MoFa-World: Motion-Factorized Latent Actions for Interactive World Models

Anonymous authors
Paper under double-blind review



Figure 1: Our motion-factorized latent actions for interactive world models. MoFa-LAM learns separate, complementary latent actions for egomotion and object interaction. Unlike prior latent action models that entangle these behaviors within a single latent action, our representations support (a) separate egomotion and interaction control after dataset-specific adaptation, (b) cross-scene egomotion transfer, and (c) cross-scene interaction transfer.

Abstract

Egocentric world models provide a promising foundation for interactive environment simulation, yet existing approaches tend to focus on egomotion control, with object-interaction control constrained by the scarcity of densely annotated interaction videos. Although latent action models (LAMs) learn action representations from videos without action labels, they often entangle object interaction and egomotion in a single representation. We believe that pixel-space motion patterns can reveal distinct action semantics, providing a basis for learning decoupled representations. Building on this insight, we introduce MoFa-LAM, which pretrains region-level latent actions through pixel-space motion factorization and aggregates them into interaction and egomotion representations through region-specific reconstruction with limited mask supervision. Probing experiments demonstrate the functional specialization of the two representations and the benefits of motion-factorized pretraining over vanilla LAM pretraining. We further introduce MoFa-World, an egocentric world model conditioned on these representations through distinct pathways. It supports cross-scene action transfer and adaptation to dataset-specific controls, including joint use of interaction and egomotion controls learned from different datasets. Together, these results support pixel-space motion factorization as a way to learn decoupled control representations for egocentric world models.

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

Picking up a cup while turning to look elsewhere combines two behaviors: interacting with an object and changing viewpoint. An interactive egocentric world model should allow users to specify and combine these behaviors through separate controls. Such a capability would broaden first-person simulation beyond navigation to environments in which agents can both explore and manipulate their surroundings. Yet existing egocentric world models largely emphasize locomotion and viewpoint control (Bar et al., 2025; Wu et al., 2026), while object-interaction control remains constrained by scarce annotations and diverse interaction behaviors (Xiong et al., 2026).

Learning controls directly from unlabeled videos offers a viable alternative. Latent action models (LAMs) infer compact representations of visual transitions and use them to condition future-frame generation (Bruce et al., 2024; Gao et al., 2025). However, existing methods commonly encode all visual changes within a transition into a single latent action, as illustrated in Figure 2(a). An egocentric transition can contain multiple sources of visual change, including object interaction, egomotion, and independent background dynamics (Wang et al., 2026b; Nam et al., 2026), which are not explicitly separated in this representation. This lack of explicit separation makes it difficult to vary one behavior while preserving another, motivating decoupled action representations for separate control of object interaction and egomotion.

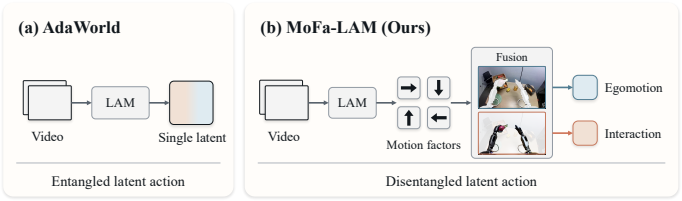

Figure 2: (a) AdaWorld encodes visual changes in a single latent action. (b) MOFA-LAM learns decoupled egomotion and interaction representations by separating their motion effects in pixel space through motion factorization and mask-guided fusion.

Our key insight is that distinct action semantics can be reflected in different motion patterns in pixel space. Learning separate latent actions to capture these motion patterns encourages each representation to encode a distinct component of the visual transition, providing a basis for semantic decoupling. Building on this insight, we introduce MOFA-LAM, a motion-factorized latent action model that first pretrains region-level latent actions through pixel-space motion factorization, as illustrated in Figure 2(b). To provide more explicit guidance for decoupling in pixel space, we use limited mask supervision to aggregate these regional factors into interaction and egomotion latent actions, with reconstruction of object-interaction regions and the surrounding scene guiding their respective functional specialization. Leveraging these two latent actions extracted by MOFA-LAM, we introduce MOFA-WORLD, an interactive world model for egocentric environments that supports both object interaction and egomotion. To account for their distinct effects on visual dynamics, we incorporate the two latent actions through separate conditioning mechanisms.

To validate latent-action decoupling, we conduct probing experiments on action classification, robot action regression, and relative camera pose estimation. The interaction and egomotion latents exhibit complementary task specialization, with each outperforming the other on its respective tasks. Under the same pixel-space supervision for decoupling, motion-factorized pretraining further improves performance on all three tasks over vanilla LAM pretraining. Beyond these probing tasks, MOFA-WORLD supports cross-scene transfer of egomotion and interaction and adaptation to dataset-specific controls, as illustrated in Figure 1. Interaction and egomotion controls learned from different datasets can also be combined, demonstrating the composability of these representations.

Our main contributions are summarized as follows:

- We reveal that pixel-space motion patterns provide a basis for latent-action decoupling, and introduce MOFA-LAM to realize this insight through motion-factorized pretraining and mask-guided fusion.
- We develop MOFA-WORLD, an egocentric world model that incorporates interaction and egomotion latent actions through distinct conditioning pathways, enabling separate and joint control of both behaviors.
- Experiments demonstrate complementary functional specialization and improved probing performance over vanilla LAM pretraining, along with cross-scene action transfer, dataset-specific control adaptation, and cross-dataset control composition.

2 Related Work

Latent Action Learning from Videos. Latent action learning derives action representations from videos without action labels. Early methods supported imitation from observation (Edwards et al., 2019) and planning (Rybkin et al., 2019). LAPO (Schmidt & Jiang, 2024) and Genie (Bruce et al., 2024) infer latent actions with an inverse model and reconstruct future observations with a forward model. These models use discrete codebooks, whereas AdaWorld (Gao et al., 2025), Olaf-World (Jiang et al., 2026), and CLAW (Ayalew et al., 2026) learn continuous latent spaces. Latent actions have since supported policy pretraining (Ye et al., 2025), action transfer and adaptation, and action-conditioned world models.

Subsequent work improves robustness to action-irrelevant variation and separates distinct transition effects. Methods suppress action-irrelevant changes through limited action supervision in LAOM (Nikulin et al., 2025) or optical-flow constraints in LAOF (Bu et al., 2026), while Garrido et al. (2026) and Motus (Bi et al., 2026) scale latent action learning with constrained continuous latents and optical-flow-derived motion signals, respectively. These methods, however, still encode each transition with a single latent action, entangling simultaneous visual changes. Recent methods further factorize transition effects to separate simultaneous sources of visual change, using object-centric representations in PlaySlot (Villar-Corrales & Behnke, 2025), Object-Centric LAM (Klepach et al., 2026), and LPWM (Daniel et al., 2026), scene factors in FLAM (Wang et al., 2026b), local transition effects in OTF-LAM (Nam et al., 2026), or a separation of motion-relevant structure from visual content in DiLA (Zhang et al., 2026). However, these methods do not separate egomotion and object interaction into independently controllable latent representations. In contrast, MoFa-LAM combines pixel-space motion factorization with mask-guided reconstruction to learn functionally specialized representations for egomotion and object interaction.

Action-Controllable World Models Action-controllable world models (Ha & Schmidhuber, 2018; Yang et al., 2024; Alonso et al., 2024; Huang et al., 2025b) predict future visual observations from context and control signals, supporting planning and interactive generation. Existing video world models remain largely navigation-centric (Bar et al., 2025), with progress in open-world control (Che et al., 2024), real-time generation (Valevski et al., 2025; Li et al., 2025; He et al., 2025), and long-horizon consistency (Sun et al., 2026a; Hong et al., 2025; Wu et al., 2026). Beyond navigation, control signals include egocentric kinematics and text (Wang et al., 2024), robot trajectories (Zhu et al., 2025), hand actions (Sun et al., 2026b), and dense action descriptions for mid-rollout interaction (Xiong et al., 2026). Such explicit controls offer clear semantics but require environment- or embodiment-specific supervision, limiting transfer across control spaces.

Latent-action world models instead learn controls from action-free videos (Bruce et al., 2024), enabling transfer and data-efficient adaptation to new action spaces (Gao et al., 2025; Jiang et al., 2026). DreamDojo (Gao et al., 2026) transfers knowledge from egocentric human videos to contact-rich robotic manipulation, while CoLA-World (Wang et al., 2026a) and CLAW (Ayalew et al., 2026) jointly learn latent actions and world models. Building on this line of work, MoFa-World conditions generation on separate interaction and egomotion latent actions through distinct pathways, supporting cross-scene transfer, data-efficient adaptation to dataset-specific controls, and joint control of both behaviors.

3 Method

3.1 Problem Formulation and Overview

Given an egocentric RGB video $\mathbf{x}_{0:T} = \{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$ with $\mathbf{x}_t \in \mathbb{R}^{H\times W\times 3}$, a standard latent action model encodes the transition from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$ as a latent action $\mathbf{z}_t$. In an inverse–forward architecture (Schmidt & Jiang, 2024; Gao et al., 2025), an inverse-dynamics encoder infers $\mathbf{z}_t$ from observed frames, and a forward decoder reconstructs $\mathbf{x}_{t+1}$ from $\mathbf{x}_t$ and $\mathbf{z}_t$.

However, an egocentric transition can contain object interaction, camera motion induced by egomotion, and background dynamics, which a single latent action does not explicitly distinguish. We aim to learn distinct action representations for separate control of object interaction and egomotion. Our key insight is that motion patterns in pixel space provide cues for learning semantically distinct latent actions. We therefore pretrain region-level latent actions through pixel-space motion

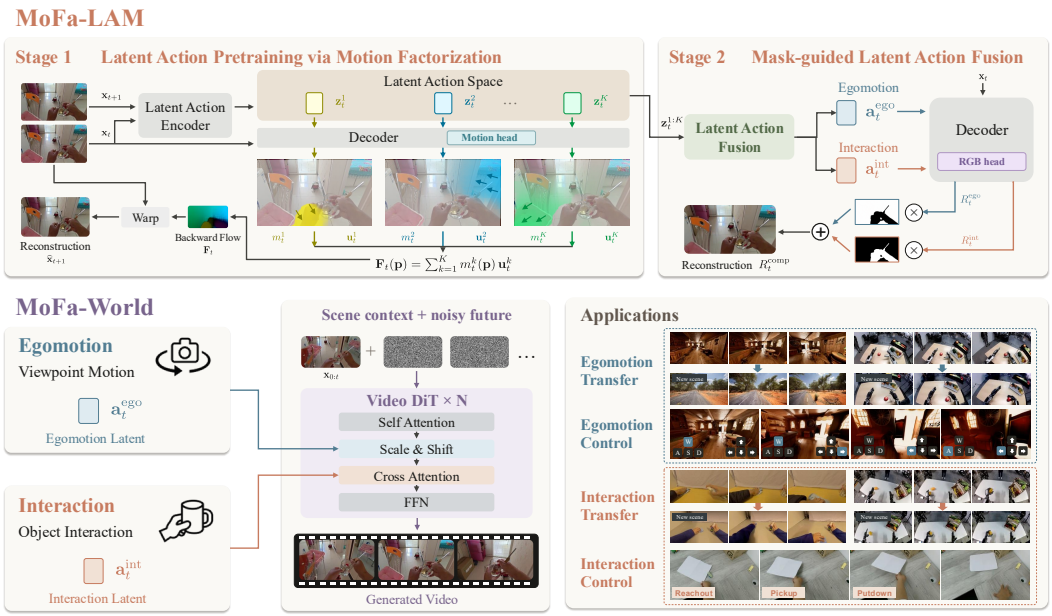

Figure 3: Overview of our framework. MoFa-LAM pretrains region-level latent actions through pixel-space motion factorization. With limited mask supervision, mask-guided fusion aggregates these motion factors into interaction and egomotion latent actions through region-specific reconstruction. MoFa-World incorporates the two latent actions through distinct conditioning pathways, supporting cross-scene transfer and explicit control of egomotion and object interaction.

factorization, encouraging them to capture distinct motion components that jointly reconstruct the next frame. Guided by the same insight, we use limited mask supervision to aggregate these factors into interaction and egomotion latent actions through reconstruction of object-interaction regions and the surrounding scene, respectively. Object-interaction regions emphasize changes associated with manipulation, while the surrounding scene provides cues about egomotion. These different reconstruction targets encourage the two latent actions to specialize in their intended roles.

Figure 3 summarizes the three-stage framework. First, MoFa-LAM pretrains K region-level latent actions $\mathbf{z}_t^{1:K}$ on a video collection $\mathcal{D}_u$ through pixel-space motion factorization. Second, mask-guided fusion aggregates these factors into interaction and egomotion latent actions, $\mathbf{a}_t^{\text{int}}$ and $\mathbf{a}_t^{\text{ego}}$, using a limited subset $\mathcal{D}_m \subset \mathcal{D}_u$ with object-interaction masks. Finally, MoFa-World incorporates the two latent-action sequences through distinct conditioning pathways to model $p_\theta(\mathbf{x}_{t+1:T} \mid \mathbf{x}_{0:t}, \mathbf{a}^{\text{int}}_{t:T-1}, \mathbf{a}^{\text{ego}}_{t:T-1})$. Here, $\mathbf{x}_{0:t}$ provides the observed context, and the future latent actions are supplied as controls rather than predicted by the world model.

3.2 Latent Action Pretraining via Motion Factorization

We seek to represent regions with different motion patterns using separate latent actions. To learn this decomposition, the model predicts how these regions move between frames and combines their motions to reconstruct the next frame from the current observation. Each latent action describes regional motion through a shared displacement, while the current frame supplies the visual content for reconstruction. This motion-factorized pretraining provides regional motion representations for subsequent mask-guided semantic decoupling.

As illustrated in Figure 3, a causal encoder extracts K region-level latent actions $\mathbf{z}_t^{1:K}$ from $\mathbf{x}_{\le t+1}$ for each transition, with k indexing the regional factors. A shared motion decoder conditions on the current frame $\mathbf{x}_t$ and each latent action $\mathbf{z}_t^k$ to predict a soft regional assignment $m_t^k(\mathbf{p})$ and pixel-wise motion proposals at target-frame pixel positions $\mathbf{p}$. For each region, we average the pixel-wise motion proposals using the assignment weights to obtain a shared displacement $\mathbf{u}_t^k \in \mathbb{R}^2$. We compose these displacements into a backward flow field $\mathbf{F}_t$ and obtain the reconstructed frame $\widehat{\mathbf{x}}_{t+1}$ through differentiable sampling:

$$\mathbf{F}_t(\mathbf{p}) = \sum_{k=1}^{K} m_t^k(\mathbf{p})\mathbf{u}_t^k, \qquad \widehat{\mathbf{x}}_{t+1}(\mathbf{p}) = \mathbf{x}_t(\mathbf{p} + \mathbf{F}_t(\mathbf{p})) \,. \tag{1}$$

Here, $\mathbf{F}_t(\mathbf{p})$ specifies the displacement from a target pixel to its sampling location in the current frame. Each regional component contributes a shared displacement, while the soft assignments determine their mixture at each pixel. The region-shared displacement constraint encourages pixels with similar motion to be represented by the same latent action.

Motion factorization is learned through next-frame reconstruction,

$$\mathcal{L}_{\text{recon}} = \mathbb{E}\Big[\|\widehat{\mathbf{x}}_{t+1} - \mathbf{x}_{t+1}\|^2\Big] \,. \tag{2}$$

The regional assignments and displacements are learned jointly through this reconstruction objective, without action labels, segmentation masks, or ground-truth optical flow.

3.3 Mask-Guided Latent Action Fusion

We aggregate the pretrained regional motion factors into two latent actions for object interaction and egomotion. Following our insight that pixel-space separation can guide latent-action decoupling, we use limited mask supervision to assign the two representations complementary reconstruction tasks. Reconstruction of object-interaction regions and the surrounding scene encourages specialization in interaction and egomotion, respectively.

As illustrated in Figure 3, a learned router assigns each regional latent action in $\mathbf{z}_t^{1:K}$ to one of two groups, and two cross-attention modules with separate parameters aggregate the assigned latents,

$$\mathcal{G}_t^g = \{k : r_t(k) = g\}, \qquad \mathbf{a}_t^g = \mathcal{A}^g(\mathbf{z}_t^{1:K}; \mathcal{G}_t^g), \quad g \in \{\text{int}, \text{ego}\}. \tag{3}$$

Here, g indexes the interaction or egomotion branch, r_t denotes the assignment given by the learned router, and $\mathcal{G}_t^g$ contains the indices assigned to branch g. The cross-attention aggregator $\mathcal{A}^g$ aggregates only the assigned latents into the fused action $\mathbf{a}_t^g$.

We retain the first-stage decoder backbone and replace its motion output head with an RGB reconstruction head. Each fused latent action $\mathbf{a}_t^g$, together with the current frame $\mathbf{x}_t$, conditions this shared decoder to produce a reconstruction proposal R_t^g. Let M_t^{int} denote the binary object-interaction mask of the target frame $\mathbf{x}_{t+1}$ for transition t. The interaction branch is supervised within this mask, and the egomotion branch within the complementary region $M_t^{\text{ego}} = 1 - M_t^{\text{int}}$,

$$\mathcal{L}_{\text{fusion}} = \sum_{g \in \{\text{int},\text{ego}\}} \mathbb{E}\Big[\|R_t^g - \mathbf{x}_{t+1}\|_{M_t^g}^2\Big] \,. \tag{4}$$

Here, $\|\cdot\|_M^2$ denotes the mean squared error within mask M, and each branch loss is averaged over transitions with nonempty supervision masks. During inference, MoFa-LAM extracts both action representations directly from RGB video without masks.

3.4 Dual-Conditioned Video Diffusion

Pretraining with Decoupled Latent Actions. We use the decoupled interaction and egomotion representations to condition MoFa-World, an egocentric world model built on a pretrained video diffusion transformer. With MoFa-LAM fixed, we extract latent-action sequences from training videos and average them within temporal windows to match the temporal resolution of the video-autoencoder latents.

As illustrated in Figure 3, the two representations enter the world model through separate conditioning pathways. Object interactions depend on local scene content, motivating cross-attention that allows video features to selectively attend to interaction latents. For egomotion, we combine the egomotion latents with spatial position features to generate position-dependent scales and shifts that modulate the video features. These pathways provide content-dependent interaction conditioning and position-dependent egomotion conditioning within one model, without restricting their effects to disjoint image regions.

During training, the observed frames provide clean context, and noise is applied only to future video latents. Let $\mathbf{y}$ denote the clean future video latents and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ Gaussian noise. At noise level $s \in [0, 1]$, we construct $\mathbf{y}_s = (1-s)\mathbf{y} + s\boldsymbol{\epsilon}$ and train the model to predict the corresponding velocity,

$$\mathcal{L}_{\mathrm{WM}} = \mathbb{E}_{\mathbf{y},\mathcal{C},\boldsymbol{\epsilon},s}\left[\left\|v_\theta(\mathbf{y}_s, s; \mathcal{C}) - (\boldsymbol{\epsilon} - \mathbf{y})\right\|^2\right], \tag{5}$$

where v_θ is the velocity predictor with parameters θ, and $\mathcal{C}$ comprises the observed frames, both latent-action sequences, and a text description of the initial scene.

Dataset-Specific Control Adaptation. After pretraining, we adapt MoFa-World to explicit action categories and motion commands by mapping them into the corresponding latent-action spaces. For object interaction, we assign each action category a learnable embedding initialized from the mean interaction latent of labeled training examples in that category. These embeddings replace the inferred interaction latents as cross-attention conditions. For egomotion, a learned mapping converts discrete motion commands into latents for the spatial modulation pathway. Each adaptation starts from the pretrained world model and optimizes the corresponding control mapping and conditioning components using $\mathcal{L}_{\mathrm{WM}}$. The two adapted pathways allow interaction and egomotion commands learned from different datasets to be combined within the same world model.

4 Experiments

4.1 Experimental Setup

Implementation Details. MoFa-LAM uses a causal Transformer to infer $K = 8$ region-level latent actions $\mathbf{z}_t^k \in \mathbb{R}^{32}$ from 16-frame clips with $H = 192$ and $W = 336$. During mask-guided fusion, we finetune the encoder and decoder backbone while learning the interaction and egomotion aggregators. MoFa-World builds on the pretrained Wan2.2-TI2V-5B video diffusion model (Team Wan et al., 2025) and is trained on 81-frame clips at 704×1280 resolution with MoFa-LAM fixed. Additional architecture and training details are provided in Appendix A.

Training Data. We pretrain MoFa-LAM on RGB videos from EgoDex (Hoque et al., 2026), GR1 (Gao et al., 2026), MIND (Ye et al., 2026), and OpenAoE (Li et al., 2026) without action, mask, or optical-flow supervision. For mask-guided fusion, we generate interaction-region masks with GMOS-S (Xie et al., 2026) and manually discard inaccurate predictions. The default setting uses 10% of the clips with available masks. We pretrain MoFa-World on videos from EgoDex (Hoque et al., 2026), GR1 (Gao et al., 2026), MIND (Ye et al., 2026), and HOI4D (Liu et al., 2022). For interaction adaptation, we use eight HOI4D (Liu et al., 2022) action labels (`reachout`/`grasp`/`pickup`/`carry`/`putdown`/`open`/`close`/`dump`). For egomotion adaptation, we use first-person MIND (Ye et al., 2026) navigation controls (`W`/`S`/`A`/`D`: forward / backward / left / right) and camera controls ($\uparrow$ / $\downarrow$ / $\leftarrow$ / $\rightarrow$: up / down / left / right).

Baselines and Model Variants. We compare with AdaWorld LAM (Gao et al., 2025), a representative continuous latent action model that supports action transfer and adaptation in world models and encodes each transition with a single latent action. We train it on the same data as MoFa-LAM, following its original training pipeline and configuration. Based on our pixel-space decoupling insight, we construct two dual-latent variants that both use mask-guided RGB reconstruction to learn separate interaction and egomotion latent actions. *Ours w/ Vanilla LAM Pretraining* initializes from the pretrained AdaWorld LAM, whereas *Ours w/ Motion-Factorized Pretraining*, our full MoFa-LAM model, pretrains region-level motion factors before mask-guided fusion. Sharing the same data and mask supervision, the two variants isolate the contribution of motion-factorized pretraining.

4.2 Functional Specialization of Latent Actions

We evaluate the semantic richness and functional specialization of the learned latent actions through three probing tasks: hand-object interaction classification on HOI4D (Liu et al., 2022), robot action regression on AgiBotWorld (Bu et al., 2025), and relative camera pose estimation on Epic-Kitchens (Damen et al., 2018). All three datasets are unseen during LAM training, allowing us to assess whether these properties generalize beyond the training data.

Evaluation Protocol. We freeze each latent-action model and train a separate MLP probe on each latent action, including $\mathbf{a}_t^{\mathrm{int}}$ and $\mathbf{a}_t^{\mathrm{ego}}$ of MoFa-LAM. For each task, we use standard 75:25 training–

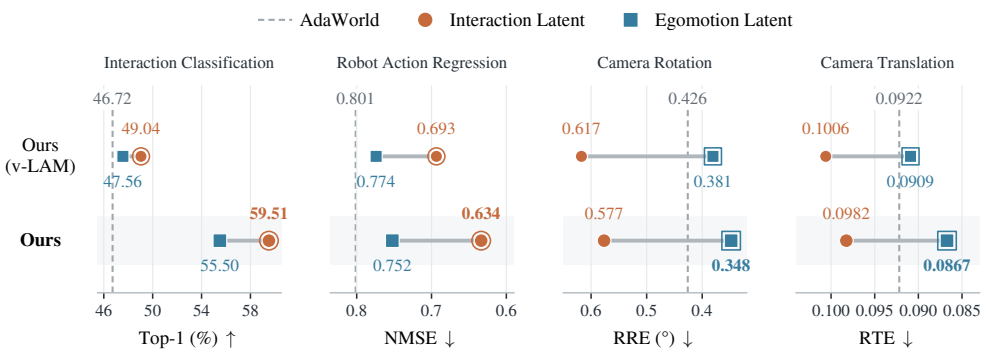

Figure 4: Task-specific latent-action probing. Our interaction latent action performs better on interaction classification and robot action regression, while our egomotion latent action performs better on relative camera pose estimation on Epic-Kitchens in both rotation and translation. Both outperform AdaWorld and the corresponding latent actions of Ours (v-LAM) on their respective tasks. Ours (v-LAM) uses vanilla LAM pretraining instead of Motion Factorization pretraining.

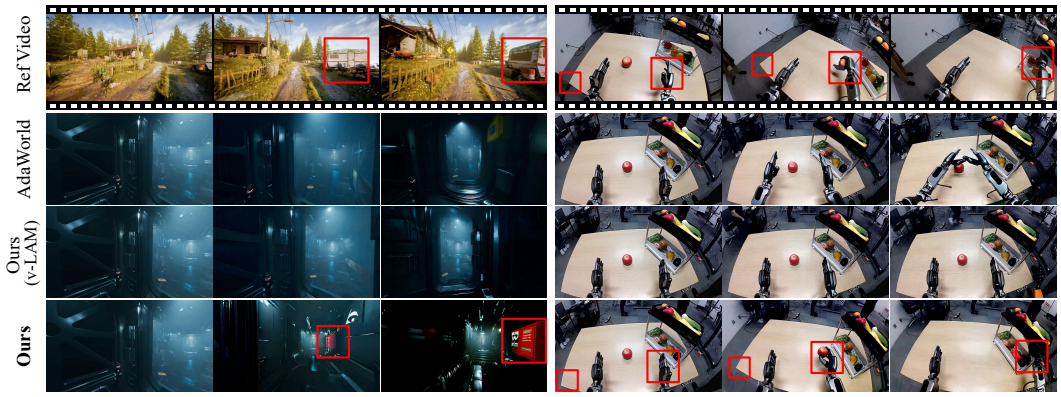

Figure 5: Cross-scene action transfer from reference videos (top row) to new scenes. The examples show egomotion transfer (left) and interaction transfer (right) with accompanying viewpoint changes. Red boxes highlight scene landmarks and interaction regions for comparison. Ours (v-LAM) uses vanilla LAM pretraining instead of Motion Factorization pretraining.

validation splits. For interaction classification, we report Top-1 accuracy over 18 action classes. For robot action regression, we predict 16-dimensional action increments between sampled frames, comprising end-effector translation, rotation, and gripper changes. We report normalized mean squared error (NMSE), with each target dimension standardized using training-set statistics. For camera-pose estimation, we report relative rotation error (RRE, in degrees) and relative translation error (RTE, in normalized dataset units).

Results and Analysis. Figure 4 shows complementary functional specialization of the interaction and egomotion latent actions. Compared with the egomotion latent action, the interaction latent action improves interaction-classification accuracy by 4.01 percentage points and reduces robot action-regression NMSE by 15.8%. Conversely, the egomotion latent action reduces camera RRE and RTE by 39.7% and 11.7%, respectively, compared with the interaction latent action. Motion-factorized pretraining further sharpens this specialization. Relative to the vanilla-LAM variant, the gap between the two latent actions widens on all four metrics: from 1.48% to 4.01% in classification accuracy, from 10.5% to 15.8% in NMSE, from 38.2% to 39.7% in RRE, and from 9.6% to 11.7% in RTE. On their respective tasks, both latent actions outperform AdaWorld and the corresponding representations of *Ours w/ Vanilla LAM Pretraining*. Under the same mask-guided reconstruction supervision, our interaction latent action improves classification accuracy from 49.04% to 59.51% and reduces action-regression NMSE by 8.6% compared with the vanilla-LAM variant. Our egomotion latent action also reduces RRE and RTE by 8.6% and 4.6%, respectively.

Table 1: World-model adaptation to egomotion commands on MIND and interaction labels on HOI4D. Ours (v-LAM) replaces Motion Factorization pretraining with vanilla LAM pretraining.

Method	MIND				HOI4D			
	Visual Quality		Pose Error (RPE)		Visual Quality		Action Accuracy	
	Img. Qual. ↑	Bkg. Cons. ↑	Trans. ↓	Rot. ↓	Img. Qual. ↑	Subj. Cons. ↑	Score ↑	Succ. ↑
AdaWorld	0.496	0.923	0.066	1.735	0.702	0.931	3.700	46%
Ours (v-LAM)	0.482	0.923	0.074	1.744	0.701	0.936	3.720	52%
Ours (MoFa-World)	**0.528**	**0.935**	**0.059**	**1.413**	**0.711**	**0.946**	**4.220**	**72%**



Figure 7: Adapted egomotion control on MIND (left) and interaction control on HOI4D (right). Blue and orange highlights indicate active egomotion and interaction commands, respectively.

4.3 Cross-Scene Action Transfer

Evaluation Protocol. We evaluate whether the pretrained world model can transfer behaviors from reference videos to new scenes without adaptation. We extract interaction and egomotion latent actions from a reference video and use them to condition generation from a target-scene input frame.

Results and Analysis. Figure 5 presents egomotion and interaction transfer examples, comparing AdaWorld, our variant with vanilla LAM pretraining (Ours (v-LAM)), and our MoFa-World. For egomotion transfer, MoFa-World follows the forward camera motion of the reference video while preserving the appearance of the target scene. For interaction transfer, it more closely follows the reference grasping and object-transport sequence despite changes in the scene layout. Our model also captures and transfers the subtle viewpoint changes accompanying the interaction in the reference video. AdaWorld produces a different interaction trajectory, while the vanilla-LAM variant shows limited manipulation progress. These qualitative comparisons suggest that motion-factorized pretraining helps preserve both object interaction and accompanying viewpoint changes during cross-scene transfer.

4.4 World-Model Adaptation and Joint Control

Evaluation Protocol. We evaluate adaptation to explicit egomotion commands on MIND and interaction labels on HOI4D, comparing AdaWorld, our vanilla-LAM variant, and MoFa-World. Generation is conditioned on an input frame and the specified control sequence. We report VBench (Huang et al., 2024) imaging quality and background consistency on MIND, and imaging quality and subject consistency on HOI4D. We evaluate control accuracy using translational and rotational relative pose error (RPE) on MIND, and action-sequence scores (Score, 1–5) and full-score rates (Succ., score = 5) on HOI4D. Details of the evaluation metrics are provided in Appendix B.2.

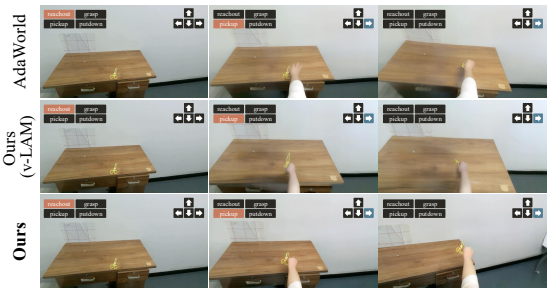

Figure 6: Cross-dataset joint control combining interaction control learned on HOI4D with egomotion control learned on MIND. Highlighted commands specify reaching, picking up the scissors while turning right, and continuing the rightward turn.

Results and Analysis. As shown in Table 1, MoFa-World outperforms both comparison methods across the metrics. On MIND, it achieves lower camera-pose errors together with higher imaging quality and background consistency. On HOI4D, it improves action-sequence scores and full-score rates while maintaining higher imaging quality and subject consistency. Figure 7 illustrates these differences: MoFa-World follows the specified viewpoint changes and more faithfully executes the reaching, pickup, and putdown sequence. The improvements over the vanilla-LAM variant support the value of motion-factorized pretraining for subsequent adaptation to explicit controls.

Cross-dataset Joint Control. We further evaluate whether control capabilities learned from different datasets can be composed within a single world model. We combine interaction control adapted on HOI4D with egomotion control adapted on MIND. As shown in Figure 6, MoFa-World executes the reaching and pickup commands while incorporating the specified rightward viewpoint change. This cross-dataset composition suggests that interaction and egomotion representations retain distinct functional roles beyond their respective adaptation datasets. Appendix C.2 further shows that conditioning on either latent action alone produces the corresponding behavior.

4.5 Ablations and Analysis

We examine the effects of pretraining, mask supervision, and the number of regional latents on interaction classification and camera-pose estimation (Table 2). The default configuration uses motion-factorized pretraining, 10% mask supervision, and $K = 8$ regional latents. We report HOI4D Top-1 accuracy using the interaction latent action and Epic-Kitchens RRE using the egomotion latent action, following the probing protocol in Section 4.2.

Pretraining. We compare motion-factorized pretraining with random initialization and vanilla LAM pretraining (Table 2). Our model outperforms both variants on HOI4D Top-1 accuracy and Epic-Kitchens RRE, supporting motion-factorized pretraining for functional specialization.

Mask Supervision. We vary the mask-supervised fraction from 1% to 10%. With only 5% mask supervision, the model achieves 59.13% HOI4D Top-1 accuracy and 0.363° Epic-Kitchens RRE, close to the default results of 59.51% and 0.348° with 10% supervision. These results indicate that only a small amount of mask supervision is needed to guide the pretrained motion factors into functionally specialized latent actions.

Table 2: Ablations of MoFa-LAM. HOI4D Top-1 and Epic-Kitchens RRE are evaluated using the interaction and egomotion latent actions.

Variant	HOI4D Top-1 (%) ↑	Epic-Kitchens RRE (°) ↓
(a) Pretraining		
w/o pretraining	49.59	0.446
w/ Vanilla LAM Pretraining	49.04	0.381
(b) Mask supervision		
1% masks	53.94	0.384
5% masks	59.13	0.363
(c) Number of regional latents		
$K = 2$	47.81	0.385
$K = 4$	59.14	0.366
Ours (default)	**59.51**	**0.348**

Number of Regional Latents. We vary the number of regional latents $K \in \{2, 4, 8\}$. Increasing K from 2 to 4 raises HOI4D Top-1 from 47.81% to 59.14%, with only a 0.37 percentage-point gain at $K = 8$. Epic-Kitchens RRE is 0.385°, 0.366°, and 0.348° for $K = 2, 4, 8$, respectively. These results support $K = 8$ as an effective default for both tasks, while interaction-classification gains diminish beyond four latents.

5 Conclusion

We presented a framework for learning decoupled latent actions based on the insight that pixel-space motion patterns provide cues about action semantics. MoFa-LAM combines motion-factorized pretraining with limited mask-guided reconstruction supervision to learn functionally specialized representations for object interaction and egomotion. Probing experiments validate this specialization and demonstrate the benefits of motion-factorized pretraining over vanilla LAM pretraining. Conditioned on these representations, MoFa-World supports cross-scene action transfer, dataset-specific control adaptation, and joint use of controls learned from different datasets. These findings support pixel-space motion factorization for learning reusable and composable control representations for interactive world models.

AI USE STATEMENT

We used large language models solely for language editing and improving the clarity of presentation. All technical content, experimental results, numerical values, and citations were independently reviewed and verified by the authors. Generative AI was not used to design experiments, generate results, select baselines, or formulate scientific conclusions.

ETHICS STATEMENT

This work uses existing research datasets for training and evaluation and does not involve new human-subject studies or personal data collection. We use the datasets and pretrained models under their respective licenses and terms of use. As interactive world models may be applied to realistic visual simulation, we encourage responsible use and appropriate safeguards against deceptive or harmful applications.

REPRODUCIBILITY STATEMENT

The main paper describes the model architectures and learning objectives of MOFA-LAM and MOFA-WORLD, including motion-factorized pretraining, mask-guided latent-action fusion, and dual-conditioned video diffusion. Training data, evaluation protocols, baselines, metrics, and ablation settings are specified in the main paper and appendix. Code and models will be released publicly upon publication.

REFERENCES

Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. In *NeurIPS*, 2024.

Tewodros W. Ayalew, Matthew Jeung, Samuel Wheeler, Xiao Zhang, Andre de la Cruz Arce, Kaylene Stocking, Michael Maire, and Matthew R. Walter. CLAW: Learning continuous latent action world models via adversarial latent regularization. *arXiv preprint arXiv:2606.04130*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. In *CVPR*, 2026.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *ICML*, 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 *IROS*, 2025.

Xizhou Bu, Jiexi Lyu, Fulei Sun, Ruichen Yang, Zhiqiang Ma, and Wei Li. LAOF: Robust latent action learning with optical flow constraints. In *CVPR*, 2026.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. GameGen-X: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling Egocentric Vision: The epic-kitchens dataset. In *ECCV*, 2018.

Tal Daniel, Carl Qi, Dan Haramati, Amir Zadeh, Chuan Li, Aviv Tamar, Deepak Pathak, and David Held. Latent particle world models: Self-supervised object-centric stochastic dynamics modeling. In *ICLR*, 2026.

Ashley Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Isbell. Imitating latent policies from observation. In *ICML*, 2019.

Shenyuan Gao, Siyuan Zhou, Yilun Du, Jun Zhang, and Chuang Gan. AdaWorld: Learning adaptable world models with latent actions. In *ICML*, 2025.

Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, Qianli Ma, Seungjun Nah, Loic Magne, Jiannan Xiang, Yuqi Xie, Ruijie Zheng, Dantong Niu, You Liang Tan, K. R. Zentner, George Kurian, Suneel Indupuru, Pooya Jannaty, Jinwei Gu, Jun Zhang, Jitendra Malik, Pieter Abbeel, Ming-Yu Liu, Yuke Zhu, Joel Jang, and Linxi Jim Fan. DreamDojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

Quentin Garrido, Tushar Nagarajan, Basile Terver, Nicolas Ballas, Yann LeCun, and Michael Rabbat. Learning latent action world models in the wild. In *ICML*, 2026.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, 2018.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, Kalyan Sunkavalli, Feng Liu, Zhengqi Li, and Hao Tan. RELIC: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

Ryan Hoque, Peide Huang, David J. Yoon, Mouli Sivapurapu, and Jian Zhang. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *ICLR*, 2026.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, Jiawei Ren, Kevin Xie, Joydeep Biswas, Laura Leal-Taixe, and Sanja Fidler. ViPE: Video pose engine for 3d geometric perception. In *NVIDIA Research Whitepapers*, 2025a.

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*, 2025b.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *CVPR*, pp. 21807–21818, 2024.

Yuxin Jiang, Yuchao Gu, Ivor W. Tsang, and Mike Zheng Shou. Olaf-World: Orienting latent actions for video world modeling. In *ICML*, 2026.

Albina Klepach, Alexander Nikulin, Ilya Zisman, Denis Tarasov, Alexander Derevyagin, Andrei Polubarov, Nikita Lyubaykin, Igor Kiselev, and Vladislav Kurenkov. Object-centric latent action learning. In *AAAI*, 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*, 2025.

Zishuo Li, Bowen Yang, Changtao Miao, Kai Zhu, Hao Chen, Qingze Guan, Zhengxing Wu, Wanke Zhan, Yang Sun, Zhiyi Huang, Zitong Shan, Zhenchao Jin, Jiadong Hong, Taowen Wang, Yushi Feng, You Liu, Yibo Wang, Yifan Yang, Zhaowen Zhou, Man Luo, Hao Cheng, Bo Zhang, Jianshu Li, Jiansheng Cai, Guocai Yao, Jize Zhang, Chenhao Lin, Renjing Xu, Lequan Yu, Chao Shen, Chunhua Shen, and Zhe Li. Open-AoE: An open egocentric manipulation dataset and toolchain for embodied learning. *arXiv preprint arXiv:2607.14183*, 2026.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. HOI4D: A 4d egocentric dataset for category-level human-object interaction. In *CVPR*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Heejeong Nam, Chandradithya S. Jonnalagadda, Harshit Aggarwal, Eric Xu, and Randall Balestriero. Latent actions from factorized transition effects under agent ambiguity. *arXiv preprint arXiv:2606.30544*, 2026.

Alexander Nikulin, Ilya Zisman, Denis Tarasov, Nikita Lyubaykin, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors. In *ICML*, 2025.

Oleh Rybkin, Karl Pertsch, Konstantinos G. Derpanis, Kostas Daniilidis, and Andrew Jaegle. Learning what you can do before doing anything. In *ICLR*, 2019.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *ICLR*, 2024.

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 modeling. In *ICML*, 2026a.

Zhihao Sun, Zhiying Du, Xitong Yang, and Zuxuan Wu. HandWorld: Hand-centric unified video action generation. In *CVPR*, 2026b.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *ICLR*, 2025.

Angel Villar-Corrales and Sven Behnke. PlaySlot: Learning inverse latent dynamics for controllable object-centric video prediction and planning. In *ICML*, 2025.

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

Yucen Wang, Fengming Zhang, De-Chuan Zhan, Li Zhao, Kaixin Wang, and Jiang Bian. Co-evolving latent action world models. In *ICML*, 2026a.

Zizhao Wang, Chang Shi, Jiaheng Hu, Kevin Rohling, Roberto Martín-Martín, Amy Zhang, and Peter Stone. Factored latent action world models. In *ICML*, 2026b.

Ruiqi Wu, Xuanhua He, Meng Cheng, Tianyu Yang, Yong Zhang, Zhuoliang Kang, Xunliang Cai, Xiaoming Wei, Chunle Guo, Chongyi Li, and Ming-Ming Cheng. Infinite-World: Scaling interactive world models to 1000-frame horizons via pose-free hierarchical memory. In *ICML*, 2026.

Junyu Xie, Tengda Han, Weidi Xie, and Andrew Zisserman. GMOS: Grounding moving object segmentation in 3d space and time. *arXiv preprint arXiv:2605.30352*, 2026.

Zhexiao Xiong, Yizhi Song, Hao Kang, Qing Yan, Liming Jiang, Jenson Yang, Zhoujie Fu, Stathi Fotiadis, Angtian Wang, Zichuan Liu, Bo Liu, Yiding Yang, Xin Lu, and Nathan Jacobs. ActWorld: From explorable to interactive world model via action-aware memory. *arXiv preprint arXiv:2606.17730*, 2026.

Sherry Yang, Yilun Du, Seyed Ghasemipour, Jonathan Tompson, Leslie Kaelbling, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *ICLR*, 2024.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, Lars Liden, Kimin Lee, Jianfeng Gao, Luke Zettlemoyer, Dieter Fox, and Minjoon Seo. Latent action pretraining from videos. In *ICLR*, 2025.

Yixuan Ye, Xuanyu Lu, Yuxin Jiang, Yuchao Gu, Rui Zhao, Qiwei Liang, Jiachun Pan, Fengda Zhang, Weijia Wu, and Alex Jinpeng Wang. MIND: Benchmarking memory consistency and action control in world models. *arXiv preprint arXiv:2602.08025*, 2026.

Tianqiu Zhang, Muyang Lyu, Yufan Zhang, Fang Fang, and Si Wu. DiLA: Disentangled latent action world models. In *ICML*, 2026.

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 *ICLR*, 2026.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. IRASim: A fine-grained world model for robot manipulation. In *ICCV*, 2025.

APPENDIX

A IMPLEMENTATION DETAILS

A.1 NETWORK ARCHITECTURE

Latent Action Backbone. MOFA-LAM consists of a causal spatio-temporal encoder and a spatial-only decoder, each with 16 Transformer blocks, hidden dimension 1,024, and 16 attention heads. Frames are tokenized into 16×16 patches, and eight learnable action queries produce 32-dimensional regional latent actions for each transition. For each regional latent, the shared decoder adds its projected embedding to the current-frame patch embeddings and predicts a regional assignment logit and two motion components per pixel. During mask-guided fusion, we retain the input projections and decoder backbone and replace the motion head with a newly initialized RGB head with sigmoid activation.

Latent Action Fusion. The router uses two learnable 64-dimensional group slots and two rounds of context-dependent assignment, with a $32 \to 128 \to 64$ GELU MLP for slot updates. Each regional latent is assigned to one group, with hard routing in the forward pass and soft surrogate gradients during training. Each group has a separate four-block cross-attention aggregator with hidden dimension 128, four attention heads, and one learnable aggregation query. The feed-forward layers use dimensions $128 \to 512 \to 128$, GELU activation, and zero dropout. Input and output projections map between the 32-dimensional action space and the 128-dimensional aggregation space, with LayerNorm before the final output projection. The two aggregators do not share parameters; the encoder, router, and reconstruction decoder are shared across the two branches.

World Model. MOFA-WORLD uses the pretrained Wan2.2-TI2V-5B backbone (Team Wan et al., 2025) with two additional action-conditioning pathways. The 32-dimensional interaction latents are projected and augmented with temporal encodings to form cross-attention context tokens. The egomotion latents are combined with spatial position features to predict position-dependent scales and shifts, applied after self-attention and before cross-attention in the DiT blocks. The final scale and shift prediction head is zero-initialized. For each 81-frame video, we average each latent-action sequence in groups of four transitions, yielding 20 conditions per pathway aligned with the future video-autoencoder time steps. For HOI4D (Liu et al., 2022) adaptation, we optimize category embeddings, the interaction-conditioning adapter, and LoRA (Hu et al., 2022) parameters in the cross-attention layers. For MIND (Ye et al., 2026), we learn an embedding-based MLP that maps motion commands to egomotion latents and optimize it together with the spatial modulation pathway. Both adaptations start from the same pretrained world model and keep all remaining parameters fixed. For joint inference, we combine the HOI4D-adapted category embeddings, interaction adapter, and cross-attention LoRA with the MIND-adapted command mapping and spatial modulation pathway, without additional training.

A.2 BASELINES AND MODEL VARIANTS

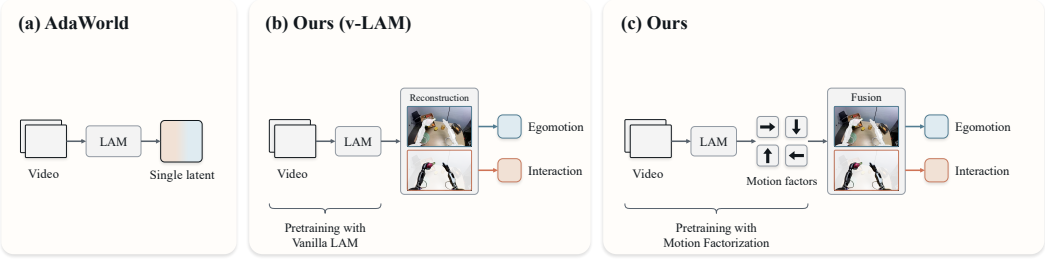

Figure 8: Comparison of latent-action model variants. (a) AdaWorld represents each transition with a single latent action. (b) Ours (v-LAM) uses vanilla LAM pretraining followed by mask-guided RGB reconstruction to learn interaction and egomotion latent actions. (c) Our full model first learns regional motion factors through motion-factorized pretraining, then aggregates them through mask-guided fusion. Blue and orange denote egomotion and interaction, respectively.

Figure 8 summarizes the baseline and the two variants used in our experiments. All three latent-action models are pretrained on the same video sources. Both dual-latent variants follow our core insight that pixel-space separation can guide latent-action decoupling, using mask-guided RGB reconstruction to learn interaction and egomotion representations.

AdaWorld. We train the AdaWorld LAM (Gao et al., 2025) following its original training pipeline and configuration. As shown in Figure 8(a), it encodes each transition into a single latent action and reconstructs the next RGB frame conditioned on the current frame.

Ours with Vanilla LAM Pretraining. This variant, denoted Ours (v-LAM), initializes from the pretrained AdaWorld LAM and learns two action queries through mask-guided RGB reconstruction. The two latent actions condition a shared decoder and are supervised on object-interaction regions and the complementary scene regions, respectively, as illustrated in Figure 8(b).

Ours with Motion-Factorized Pretraining. Our full MoFA-LAM model first learns K regional latent actions through motion-factorized reconstruction, then aggregates them into interaction and egomotion latent actions using mask-guided fusion (Figure 8(c)). Both dual-latent variants use the same mask-guided RGB reconstruction supervision, allowing their comparison to assess the benefit of motion-factorized pretraining.

A.3 Training Details

Latent Action Pretraining. We pretrain MoFA-LAM on 16-frame RGB clips at 192×336 resolution, with eight 32-dimensional regional latent actions per transition. We use AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-5}, weight decay 10^{-2}, and an effective batch size of 32 across 16 A100 GPUs. We pretrain for approximately 150k steps and use the resulting checkpoint to initialize mask-guided fusion.

Mask-Guided Fusion. We train for 15k optimizer steps using AdamW with weight decay 10^{-2} and an effective batch size of 32 across 16 A100 GPUs. The pretrained encoder and decoder backbone use a learning rate of 10^{-5}, while the router, aggregators, and newly initialized RGB head use 10^{-4}. Learning rates remain constant throughout training. The default setting uses 10% of the clips with available masks. The 1% and 5% mask-supervision variants are initialized from the same pretrained checkpoint and trained for the same 15k steps.

World-Model Training and Adaptation. We train MoFA-World on 81-frame clips at 704×1280 resolution, keeping MoFA-LAM fixed and updating the LoRA parameters and action-conditioning modules with the video-generation objective. The initial frame provides clean visual context, and its description generated by Qwen3-VL (Bai et al., 2025) provides scene-level text conditioning during pretraining. We use AdamW with a learning rate of 10^{-4}, weight decay 10^{-2}, and a total batch size of 16 across 16 A100 GPUs. We train the world model for 20k steps. For dataset-specific adaptation, we initialize each model from the pretrained world model and update the parameters specified in Appendix A.1. Interaction-category embeddings are initialized from training-set latent means, whereas the egomotion-command mapping is randomly initialized. Both adaptation stages use AdamW with zero weight decay. We use learning rates of 5×10^{-5} for HOI4D and 10^{-4} for MIND. Both adaptations run for 2k steps with a total batch size of 16 across 16 A100 GPUs. World-model pretraining and both adaptations use one clip per GPU without gradient accumulation.

B Evaluation Details

B.1 Latent-Action Probing

We evaluate frozen latent-action models on interaction classification using HOI4D (Liu et al., 2022), robot action regression using AgiBotWorld-Beta (Bu et al., 2025), and relative camera pose estimation using Epic-Kitchens (Damen et al., 2018) with OmniWorld (Zhou et al., 2026) annotations. Inputs contain 16 RGB frames at 192×336 resolution. We use standard 75:25 training–validation splits, disjoint by clip for HOI4D, by task for AgiBot, and by participant for Epic-Kitchens.

Each latent representation is evaluated with a separate two-layer MLP with hidden width 256, LayerNorm, GELU, and dropout of 0.1. We train with AdamW for 60 epochs using learning rate 10^{-3},

weight decay 10^{-4}, batch size 1,024, and cosine learning-rate decay. Features and robot regression targets are standardized using training-set statistics.

HOI4D classification retains 18 action classes after excluding `unknown`, `Lookaround`, and `go`, and uses cross-entropy with label smoothing of 0.1. Robot regression predicts 16-dimensional bimanual action increments from frame-aligned proprioceptive annotations using MSE. Camera targets are relative poses derived from OmniWorld Epic-Kitchens reconstructed extrinsics. We normalize translation by the RMS radius of the camera centers over each reconstruction segment. Rotation and translation probes have separate parameters; translation targets are additionally standardized using training-set statistics for optimization and converted back before evaluation. We report Top-1 accuracy, robot action NMSE, rotation geodesic error in degrees, and full-vector translation L2 error in these segment-normalized units.

B.2 World-Model Evaluation Metrics

Egomotion Accuracy. We use ViPE (Huang et al., 2025a) to estimate camera trajectories from both generated and reference videos and compute translational and rotational relative pose error (RPE) against the estimated reference trajectories. For translational RPE, we apply per-clip Sim(3) alignment to the generated trajectory before comparison. Translation is measured in reference trajectory units and rotation in degrees.

Interaction Accuracy. We use Qwen3-VL to assess adherence to the specified action sequence over the entire generated clip, assigning a sequence-level score from 1 to 5. Scores are assigned based on visible action execution and ordering:

- **1:** The requested actions are essentially absent.
- **2:** Only a small part of the sequence is performed, or a major ordering error occurs.
- **3:** A substantial part of the sequence is performed in broadly correct order, but major actions are missing or incomplete.
- **4:** Most requested actions are completed in the correct order, with only minor omissions or incomplete execution.
- **5:** All requested actions are visibly completed in the specified order.

Score denotes the mean sequence-level score, and Succ. denotes the percentage of clips receiving a score of 5.

Video Quality. We evaluate visual quality using three VBench metrics (Huang et al., 2024). *Imaging quality* assesses frame-level perceptual quality using MUSIQ, with scores averaged over the generated frames. *Background consistency* measures temporal consistency of scene appearance using CLIP image features, while *subject consistency* measures consistency of subject appearance using DINO features. Both consistency metrics compare each frame with the first frame and the preceding frame using cosine similarity. We report imaging quality on both datasets, background consistency on MIND to assess scene coherence during viewpoint changes, and subject consistency on HOI4D to assess appearance preservation during interaction. Higher values are better for all three metrics.

C Additional Visualizations

C.1 Video Demo

We provide a supplementary demo video (MoFA-World_demo.mp4) showcasing the qualitative results of MoFA-World.

C.2 Separate Control

We further examine the functional roles of the two latent actions through conditional video generation. Figure 9 compares generation using only the egomotion latents, only the interaction latents, and both, with the target video shown for reference. With egomotion conditioning alone, the generated sequence exhibits viewpoint changes, as indicated by the highlighted table corner, while the object

remains near its initial position in the hand. With interaction conditioning alone, the hand moves the object toward the shelf while the viewpoint remains relatively stable. Using both latent-action sequences combines the object manipulation with the accompanying viewpoint changes, more closely matching the target behavior. These examples illustrate separate and combined control of object interaction and egomotion.

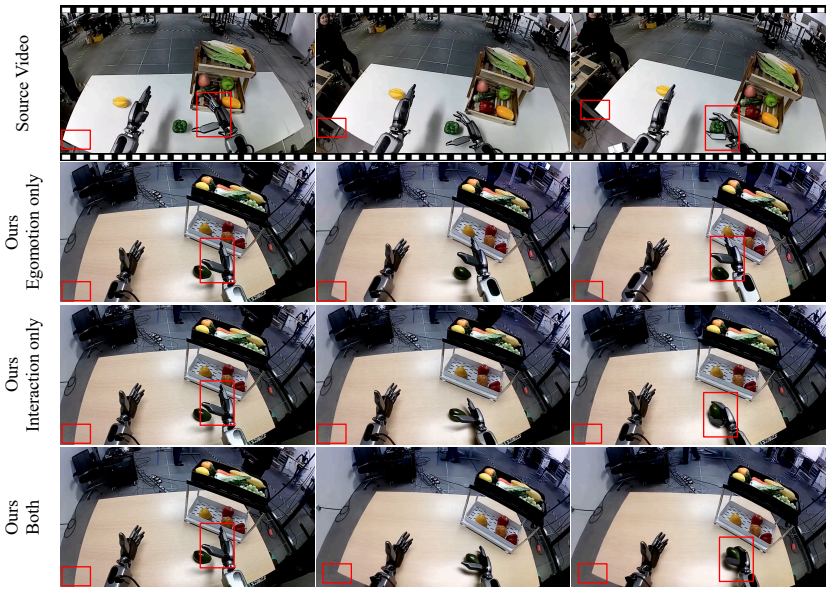

Figure 9: Separate interaction and egomotion control. Rows show the target video and generations conditioned on egomotion latents only, interaction latents only, and both. Red boxes highlight the table corner and the manipulated object to illustrate viewpoint changes and object motion.

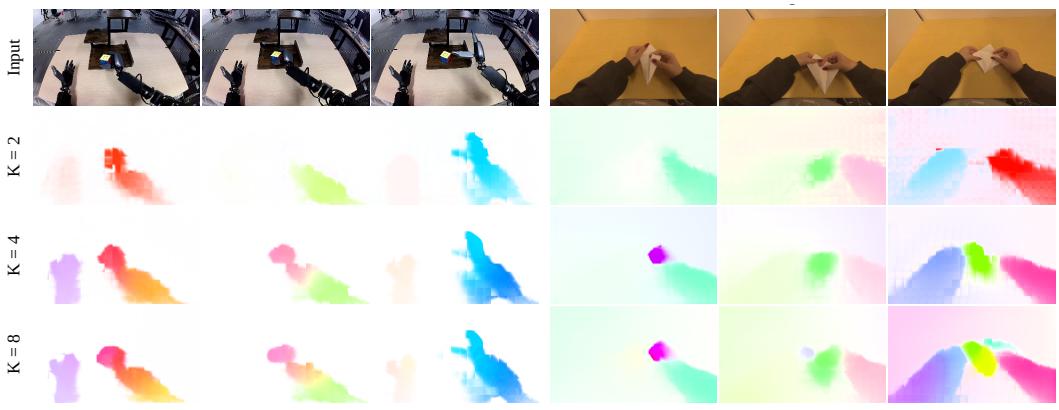

Figure 10: First-stage flow predictions with different numbers of regional latents. Left: GR1; right: EgoDex. Rows show the input target frames and predicted target-to-source flow for $K = 2, 4, 8$; each scene contains three time points. Hue encodes flow direction and saturation encodes magnitude, with white denoting zero motion. With $K = 8$, motion-factorized pretraining produces sharper and more coherent flow that separates the moving hands, robot arms, and manipulated objects from the surrounding scene.

C.3 Flow Visualization

Figure 10 compares first-stage flow predictions with $K \in \{2, 4, 8\}$ regional latents on GR1 (Gao et al., 2026) and EgoDex (Hoque et al., 2026) sequences. All models receive the same frames

within each example. With $K = 2$, the predicted flow exhibits coarse regions and block artifacts; $K = 4$ captures more localized motion, while $K = 8$ shows finer structure around the hands and manipulated objects, delineating dynamic objects as distinct motion regions without any mask or flow supervision.

C.4 Additional Qualitative Results

Cross-Scene Action Transfer. Figures 11 and 12 provide additional cross-scene transfer comparisons with AdaWorld and Ours (v-LAM). For egomotion transfer, both comparison methods make little forward progress in the target scenes, whereas MoFa-World reproduces the forward motion and turns of the reference trajectories while preserving the appearance of each new environment. For interaction transfer, MoFa-World completes the reference manipulation on the target objects, including grasping and carrying a toy car, manipulating a folder lid, and a robotic pick-and-place. It also transfers the viewpoint change that accompanies the toy-car transport, which AdaWorld does not reproduce, while Ours (v-LAM) shows limited manipulation progress. These results show that the two latent actions capture behaviors that transfer regardless of scene appearance, and that motion-factorized pretraining is important for transferring both behaviors faithfully.

Dataset-Specific Control Adaptation. Figures 13 and 14 provide additional comparisons after adaptation to MIND egomotion commands and HOI4D interaction labels, respectively. For each example, AdaWorld, Ours (v-LAM), and MoFa-World receive the same initial frame and control sequence. In these examples, MoFa-World produces clearer manipulation progress and more coherent scene appearance during commanded viewpoint changes, while the comparison methods exhibit incomplete interactions or visual distortions.

Joint Interaction and Egomotion Control. Figure 15 shows two examples combining interaction controls learned on HOI4D with egomotion controls learned on MIND within a single rollout.

D Limitations

Our evaluation focuses on egocentric world modeling, with broad applications in interactive simulation and embodied intelligence. The evaluated scenes primarily involve object interaction and egomotion, with limited independent background motion. Evaluation across other viewpoints and more complex background dynamics remains part of our future work.

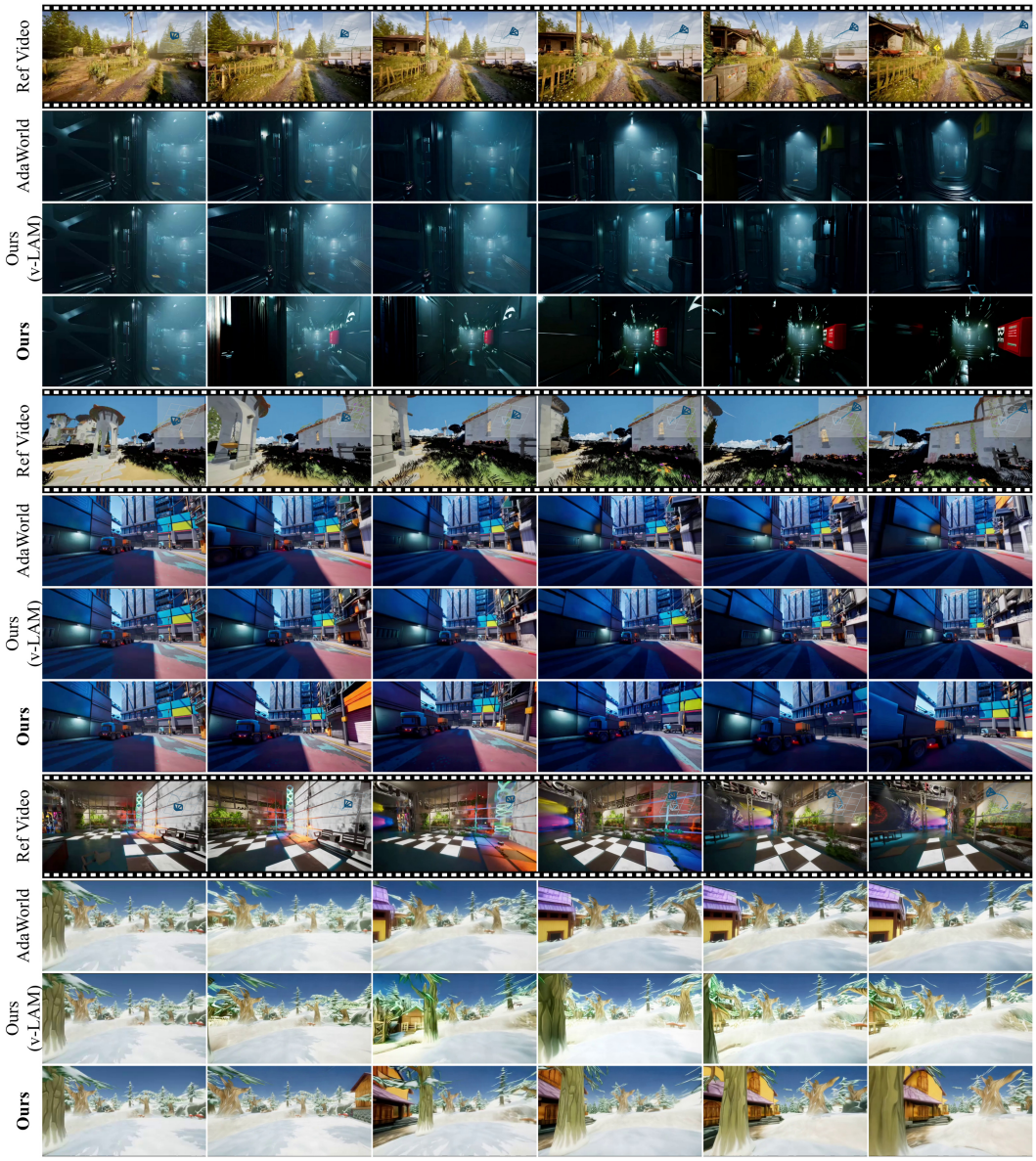

Figure 11: Additional cross-scene egomotion transfer results. The three examples are arranged from top to bottom; within each example, rows show the reference video, AdaWorld, Ours (v-LAM), and Ours. Six uniformly sampled frames, including the first and last frames, show the progression of the reference video and each generated sequence in the target scene; Figure 12 follows the same layout. Insets in the reference row illustrate the recorded camera trajectories relative to the initial pose. Gray shows the full trajectory, blue the traversed path and current pose, and the pale marker the initial pose. AdaWorld and Ours (v-LAM) make little forward progress, whereas Ours follows the forward motion and turns of the reference, with scene structures approaching consistently in each new environment.

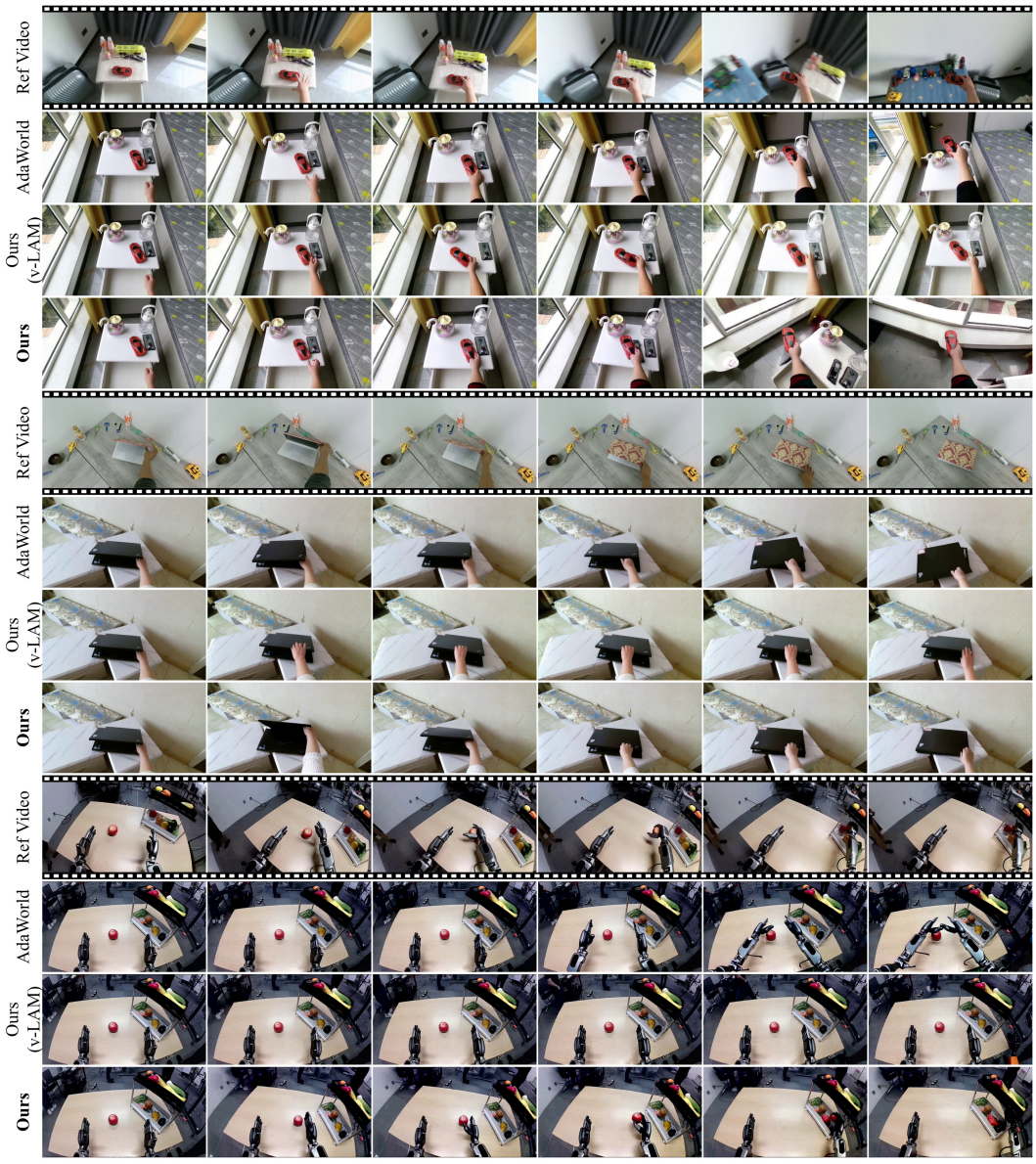

Figure 12: Additional cross-scene interaction transfer results. The three examples, arranged from top to bottom, show toy-car manipulation, folder manipulation, and robotic object manipulation. Within each example, rows show the reference video, AdaWorld, Ours (v-LAM), and Ours. Ours (v-LAM) shows little manipulation progress, while AdaWorld misses the viewpoint change accompanying the toy-car transport, deforms the folder, and produces a different bimanual motion in the robot scene. Ours reproduces the grasp-and-carry with its accompanying viewpoint change, the lid manipulation, and the robotic pick-and-place, adapting each interaction to the layout of the target scene.



Figure 13: Additional egomotion control results after adaptation on MIND. Three examples are arranged from top to bottom, with rows comparing AdaWorld, Ours (v-LAM), and MoFa-World (Ours) within each example. Columns show successive video frames. Blue highlights indicate the active navigation and camera commands.



Figure 14: Additional interaction control results after adaptation on HOI4D. From top to bottom, the examples show chair manipulation, drawer opening and closing, and laptop opening. Within each example, rows compare AdaWorld, Ours (v-LAM), and MoFa-World (Ours), and columns show successive video frames. Orange highlights indicate the active interaction commands.

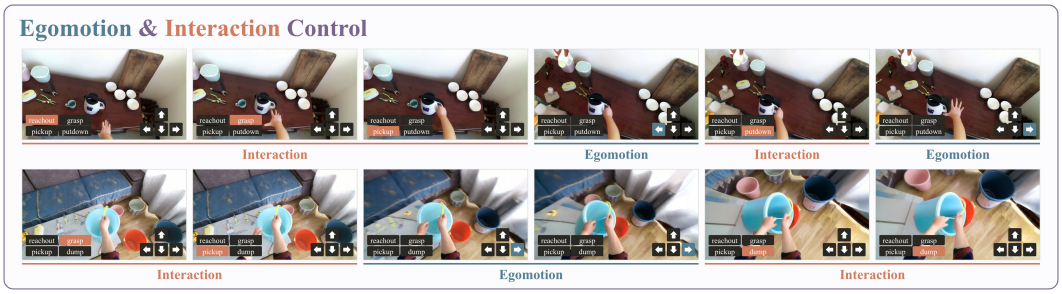

Figure 15: Additional cross-dataset joint control results of MoFa-World. Each row shows a generated sequence combining HOI4D interaction commands with MIND camera commands. Orange and blue highlights mark interaction and egomotion commands, respectively; the colored bars indicate the corresponding control phases. The upper sequence interleaves object manipulation with leftward and rightward turns, while the lower sequence combines picking up a container, turning right, and dumping.