstrations, these sources do not consistently provide embodied action signals, although some include camera poses and geometric annotations. We additionally incorporate unlabeled internet videos.

Processing for this category focuses on obtaining reliable camera controls. We retain available camera annotations and otherwise estimate scene geometry and camera trajectories from RGB sequences using monocular depth estimation and camera reconstruction methods. Reliable trajectories are transformed into the shared reference-camera coordinate system. Videos without reliable camera poses are retained as unlabeled visual data for learning appearance and motion distributions, without imposing camera-control supervision. This allows us to exploit diverse video sources while avoiding unreliable geometric supervision.

C RELATED WORK

C.1 ACTION-CONDITIONED WORLD MODELS

Action-conditioned world models predict how the environment evolves in response to embodied actions, with early work ranging from visual foresight for planning (Finn et al., 2016; Ebert et al., 2018) to compact latent dynamics (Hafner et al., 2019b) for imagination-based control (Hafner et al., 2019a; 2023). With the rise of large-scale generative priors, recent methods have shifted toward visually grounded interactive simulators and video world models that generate future observations from explicit or latent controls (Hu et al., 2023; Yang et al., 2023; Bruce et al., 2024; Rigter et al., 2024; Huang et al., 2025; Parker-Holder et al., 2024; Agarwal et al., 2025). In robotics, recent world models further adapt generative priors to manipulation, supporting compositional prediction, fine-grained action-conditioned rollout, synthetic experience generation, and policy evaluation (Zhou et al., 2024; Gao et al., 2025; Zhu et al., 2025; Jang et al., 2025; Guo et al., 2026). Moving beyond action-conditioned formulations, recent works such as Uni3C (Cao et al., 2025), EgoHOI (Li et al., 2026), and RynnWorld-Teleop (Zhao et al., 2026) jointly model embodied motion and viewpoint variation, making them more closely related to our setting. Although these methods support both action and camera control, their signals are often encoded as view-dependent visual representations, such as 2D skeletons or rendered conditions. These representations implicitly entangle action with camera viewpoint. We instead represent action and camera trajectories in a shared 3D camera-frame coordinate system relative to a reference frame, enabling independent and compositional control.

C.2 CAMERA CONTROL VIDEO GENERATION

Camera control video generation exposes viewpoint evolution as an explicit control signal, with early methods such as MotionCtrl (Wang et al., 2024c) and CameraCtrl (He et al., 2024) introducing controllable camera motion into pretrained video diffusion models. Subsequent approaches improve geometric controllability through camera poses, 3D rays, epipolar constraints, and camera injection mechanisms (Xu et al., 2024; Bahmani et al., 2025b; Zheng et al., 2024; Bahmani et al., 2025a). Other methods reduce camera annotation requirements through 3D reconstruction, optical flow, or rendered context (Hou & Chen, 2024; Jin et al., 2025; Ren et al., 2025; Li et al., 2025c). Recent systems further support large-range scene exploration and re-rendering of dynamic videos under novel camera trajectories (He et al., 2025; Yu et al., 2025b; Bai et al., 2025; Zhou et al., 2025a; Yu et al., 2024). A parallel line of work reduces the dependence on paired multi-view supervision by exploiting monocular videos, pseudo-pair construction, or self-supervised learning (Hu et al., 2025; Paliwal et al., 2026; Liu et al., 2026a;b). Unlike these camera-centric data-scaling approaches, our cross-segment self-supervision enables joint learning from action-controlled, camera-controlled, and unannotated videos within a unified heterogeneous world model. Beyond viewpoint control, interactive game world models jointly condition generation on camera motion and player actions, including navigation, jumping, attacking, and keyboard–mouse commands (Guo et al., 2025; Yu et al., 2025a; Zhang et al., 2025; Li et al., 2025a). These works share our motivation of modeling

multiple control signals, but their action interfaces remain largely game-oriented and are typically expressed as discrete or low-dimensional commands. In contrast, we jointly represent continuous 3D action trajectories and camera motion within a shared physical representation, making the resulting controls directly applicable to embodied evaluation and closed-loop learning.

C.3 Embodied Foundation Models and Cross-Embodiment Learning

A growing number of embodied foundation models use shared architectures and large-scale co-training to transfer across tasks, sensors, action spaces, and robot platforms (O'Neill et al., 2024; Team et al., 2024; Kim et al., 2024; Doshi et al., 2024). Recent approaches further explore embodiment-invariant interfaces through universal action representations, latent spaces, geometric motion, and human–robot co-training (Zheng et al., 2025b; Bauer et al., 2025; Yuan et al., 2024; Chen et al., 2024; Kim et al., 2025; Dan et al., 2025; Zhang et al., 2026a; Kareer et al., 2025). Recent cross-embodiment research increasingly focuses on aligning action representations across heterogeneous human and robot data, allowing them to be jointly trained within a unified model. Cosmos 3 (Agarwal et al., 2026) decomposes actions into shared geometric components but retains domain-specific projections. Qwen-RobotManip (Yuan et al., 2026) takes a closely related direction by representing heterogeneous actions as end-effector deltas in the camera coordinate frame. However, it relies on human-to-robot retargeting for egocentric demonstrations.

In contrast, we directly learn from raw heterogeneous videos without explicit conversion, using a shared physical representation that aligns embodied actions and camera trajectories in a canonical 3D camera-frame coordinate system. Building on this alignment, we extend cross-embodiment unification from policy learning to world modeling (Chen et al., 2026a; Zhang et al., 2026b; Kim et al., 2026; Chen et al., 2026b). Unlike prior formulations based on latent tokens, language interfaces, rendered geometry, or factorized controls, our representation uses fine-grained numerical trajectories directly corresponding to embodied control signals.

D More Results

We provide additional qualitative results for compositional control, camera-controlled generation, and human-to-robot transformation, followed by an expanded robot-action comparison and examples across diverse visual domains. These results complement the main-paper evaluations of control accuracy and visual consistency.

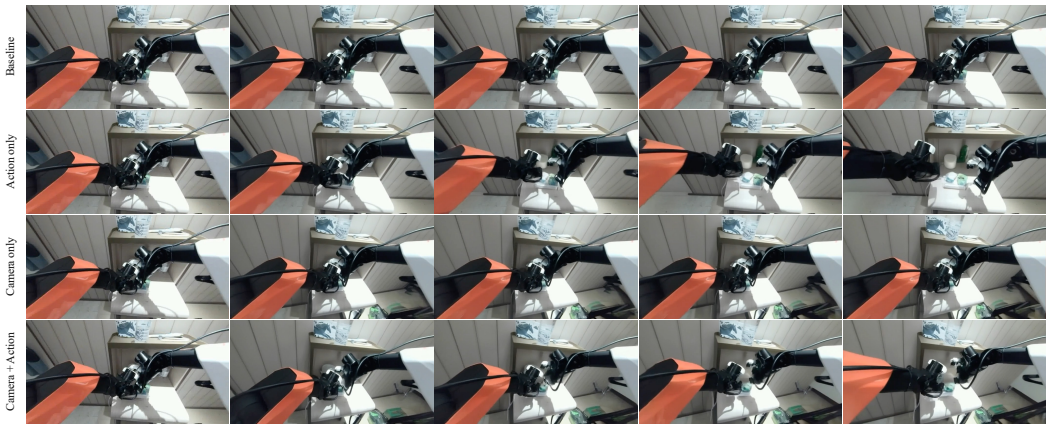

Figure 5: Qualitative examples of compositional action–camera control. Rows correspond to action trajectories and columns to camera trajectories.

Robot motion and scene consistency. Fig. 10 expands the AgiBot example in Fig. 4, showing more intermediate frames of the refrigerator-opening interaction. As the gripper moves and the door opens, our prediction reveals a populated interior with a more coherent arrangement of shelves and



Figure 6: Qualitative comparison of camera-controlled image-to-video generation. The input image and camera trajectory are shown on the left, and generated frames from different methods are shown on the right.

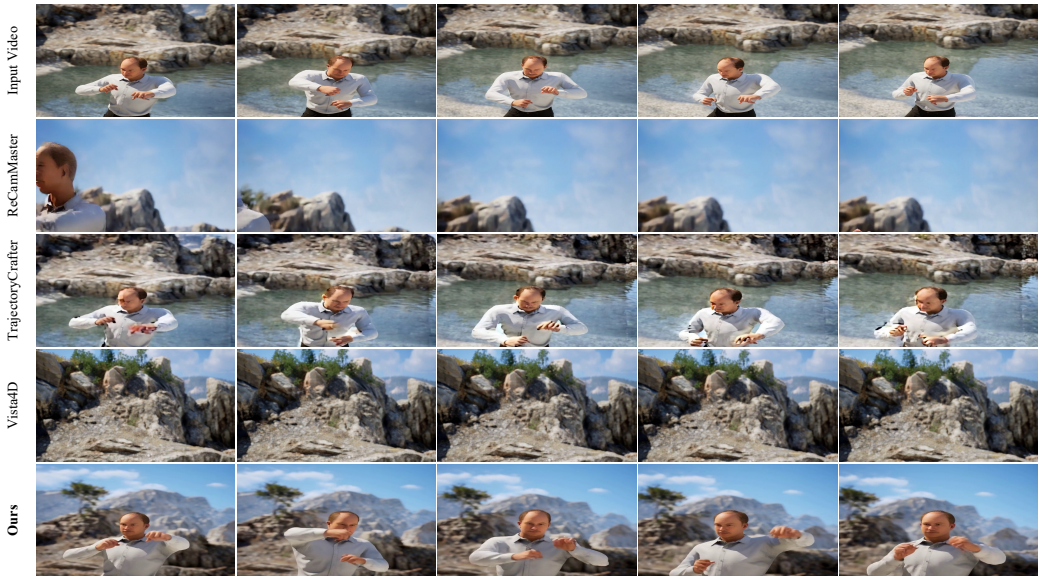

Figure 7: Qualitative comparison of camera-controlled video-to-video generation. Our model generates a temporally coherent sequence from a novel viewpoint while preserving the subject's appearance and action dynamics from the reference video.

objects. The highlighted baseline regions exhibit inconsistencies in gripper geometry, door appearance, or interior layout. This example complements the robot-action results in Tab. 1: predicting the visual consequences of actions requires consistency between the moving effector and the surrounding scene, beyond plausible motion alone.

Generation across visual domains. Fig. 11 presents image-conditioned examples spanning landscapes, architectural interiors, first-person riding, and robot manipulation. Across the displayed frames, recognizable scene appearance is maintained as viewpoints and scene elements evolve. These examples illustrate the visual coverage motivated by heterogeneous pretraining in Sec. 2.3, while the shared conditioning interface in Sec. 2.2 accommodates the controls available in each domain. Together with the compositional examples in Fig. 5, they illustrate the model's intended scope across visual environments and control settings.

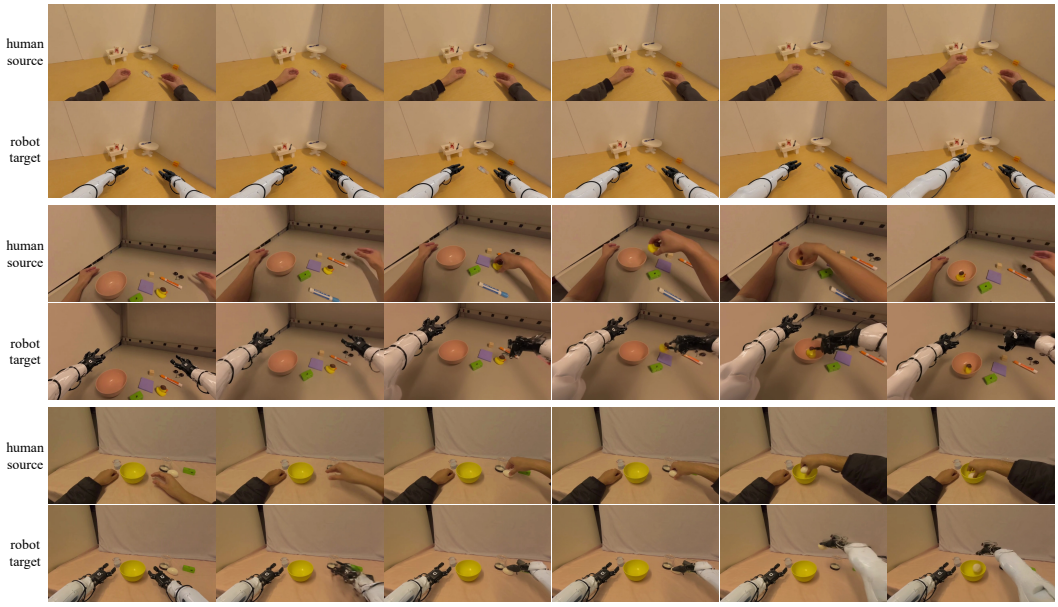

Figure 8: Given a human demonstration video and a rendered initial frame of the robot, our model predicts the robot's subsequent execution video. The predicted sequences exhibit plausible robot motions that broadly follow the demonstrated actions.



Figure 9: Our model generates videos conditioned on varying camera poses, robot actions, and hand poses, producing corresponding variations in viewpoint and motion.

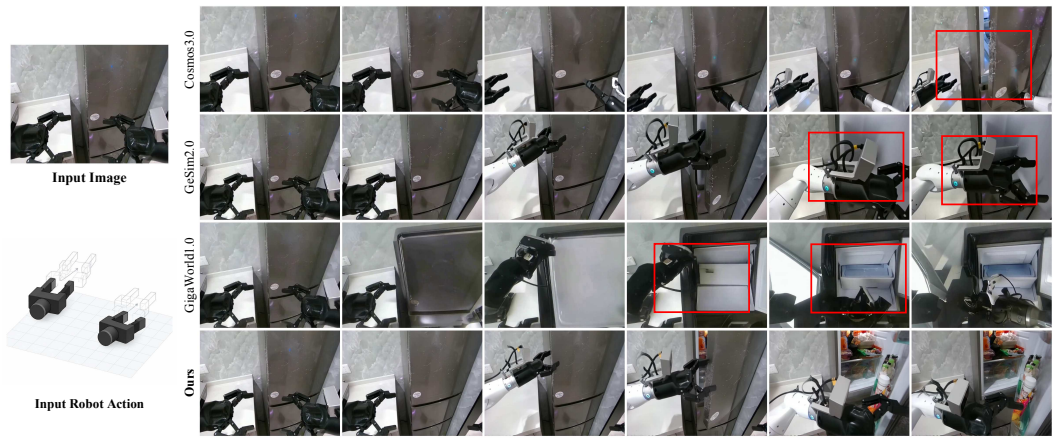

Figure 10: Expanded action-conditioned comparison on AgiBot. The input image and robot action are shown on the left, followed by generated frames. Red boxes highlight geometric and appearance inconsistencies in baseline predictions.

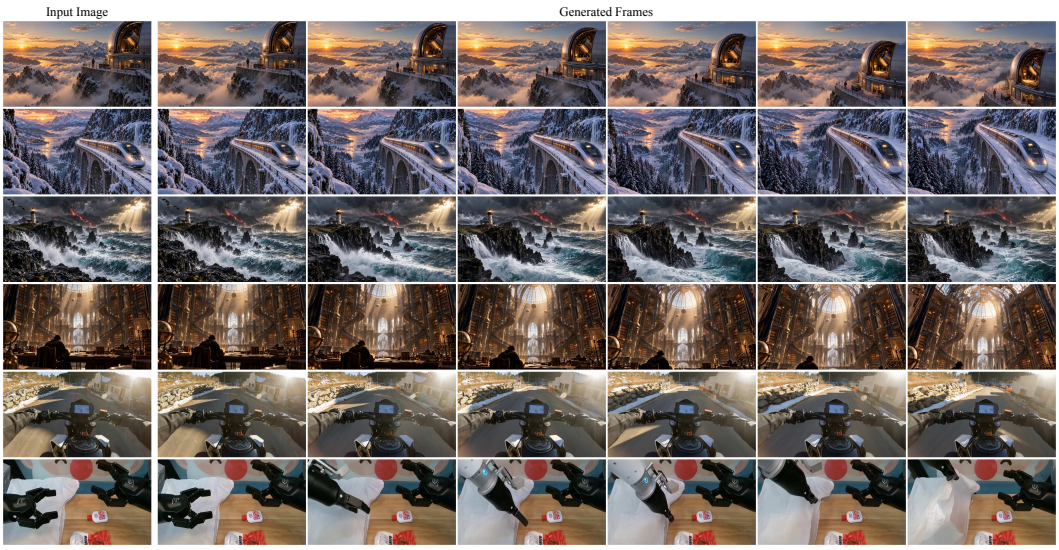

Figure 11: Image-conditioned generation across diverse visual domains. Each row shows an input image on the left and generated frames on the right, spanning landscapes, architectural interiors, first-person riding, and robot manipulation.