HANDWM: INTERACTIVE VIDEO WORLD MODELS WITH EXPLICIT HAND-OBJECT DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video world models aim to predict future visual observations conditioned on predefined controlling signals. While most existing methods focus on exploration (i.e., controlling the observation pose), the recent emergence of interactive world models begins to control the behavior of an agent in the video and model their interaction with the world. In this paper, we focus on hand–object interaction which poses a particular challenge because future observations depend on both 3D interaction dynamics and complex visual appearance. Direct action-conditioned video generation couples dynamics and appearance, leading to inaccurate interaction predictions. We propose a factorized interactive world modeling framework that disentangles explicit 3D interaction dynamics from visual observation synthesis. Given an initial observation and future hand controls, we first predict the 3D object geometry and the initial 3D hand pose and then leverage a simulator-based interaction module to produce an explicit 3D hand–object trajectory. We introduce pixel-aligned surface conditioning to rasterize future hand–object surfaces into the target view. The resulting maps provide pixel-aligned geometric guidance with explicit visibility and occlusion information. We further propose a bidirectional geometry–appearance interaction module to enable spatially aligned control and generation features to exchange information during synthesis. On DexYCB, our model reduces FVD from 31.9 to 16.9 and hand-pose MPJPE from 31.1 to 21.5 mm relative to the state of the art while preserving motion fidelity.

1 INTRODUCTION

Video world models aim to predict future visual observations conditioned on an initial scene and predefined control signals (Bar et al., 2025; Akkerman et al., 2025). A prominent line of work focuses on navigation and scene exploration, controlling the observation viewpoint through camera trajectories or navigation actions (Bar et al., 2025; He et al., 2025). Recent interactive world models extend this capability to controlling the behavior of an agent within the video and predicting its interactions with the surrounding world (Akkerman et al., 2025; Wang et al., 2026).

In particular, hand–object interaction (HOI) world models focus on predicting how articulated hand motions and contact change object states while maintaining realistic visual appearance (Wang et al., 2026; Li et al., 2026). Most existing action-conditioned HOI video models directly generate future frames from an initial observation and hand controls, coupling interaction dynamics with visual observation synthesis (Akkerman et al., 2025; Wang et al., 2026; Li et al., 2026). Under this formulation, the generator must jointly learn interaction dynamics and visual appearance from the same pixel-space objective. Also, geometrically determined structure is therefore inferred jointly with appearance variation. Learning 3D interaction dynamics only through pixel-level supervision is inefficient. The model is supervised through image differences, so errors in 3D hand–object states are only reflected indirectly in the training loss. Since HOI videos with accurate 3D hand–object annotations are valuable and expensive to obtain, the available supervision should be used to expose structured geometry rather than forcing the generator to rediscover it from RGB losses alone.

We propose a factorized interactive world modeling framework that disentangles explicit 3D interaction dynamics from visual observation synthesis as shown in Figure 1. Given an initial observation and future hand controls, we estimate the initial 3D hand–object state and use a simulator-based interaction module to generate a future 3D trajectory. We introduce Pixel-Aligned Surface Condi-

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



Figure 1: **Factorized interactive world modeling.** **(a)** Conventional HOI video world models directly map an initial observation and future hand controls to RGB video, entangling interaction dynamics with visual appearance. **(b)** HandWM explicitly evolves hand–object interactions in a structured 3D state space and then projects the resulting trajectory into the observation space to guide photorealistic video synthesis.

tioning (PASC) to rasterize future hand–object surfaces into the target view. The resulting maps provide pixel-aligned geometric guidance with explicit visibility and occlusion information. Since the rendered geometry and target RGB video are spatially aligned observations of the same underlying 3D state, we further introduce bidirectional geometry–appearance interaction between the control and generation streams. Finally, we address motion-induced degradation in hand–object videos through interaction-aware data enhancement to improve the geometric supervision used for training. HandWM treats explicit 3D interaction structure as an intermediate representation to be exploited, rather than requiring the video generator to recover it implicitly from pixel-level supervision.

We evaluate whether the observation model in our factorized framework improves both video fidelity and interaction-centered fidelity. Under a unified DexYCB protocol, our observation model reduces FVD from 31.9 to 16.9 and hand-pose MPJPE from 31.1 to 21.5 mm relative to PAM (Gao et al., 2026), while also improving hand-crop FID and LPIPS. Ablations also show that replacing bidirectional geometry–appearance interaction with one-way full-field injection severely degrades performance, and a condition-detail probe shows that generated observations respond measurably to fine-grained PASC information.

2 Related Work

HOI Video World Models. Recent video generative models have increasingly been used to model interactive dynamics. InterDyn (Akkerman et al., 2025) treats video diffusion models as implicit physics simulators and generates interaction dynamics from an initial image and the motion of a driving entity. More recent works specialize this paradigm to hand–object interaction (HOI). SCAR (Yan et al., 2025) introduces structure- and contact-aware representations to improve the physical consistency of generated interactions, while HandWorld (Sun et al., 2026) jointly models egocentric videos and hand actions through shared cross-domain conditioning. Hand2World (Wang et al., 2026) generates long-horizon egocentric interactions from free-space hand gestures, and Ego-HOI (Li et al., 2026) incorporates geometry- and kinematics-derived priors to predict interaction rollouts without access to future object states. They demonstrate the potential of video models for interactive prediction, but generally integrate interaction reasoning and visual synthesis within the generative model. Our factorized interactive world modeling framework disentangles explicit 3D interaction dynamics from visual observation synthesis. Interaction dynamics evolve in a structured 3D state space, while a dedicated observation model translates the resulting states into video.

Geometry-Conditioned Video Generation. Spatial conditioning has been widely explored for controllable visual generation. ControlNet (Zhang et al., 2023) conditions diffusion models on image-space signals such as depth, segmentation, and pose, while subsequent video methods extend structured control to motion and 3D geometry (Zhang et al., 2024; Peng et al., 2024). In HOI generation, Re-HOLD (Fan et al., 2025) uses specialized hand and object layouts for interaction reenactment, and PAM (Gao et al., 2026) combines depth, segmentation, and keypoint conditions for control-

lable HOI video synthesis. Recent methods further exploit explicit 3D information. HVG-3D (Chen et al., 2026) introduces a 3D-aware ControlNet to encode geometric and motion cues from real or simulated 3D conditions, while DWM (Kim et al., 2026) conditions interaction generation on static scene renderings and egocentric hand-mesh renderings. We focus on a complementary perspective: the key question is not only *what* geometric cues to provide, but *which 3D-to-2D correspondences should be resolved before generation* and how the generator should use them. Pixel-Aligned Surface Conditioning (PASC) rasterizes posed hand–object surfaces into target-image coordinates, resolving projection, visibility, and occlusion while leaving appearance synthesis to the video model.

3D Hand–Object Interaction and Data Annotation. Structured 3D representations are fundamental to modeling hand–object interactions. MANO (Romero et al., 2017) provides a compact parametric representation of articulated hands, while datasets such as DexYCB (Chao et al., 2021), H2O (Kwon et al., 2021), OakInk2 (Zhan et al., 2024), and EgoDex (Hoque et al., 2025) provide increasingly rich supervision for hand pose, object pose, and dexterous interaction geometry. Such annotations enable explicit interaction states, but the quality of their corresponding RGB observations has received less attention. Motion blur is a long-standing challenge in video restoration (Su et al., 2017; Nah et al., 2017) and is particularly relevant to HOI, where rapid hand motion can degrade fine geometric evidence around fingers, contacts, and object boundaries. We study this geometry-sensitive supervision problem across multiple HOI datasets and use motion-aware restored supervision as a targeted training intervention.

3 Method

3.1 Problem Formulation and Overview

Given an initial RGB observation I_0 and future hand controls $A_{1:T}$, the goal of an HOI video world model is to predict future visual observations $I_{1:T}$. Existing action-conditioned models directly synthesize the future video (Akkerman et al., 2025; Wang et al., 2026; Li et al., 2026):

$$I_{1:T} \sim G_{\text{direct}}(I_0, A_{1:T}), \tag{1}$$

where G_{direct} is an action-conditioned video generator.

This formulation couples interaction dynamics and visual appearance within the same pixel-space objective. Image reconstruction errors mix discrepancies in motion and appearance, providing only indirect supervision for the underlying 3D state evolution. We therefore propose a factorized interactive world modeling framework that disentangles explicit 3D interaction dynamics from visual observation synthesis. This allows the interaction module to model state evolution explicitly and the observation model to synthesize video conditioned on the resulting trajectory.

Given the initial observation, we estimate the object geometry and initial hand pose to obtain the initial 3D state S_0. A simulator-based interaction module then uses future hand controls to produce a 3D hand–object trajectory $S_{1:T}$. We convert this trajectory into pixel-aligned geometric observations $R_{1:T}$, which guide the video generator together with the initial RGB observation:

$$\begin{aligned} S_{1:T} &= F_{\text{interaction}}(S_0, A_{1:T}), \\ R_t &= \mathcal{R}(S_t; K), \quad t = 1, \dots, T, \\ I_{1:T} &\sim G_{\text{obs}}(I_0, R_{1:T}). \end{aligned} \tag{2}$$

Here, $F_{\text{interaction}}$ evolves the 3D state, $\mathcal{R}$ projects it into the target view using camera parameters K, and G_{obs} synthesizes the video.

3.2 Explicit Hand–Object Interaction Dynamics

Given an initial observation I_0 and future hand controls $A_{1:T}$, we first estimate the 3D object geometry and initial 3D hand pose to obtain the initial state S_0, and then simulate the hand–object trajectory $S_{1:T}$ under the controls. Each state $S_t = \{H_t, O_t\}$ consists of the hand state H_t (MANO (Romero et al., 2017) shape β, finger articulation θ_t, and wrist pose $g_t \in SE(3)$) and the object state O_t (mesh M_o and pose $P_t \in SE(3)$). Each control A_t specifies a target articulation and wrist pose. Besides I_0, we use the aligned depth, camera calibration, and object labels provided by DexYCB.

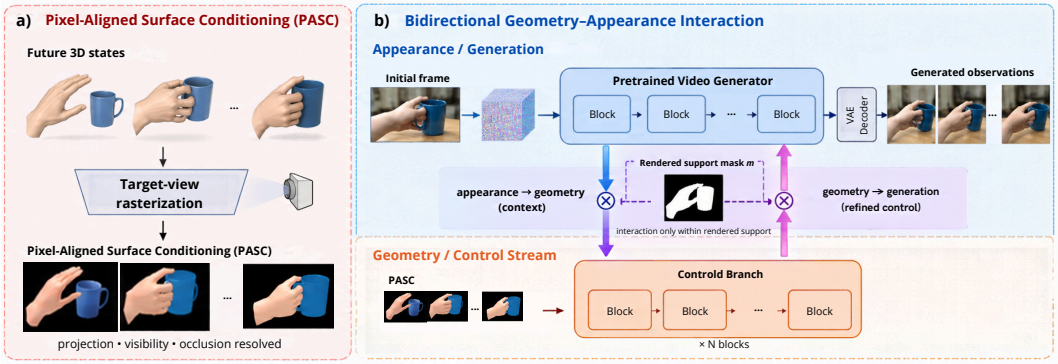

Figure 2: **Overview of the observation model. (a)** PASC rasterizes future 3D hand–object states into the target view, resolving projection, visibility, and occlusion before generation. **(b)** The resulting pixel-aligned geometry drives a control branch that interacts bidirectionally with the pretrained video generator. Appearance features contextualize the geometric control, which is then injected back into the generation stream.

3D Object Geometry Estimation. We segment the object in I_0 by detecting it with Grounding DINO (Liu et al., 2024), prompted with its category name, and refining the detected box into a mask with SAM 2 (Ravi et al., 2025). TRELLIS.2 (Xiang et al., 2025) then reconstructs a textured mesh M_o from the masked object crop. Because the generated mesh lies in a normalized canonical frame, we recover its metric scale and initial pose P_0 from the depth aligned with I_0: masked depth pixels are back-projected to world coordinates, and the mesh is registered to this partial point cloud by a trimmed ICP (Besl & McKay, 1992) with closed-form similarity updates (Umeyama, 1991), initialized by aligning its principal axis with the calibrated up direction. For simulation, M_o is decomposed into convex parts with CoACD (Wei et al., 2022) and assigned its YCB catalog mass.

3D Hand-Pose Estimation. We estimate H_0 with WiLoR (Potamias et al., 2025), which detects the hand and regresses its MANO parameters. We then re-fit the translation to the predicted 2D keypoints under the calibrated intrinsics and map the hand to world coordinates. Applied to the subsequent frames of the source clip, the same estimator yields a reference trajectory with a shared shape β (the median of per-frame estimates). In our experiments, $A_{1:T}$ is either this trajectory or an approach–closure–lift sequence built from one of its grasp configurations.

Simulator-based Interaction. We instantiate $F_{\text{interaction}}$ with the MuJoCo physics simulator (Todorov et al., 2012). The hand is an articulated rigid body with a free wrist and 15 ball joints, built from the MANO skeleton with shape β. The palm and distal finger segments collide through convex hulls of their skinning-assigned vertices, and the other phalanges through capsules. Rather than replaying the controls kinematically, we track their time interpolation with gravity-compensated PD controllers, whose outputs are clipped to bound the wrist force, wrist torque, and each finger-joint torque by 50 N, 3 N·m, and 0.6 N·m in norm. The object is a free rigid body driven only by contact and gravity, without pose replay, attachment to the hand, or external force. Contacts use a friction coefficient of 1 with torsional and rolling resistance, and the table is a finite slab whose height is fitted to the background depth around the object.

Because independently estimated hand and object geometries can interpenetrate, grasping controls are first made contact-consistent: a bounded projected least-squares fit moves the distal finger segments of the grasp configuration into light contact with the object at its initial pose while keeping the other hand parts clear. A second fit with fingertip clearance gives the pregrasp, which the hand reaches through an intermediate waypoint. Optionally, a linear program over linearized friction cones adds hand-only feedforward torques that support the object weight within these limits. We integrate with a 0.1 ms time step and record S_t every 5 ms, converting hand joints back to MANO parameters. The observation model G_{obs} receives these states only through PASC (Secs. 3.3 and 4.3).

3.3 From Dynamics to Appearance

As illustrated in Figure 2, the observation model converts the explicit 3D trajectory into RGB video through pixel-aligned geometric conditioning:

$$I_{1:T} \sim G_\theta(I_0, R_{1:T}), \tag{3}$$

where G_θ instantiates G_{obs} as a latent video diffusion model built on CogVideoX-5B-I2V (Yang et al., 2024). The initial observation provides appearance context, while the PASC sequence specifies the projected interaction geometry. Both RGB and PASC are encoded by the same VAE, producing spatially aligned latent representations for the generation and control streams.

Pixel-Aligned Surface Conditioning. Pixel-Aligned Surface Conditioning (PASC) converts the predicted 3D hand–object trajectory into conditioning maps for video generation. Let $S_t = \{H_t, O_t\}$ denote the 3D interaction state at frame t, where H_t contains the articulated hand geometry and pose and O_t contains the object geometry and pose. Given camera parameters K, we rasterize the state into the target view:

$$R_t = \mathcal{R}(S_t; K), \qquad R_t \in \mathbb{R}^{H \times W \times C}, \tag{4}$$

where $\mathcal{R}$ denotes the rasterizer and H, W, and C are the image height, width, and number of control channels, respectively. The resulting PASC map R_t is pixel-aligned with the target frame. Because the surfaces, poses, and camera parameters are known, rasterization directly determines their projected locations and visibility. This resolves hand–object occlusions before generation.

We render PASC on a black background without scene lighting. Object vertex colors are baked from I_0 to retain the observed object appearance. The hand uses temporally consistent colors derived from fixed normal-based shading. Each foreground pixel thus corresponds to a visible 3D surface point, allowing the sequence $R_{1:T}$ to provide pixel-aligned geometric guidance for the video generator.

Bidirectional Geometry–Appearance Interaction. We augment the diffusion transformer (Peebles & Xie, 2023) with a parallel 12-block control branch. Projection layers connect the control and generation streams. PASC and RGB provide complementary observations of the same projected interaction. A one-way full-field control pathway does not explicitly exploit this correspondence and can inject control into unsupported background regions. At control block l, let h_l and c_l denote the generation and control token features, respectively. Their interaction is

$$\tilde{c}_l = c_l + \phi_l(h_l) \odot m, \qquad h'_l = h_l + \psi_l(\tilde{c}_l) \odot m, \tag{5}$$

where $\tilde{c}_l$ is the appearance-refined control feature, h'_l is the geometry-conditioned generation feature, ϕ_l and ψ_l are linear projections, $\odot$ denotes element-wise multiplication, and m is a binary token-level mask derived from the rendered PASC support. The first update transfers appearance context from the generation stream to the control stream, while the second injects the refined geometric control back into generation. Masking confines both interactions to the spatial support of PASC.

Diffusion Video Generation. We encode the initial observation and PASC sequence as conditioning inputs and initialize the video latent with Gaussian noise. At each sampling step τ, the generation stream processes the current video latent z_τ with the initial-frame conditioning, while the control stream processes the PASC features. The two streams exchange geometry and appearance information through the bidirectional interactions described above. The transformer predicts a denoising velocity $v_\theta(z_\tau, \tau, I_0, R_{1:T})$, which the diffusion sampler uses to update the video latent for the next step. We repeat this process over the sampling schedule with fixed conditioning inputs. The VAE decoder maps the final denoised latent to the predicted RGB video $I_{1:T}$.

3.4 Training

Training Objective. During training, we encode the target RGB video into a clean latent z_0 and sample a diffusion timestep τ and Gaussian noise ϵ. We obtain the noisy latent z_τ using the backbone's diffusion noise schedule. We train the observation model with the standard diffusion velocity-prediction objective (Salimans & Ho, 2022)

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{z_0, \epsilon, \tau}\left[\left\| v - v_\theta(z_\tau, \tau, I_0, R_{1:T}) \right\|_2^2\right], \tag{6}$$

where v is the target diffusion velocity and v_θ is the velocity predicted by the transformer conditioned on the initial observation and PASC sequence. The control branch is initialized from the corresponding pretrained transformer blocks, and the projection layers connecting the two streams are initialized to zero (Zhang et al., 2023).

Interaction-Aware Data Enhancement. To improve supervision in fast-moving hand regions, we use interaction-aware data enhancement based on video restoration. We replace hand windows with

Table 1: **Quantitative results on DexYCB s0.** Upper: published results compiled in the PAM publication under different resolutions and protocols. Lower: PAM and our model under the unified protocol in Sec. 4.1. "–" denotes unavailable results. MPJPE is in mm.

	Full frame					Hand crop					
Method	FVD↓	MF↑	LPIPS↓	SSIM↑	PSNR↑	FID↓	LPIPS↓	SSIM↑	PSNR↑	MPJPE↓	Res.
CosHand	58.5	0.591	0.139	0.767	23.2	–	–	–	–	30.1	256×256
InterDyn	38.8	0.680	0.119	0.848	24.9	–	–	–	–	–	256×384
ManiVideo	–	–	0.079	0.913	30.1	–	–	–	–	57.3	–
PAM	31.9	**0.834**	0.093	0.861	26.3	3.81	0.136	0.772	22.6	31.1	480×720
Ours	**16.9**	**0.834**	**0.071**	**0.882**	**27.1**	**3.14**	**0.107**	**0.803**	**23.6**	**21.5**	480×720

high restoration gains by their restored versions, while retaining the original RGB targets elsewhere. The 3D states and PASC remain fixed, so only the visual supervision changes.

Training Datasets. We train the observation model on the DexYCB (Chao et al., 2021) s0 training split, using 6,400 videos of 49 frames at 480×720. Training pairs consist of RGB videos and PASC sequences rendered from the annotated hand–object geometry.

4 Experiments

4.1 Experimental Setup

Datasets. We use DexYCB (Chao et al., 2021) split s0 as the primary benchmark, containing 6,400 training and 1,600 validation videos of 49 frames at 480×720. Approximately 85.4% of evaluated frames contain a visible hand. For interaction-centered evaluation, we compute hand metrics on 200×200 crops centered at the main-hand keypoint centroid and restrict them to frames in which the hand is visible. The motion-associated degradation analysis additionally uses the DexYCB training set, EgoDex (Hoque et al., 2025), and OakInk2 (Zhan et al., 2024).

Evaluation Metrics. We evaluate complementary aspects of video prediction. For appearance fidelity, we report LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004), and PSNR on both full frames and hand crops. We use FVD (Unterthiner et al., 2019) and hand-crop FID (Heusel et al., 2017; Szegedy et al., 2016) to measure video-level and hand-region distribution quality, respectively. Motion fidelity (MF) (Yatim et al., 2024) measures directional consistency between generated and ground-truth point trajectories from a point tracker (Karaev et al., 2024). To assess interaction geometry from the generated observations, we estimate 3D hands from the generated videos (Pavlakos et al., 2024) and report MPJPE against the annotated DexYCB joints, excluding out-of-view joints. FVD uses I3D (Carreira & Zisserman, 2017) features over all 1,600 validation videos.

Training and Implementation. We train with AdamW (Loshchilov & Hutter, 2019) at a learning rate of 10^{-4} and a batch size of 32. Training uses pre-encoded latents of $T = 49$ frames at 480×720, bfloat16 precision, gradient checkpointing, and first-frame conditioning dropout of 0.05. For PAM and our model, we use identical inference settings: 49 frames, 50 sampling steps, and dynamic classifier-free guidance (Ho & Salimans, 2022) of 6.0. We evaluate the 8k-step checkpoint.

4.2 Main Results

Table 1 compares our model with prior HOI video generators, including CosHand (Sudhakar et al., 2024), InterDyn (Akkerman et al., 2025), and ManiVideo (Pang et al., 2025). The upper block reproduces results from PAM (Gao et al., 2026) for literature context but is not directly comparable because of different resolutions and evaluation protocols. The lower block evaluates released PAM and our model under the identical protocol in Sec. 4.1 and supports the controlled claims below.

Under this unified protocol, our observation model improves both global video quality and interaction-centered fidelity relative to PAM. FVD decreases from 31.9 to 16.9, accompanied by improvements in full-frame LPIPS, SSIM, and PSNR. More importantly for HOI prediction, the gains persist when evaluation is restricted to the interaction region: hand-crop FID decreases from 3.81 to 3.14, hand-crop LPIPS from 0.136 to 0.107, and hand-pose MPJPE from 31.1 to 21.5 mm. The reported MF remains unchanged at 0.834. These results indicate that the improvement is not

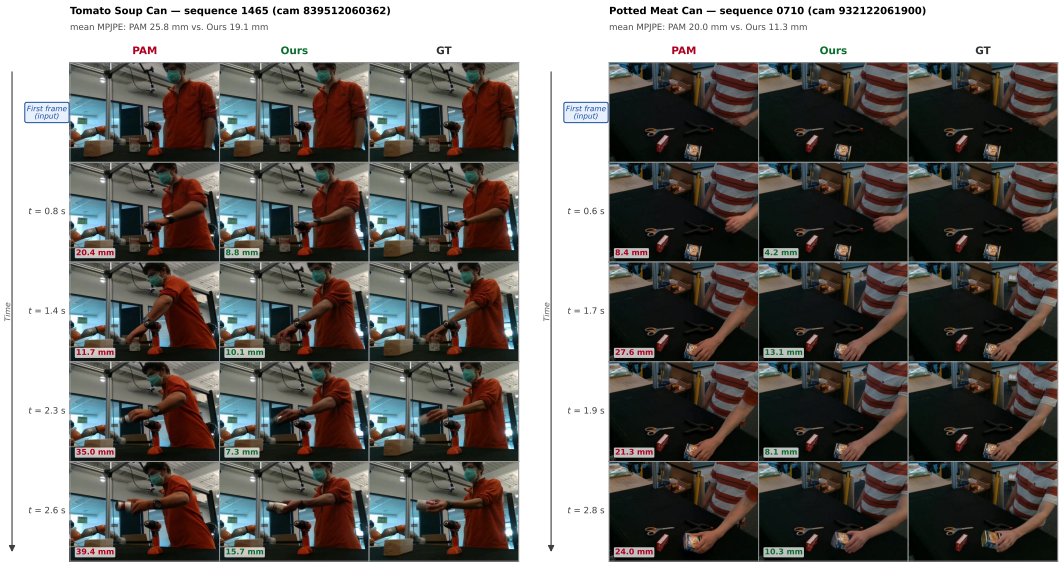

Figure 3: **Qualitative comparison with PAM on DexYCB.** Two validation sequences on which released PAM exhibits large hand-pose errors. For each sequence, columns show PAM, our model, and the ground truth, and rows show the conditioning first frame and four sampled frames (top to bottom). Per-frame hand MPJPE (mm) is annotated on the PAM and Ours tiles. Our model stays close to the ground-truth hand trajectory along the clip, while PAM drifts or loses contact. Further evaluated sequences are shown in Appendix F.

Table 2: **Ablation of geometry–appearance interaction and training targets on DexYCB s0.** One-way models replace bidirectional geometry–appearance interaction with one-way full-field control. Motion-aware (m-aware) variants differ only in supervision targets. All variants are evaluated against the original references. MPJPE is in mm.

	Full frame					Hand crop				
Method	FVD↓	MF↑	LPIPS↓	SSIM↑	PSNR↑	FID↓	LPIPS↓	SSIM↑	PSNR↑	MPJPE↓
Ours	**16.9**	**0.834**	**0.071**	**0.882**	**27.1**	**3.14**	**0.107**	**0.803**	23.6	21.5
One-way	98.2	0.755	0.143	0.799	22.5	3.80	0.163	0.734	21.1	29.0
Ours (m-aware)	17.5	0.831	0.075	0.879	27.0	3.42	0.110	0.794	**23.7**	**21.0**
One-way (m-aware)	93.5	0.758	0.140	0.802	22.6	4.05	0.164	0.734	21.1	28.9

confined to full-frame appearance. The generated observations also better preserve hand-region appearance and the 3D hand geometry recovered from the generated video.

Figure 3 illustrates this difference on validation sequences selected for large PAM hand-pose errors. In these failure cases, PAM can drift from the ground-truth hand trajectory or lose the intended hand–object contact, whereas our predictions remain closer to the reference trajectory across the displayed frames. These examples are intended to visualize the interaction errors reflected by MPJPE, and additional evaluated sequences are provided in Appendix F.

4.3 Experimental Analysis

Ablation on Geometry–Appearance Interaction. Given dense PASC control, we test whether the way geometric control interacts with generation features matters. We replace the proposed bidirectional geometry–appearance interaction with one-way full-field injection, which removes generation-to-control feedback and rendered-support gating while keeping the remaining settings fixed. Motion-aware rows are analyzed below.

Despite receiving the same PASC sequence, the one-way variant degrades substantially across the evaluation metrics (Table 2). FVD increases from 16.9 to 98.2, MF decreases from 0.834 to 0.755, and hand-pose MPJPE increases from 21.5 to 29.0 mm, with the appearance metrics degrading similarly. This shows that providing dense geometric information alone is insufficient: performance depends strongly on how the pixel-aligned control is coupled to the generation stream. Our interaction design exploits the correspondence established by PASC in two ways: generation features

Table 3: **Relationship between hand motion and local observation quality.** Spearman correlation ρ between hand speed and original hand-window quality or restoration gain Δ.

Spearman ρ	DexYCB-val	DexYCB-train	EgoDex	OakInk2
Speed vs. MUSIQ (Ke et al., 2021) (orig.)	-0.165	-0.230	-0.534	-0.458
Speed vs. MANIQA (Yang et al., 2022) (orig.)	-0.246	-0.253	-0.528	-0.504
Speed vs. ΔMUSIQ	$+0.123$	$+0.260$	$+0.357$	$+0.554$
Speed vs. ΔMANIQA	$+0.114$	$+0.246$	$+0.370$	$+0.595$

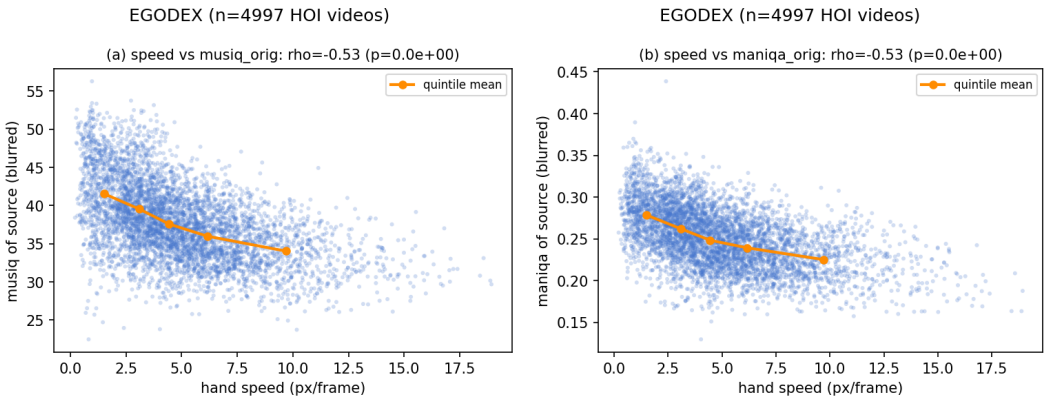

Figure 4: **Hand speed versus source observation quality on EgoDex.** Hand speed (px/frame) against pre-restoration MUSIQ (left) and MANIQA (right). Orange lines connect speed-quintile means, and panel titles report Spearman ρ. Panels for DexYCB and OakInk2 are in Appendix C.

contextualize the geometric control before it is returned to the generation stream, and rendered-support gating confines this exchange to spatial locations supported by the projected geometry. The ablation evaluates these choices jointly, and the same gap persists under motion-aware supervision.

Motion-Associated Supervision Degradation. Even with a geometry-resolved interface, the observation model still learns from RGB targets. We test whether faster hand motion is associated with lower hand-region quality and larger restoration gains. We analyze DexYCB validation (n=1470), DexYCB training (n=5288), EgoDex (n=4997 sampled), and OakInk2 (n=2508). Each clip is restored with the same model (Jin et al., 2026). On 200×200 hand windows, we compute Spearman correlations between hand speed and both original quality and restoration gain Δ.

Table 3 and Figure 4 show a consistent pattern. Hand speed is negatively correlated with original hand-region quality (ρ from -0.165 to -0.534) and positively correlated with restoration gain (ρ from 0.114 to 0.595). Thus, faster hand motion is associated with lower-quality hand-region supervision and with greater potential benefit from restoration across all four datasets. These correlations do not by themselves isolate motion blur as the sole cause of the degradation, but they identify a reproducible motion-associated weakness in the supervision of interaction-critical regions.

Motion-Aware Restored Supervision. Motivated by this analysis, we replace only high-gain hand windows with their restored targets during training, leaving the remaining RGB supervision, PASC sequences, and architecture unchanged. As shown in Table 2, motion-aware supervision reduces hand-pose MPJPE from 21.5 to 21.0 mm. The improvement is modest, and the conventional appearance metrics do not improve correspondingly: FVD changes from 16.9 to 17.5, while the reconstruction metrics remain similar or slightly worse. This is partly expected, as the restored targets differ from the evaluation references (Appendix E). We thus read this result narrowly: selective restoration yields a small hand-pose gain but does not improve overall video fidelity.

Fine-Grained Control Probe. We next ask whether generated pixels measurably respond to fine-grained control details. We degrade every condition slot—depth, mask, and keypoints for PAM, the PASC render for ours—by $8\times$ and $16\times$ mean-pooled downsampling and upsampling, which removes condition high frequencies by two to three orders of magnitude in Laplacian variance while preserving coarse position, silhouette, and mean color. All levels share identical sampling noise and inference settings, and we evaluate 128 paired validation samples. Figure 6 in Appendix B visualizes the degraded conditions, which are best compared visually. Table 4 reports the results.

Table 4: **Condition-degradation probe on DexYCB.** Generated sharpness is the Laplacian variance averaged over the 49 output frames. We report its relative change with respect to the same sample under the original condition (mean ± std over 128 paired validation samples) and the number of samples whose sharpness decreases.

Model	ΔLapVar ds8 (%)	ΔLapVar ds16 (%)	Neg. samples
PAM (multi-slot one-way)	-1.9 ± 4.1	-1.4 ± 4.7	94/128, 95/128
Ours	**-14.2** $\pm$ 6.7	**-19.2** $\pm$ 10.6	128/128, 127/128

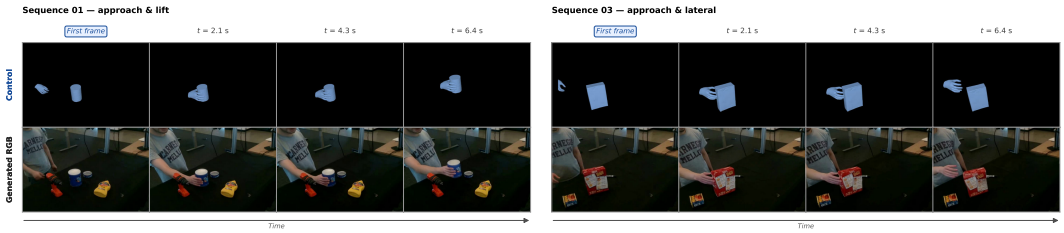

Figure 5: **Simulator-to-observation transfer.** Two simulator trajectories with different interaction outcomes: in Sequence 01 the hand grasps and lifts the can, whereas in Sequence 03 the hand approaches the box laterally, makes contact, and withdraws without lifting it. Top: rendered control. Bottom: generated RGB video.

PAM responds weakly, with generated sharpness changing by -1.9% and -1.4% under $8\times$ and $16\times$ degradation, respectively, with substantial sample-level variation. In contrast, our model decreases by -14.2% and -19.2%, with decreases observed for 128/128 and 127/128 samples. Our outputs are thus markedly more sensitive to fine-grained condition content. Because the rendered-support mask and sampling factors are fixed across degradation levels, this response cannot be attributed to changes in gating or sampling. The probe does not identify which internal features are used, but shows behaviorally that the model responds to condition information beyond coarse support and silhouette. This supports resolving pixel-aligned geometry before generation.

Simulator-to-Observation Transfer. Finally, we test whether the interface remains usable when future 3D states come from an external interaction simulator F_{sim} instead of the dataset. The simulated states are rasterized into PASC as before, and G_{obs} is used without retraining. We consider two DexYCB scenes whose simulated trajectories lead to different outcomes.

As shown in Fig. 5, the observation model produces videos that qualitatively follow the hand–object configurations specified by the simulator trajectories, both when the object is lifted and when it is only touched. This experiment does not evaluate the accuracy of F_{sim}, but shows that the source of 3D states can be changed without modifying the representation consumed by G_{obs}.

5 Conclusion

We have presented a factorized formulation for interactive video world modeling that separates explicit 3D hand–object dynamics from visual observation synthesis. Through Pixel-Aligned Surface Conditioning, future 3D states are rendered into the target view before generation, and bidirectional geometry–appearance interaction couples this geometry with appearance features in the observation model. Under a unified DexYCB protocol, our model has reduced FVD from 31.9 to 16.9 and hand-pose MPJPE from 31.1 to 21.5 mm relative to PAM, and our analyses have shown that performance depends strongly on how geometric control is coupled to generation. We have also identified a consistent association between fast hand motion and degraded hand-region supervision, and have shown that simulator-generated 3D states can drive the unchanged observation model through the same interface. These results suggest that explicit 3D state-to-observation interfaces provide a practical basis for separating interaction dynamics from visual synthesis in interactive world models.
Limitations. Our quantitative evaluation uses dataset-provided 3D states and thus measures the observation model rather than the full pipeline, while simulator transfer is assessed only qualitatively. The model can also produce locally plausible but interaction-inconsistent outcomes for atypical grasps or occluded objects, which hand-only metrics such as MPJPE do not capture (Appendix G).

AI USE STATEMENT

We used generative AI tools for manuscript writing assistance and for auxiliary tooling scripts, but not for research ideation, experiment design, or producing any reported result. All AI-assisted output was reviewed by authors, who take full responsibility for the final content of the wor

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, the main paper and Appendix A provide detailed descriptions of our methodology and implementation. We will release the complete source code and model weights.

REFERENCES

Rick Akkerman, Haiwen Feng, Michael J. Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. InterDyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12467–12479, 2025.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15791–15801, 2025. doi: 10.1109/CVPR52734.2025.01472.

Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992. doi: 10.1109/34.121791.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733, 2017. doi: 10.1109/CVPR.2017.502.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9044–9053, 2021.

Mingjin Chen, Junhao Chen, Zhaoxin Fan, Yujian Lee, Zichen Dang, Lili Wang, Yawen Cui, Lap-Pui Chau, and Yi Wang. HVG-3D: Bridging real and simulation domains for 3d-conditional hand-object interaction video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15986–15997, 2026.

Yingying Fan, Quanwei Yang, Kaisiyuan Wang, Hang Zhou, Yingying Li, Haocheng Feng, Errui Ding, Yu Wu, and Jingdong Wang. Re-HOLD: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 17550–17560, 2025.

Mingju Gao, Kaisen Yang, Huan-ang Gao, Bohan Li, Ao Ding, Wenyi Li, Yangcheng Yu, Jinkun Liu, Shaocong Xu, Yike Niu, Haohan Chi, Hao Chen, Hao Tang, Yu Zhang, Li Yi, and Hao Zhao. PAM: A pose-appearance-motion engine for sim-to-real HOI video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15954–15965, 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13416–13426, 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 (NeurIPS)*, volume 30, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

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

Renbiao Jin, Mingxin Yang, Yutian Chen, Junhao Zhuang, Xin Cai, Mulin Yu, Linning Xu, Wenxian Yu, Danping Zou, Shi Guo, and Tianfan Xue. RealVDeblur: One-step diffusion for generalizable real-world video deblurring. *arXiv preprint arXiv:2607.20628*, 2026.

Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, October 2021.

Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29663–29673, 2026.

Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10138–10148, 2021.

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.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 38–55, 2024. doi: 10.1007/978-3-031-72970-6_3.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019.

Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3883–3891, 2017.

Youxin Pang, Ruizhi Shao, Jiajun Zhang, Hanzhang Tu, Yun Liu, Boyao Zhou, Hongwen Zhang, and Yebin Liu. ManiVideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 12209–12219, 2025.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Bohao Peng, Jian Wang, Yuechen Zhang, Wenbo Li, Ming-Chang Yang, and Jiaya Jia. ControlNeXt: Powerful and efficient control for image and video generation. *arXiv preprint arXiv:2408.06070*, 2024.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6):245:1–245:17, 2017. doi: 10.1145/3130800.3130883.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022.

Shuochen Su, Mauricio Delbracio, Jue Wang, Guillermo Sapiro, Wolfgang Heidrich, and Oliver Wang. Deep video deblurring for hand-held cameras. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1279–1288, 2017.

Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

Zhihao Sun, Zhiying Du, Xitong Yang, and Zuxuan Wu. HandWorld: Hand-centric unified video action generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15976–15985, 2026.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, 2016.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012. doi: 10.1109/IROS.2012.6386109.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric and challenges. *arXiv preprint arXiv:1812.01717*, 2019.

Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2World: Autoregressive egocentric interaction generation via free-space hand gestures. *arXiv preprint arXiv:2602.09600*, 2026.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004.

Xinyue Wei, Minghua Liu, Zhan Ling, and Hao Su. Approximate convex decomposition for 3d meshes with collision-aware concavity and tree search. *ACM Transactions on Graphics*, 41(4), 2022. doi: 10.1145/3528223.3530103.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3D generation. *arXiv preprint arXiv:2512.14692*, 2025.

Haodong Yan, Hang Yu, Zhide Zhong, Weilin Yuan, Xin Gong, Zehang Luo, Chengxi Heyu, Junfeng Li, Wenxuan Song, Shunbo Zhou, and Haoang Li. Open-world hand-object interaction video generation based on structure and contact-aware representation. *arXiv preprint arXiv:2512.01677*, 2025.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5325–5335, 2022.

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

Danah Yatim, Rafail Fridman, Omer Bar-Tal, Yoni Kasten, and Tali Dekel. Space-time diffusion features for zero-shot text-driven motion transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8466–8476, 2024.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. OAKINK2: A dataset of bimanual hands-object manipulation in complex task completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024.

Appendix

A Experimental and Implementation Details

Training. Our model is trained on 8 NVIDIA A800 GPUs with a total batch size of 32, using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 10^{-4}. Following Sec. 4.1, training operates on pre-encoded latents of $T = 49$ frames at 480×720, in bfloat16 precision with gradient checkpointing and first-frame conditioning dropout of 0.05. All quantitative results use the 8k-step checkpoint unless stated otherwise.

Inference. Both PAM and our model use identical inference settings: 49 generated frames per sample, 50 sampling steps, and dynamic classifier-free guidance (Ho & Salimans, 2022) of 6.0, with the first frame given as conditioning.

Evaluation protocol. Unless otherwise specified, all metrics are computed on the 1,600 validation videos of DexYCB s0. FVD uses I3D (Carreira & Zisserman, 2017) features, motion fidelity uses trajectories from CoTracker3 (Karaev et al., 2024), and 3D hands are estimated from generated videos with HaMeR (Pavlakos et al., 2024) against annotated DexYCB joints, excluding out-of-view joints. Hand-level metrics are computed on 200×200 crops centered at the main-hand keypoint centroid and restricted to frames with a visible hand. The condition-degradation probe of Sec. 4.3 uses 128 paired validation samples with identical sampling noise and inference settings across degradation levels.

Motion-associated degradation analysis. Hand speed and hand-window quality are computed on 200×200 hand windows with MUSIQ (Ke et al., 2021) and MANIQA (Yang et al., 2022). The analysis covers DexYCB validation (n=1470), DexYCB training (n=5288), EgoDex (n=4997 sampled), and OakInk2 (n=2508). Clips whose hand segmentation covers fewer than 1,000 pixels, or for which HaMeR (Pavlakos et al., 2024) fails to detect the hand, are treated as containing no usable hand region and are excluded from this analysis; the counts above are reported after this filtering. Each clip is restored with the same video deblurring model (Jin et al., 2026), and restored targets are used only in the motion-aware variants of Table 2.

B Visualization of the Condition-Degradation Probe

Figure 6 visualizes the degraded conditions used in the fine-grained control probe (Section 4.3). The top three rows show the condition slots of PAM—depth, segmentation mask, and hand keypoints—and the bottom row shows our PASC render. Columns correspond to the original condition and its $8\times$ and $16\times$ mean-pooled downsampled–upsampled versions, with panel titles reporting the condition-side Laplacian variance (LapVar). The degradation preserves coarse position, silhouette, and mean color, while the LapVar drops by two to three orders of magnitude and saturates near the floor already at $8\times$, so the levels are best compared visually. Red boxes mark zoomed regions.

C Additional Panels for the Motion-Associated Degradation Analysis

Figure 7 complements the motion-associated degradation analysis in the main text by plotting hand speed against the source (pre-restoration) MUSIQ (top row) and MANIQA (bottom row) of the hand window on DexYCB-val, DexYCB-train, and OakInk2. The negative association is present in all three datasets shown here. Spearman correlations for all four datasets are reported in Table 3 of the main text.

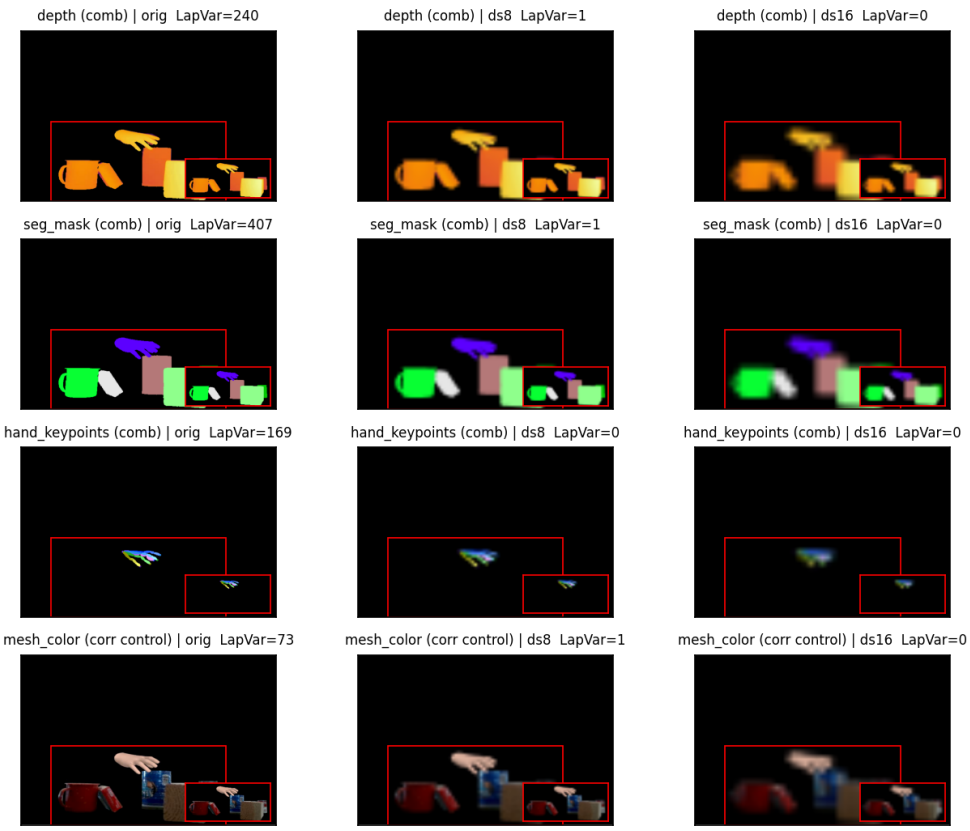

Figure 6: **Condition degradation for the fine-grained control probe.** The top three rows show the condition slots of PAM (depth, segmentation mask, hand keypoints) and the bottom row shows our PASC render on a DexYCB validation sample. Columns show the original and the $8\times$ and $16\times$ mean-pooled downsampled–upsampled versions, and panel titles report the condition-side Laplacian variance (LapVar). Red boxes mark zoomed regions.

D Qualitative Examples of Restoration Gains on Blur-Affected Clips

Figures 8–10 visualize the restoration analysis of Section 4.3 on individual clips from DexYCB (train), EgoDex, and OakInk2. We show clips among those with the largest hand-window restoration gains when ranked by ΔMUSIQ or ΔMANIQA (the same hand-window scores as in Table 3). The left column shows original frames and the right column shows restored frames. The row label reports the clip-level hand-window MUSIQ and MANIQA before and after restoration. These clips are raw dataset observations and their restored versions, not model outputs. In these selected examples, restoration primarily changes the hand and contact regions, and the displayed frames with larger quality-score gains also exhibit visible motion-associated degradation in the original observations. These examples illustrate the degradation pattern quantified in the main text.

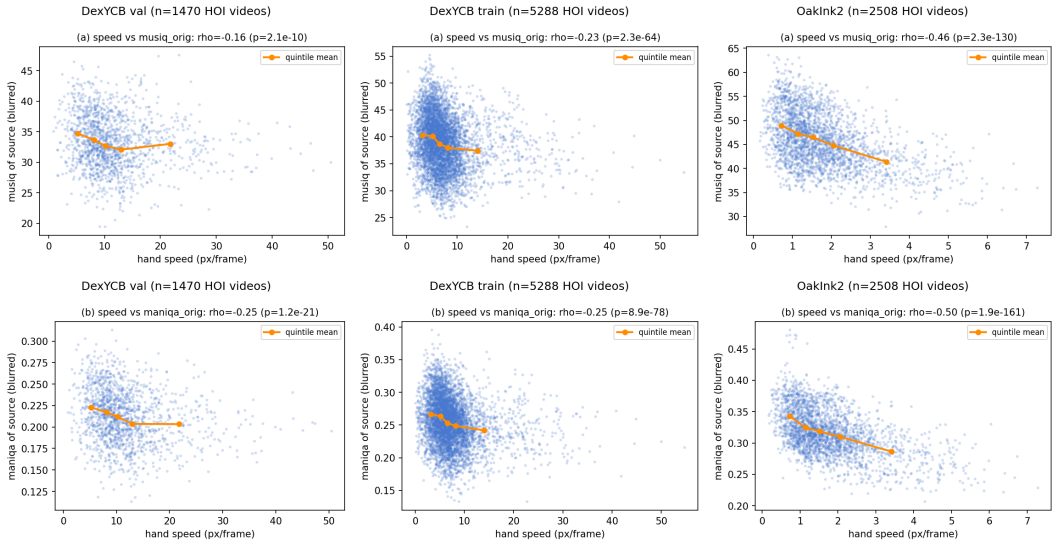

Figure 7: **Hand speed versus source observation quality on DexYCB and OakInk2.** Hand speed (px/frame) against pre-restoration MUSIQ (top row) and MANIQA (bottom row). Orange lines connect speed-quintile means, and panel titles report Spearman ρ.

E Original and Restored Ground-Truth Mismatch

To quantify the evaluation offset introduced by restored supervision itself, we compare original DexYCB ground truth and restored ground truth directly with the same evaluation code, without any generated video. Tables 5 and 6 quantify the nonzero discrepancy between original and restored targets under the same evaluation pipeline. Consequently, a model trained with selectively restored targets is optimized against supervision that differs from the original validation references. We therefore interpret small changes in conventional reference-based metrics together with the hand-pose result.

Table 5: **Cross-reference evaluation between original and restored ground truth: video-level metrics.** The two rows use the original and restored DexYCB ground-truth videos as generated/reference inputs to the same evaluation pipeline. No generated model output is involved.

Generated / Reference	LPIPS↓	SSIM↑	PSNR↑	MF↑	FVD↓
Original / Restored	0.0111	0.949	38.2	0.906	9.33
Restored / Original	0.0111	0.949	38.2	0.906	9.33

Table 6: **Cross-reference evaluation between original and restored ground truth: hand-level metrics.** Setup follows Table 5.

Generated / Reference	H-LPIPS↓	H-SSIM↑	H-PSNR↑	H-FID↓
Original / Restored	0.0332	0.867	32.1	2.40
Restored / Original	0.0332	0.867	32.1	2.40

F Additional Qualitative Comparisons with PAM

Figures 11–14 extend the qualitative comparison of Figure 3 in the main text to 12 additional evaluated DexYCB validation sequences. The layout is identical. Rows show PAM, our model, and the ground truth, columns show the conditioning first frame and four sampled frames, and per-frame hand MPJPE (mm) is annotated on the PAM and Ours tiles.

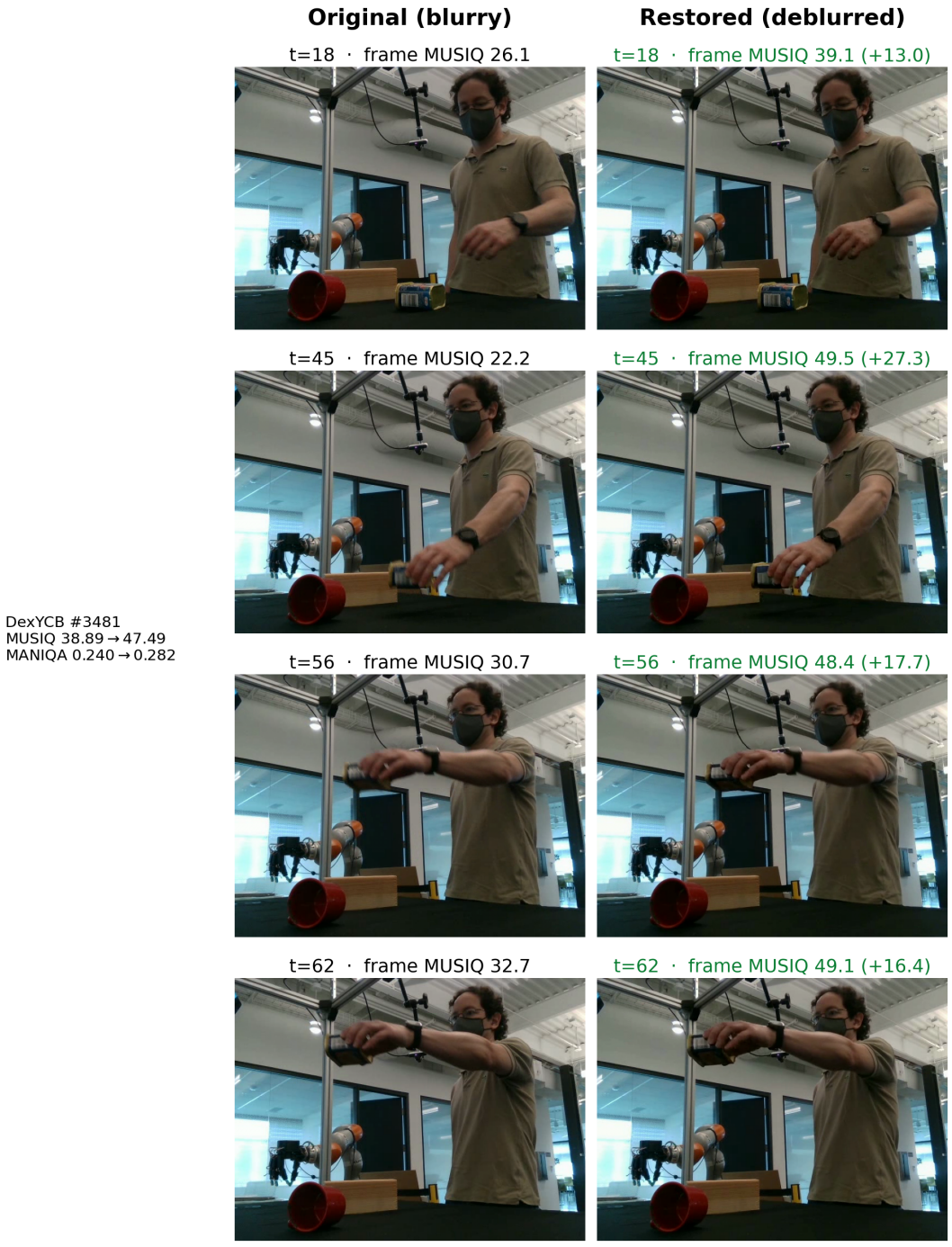

Figure 8: **Restoration examples on DexYCB (train).** Four frames of clip #3481 at different stages of the clip. Left: original frames. Right: restored frames. The row label reports clip-level hand-window MUSIQ/MANIQA before and after restoration, and tile titles report the frame index and the per-frame hand-window MUSIQ with the change in parentheses.

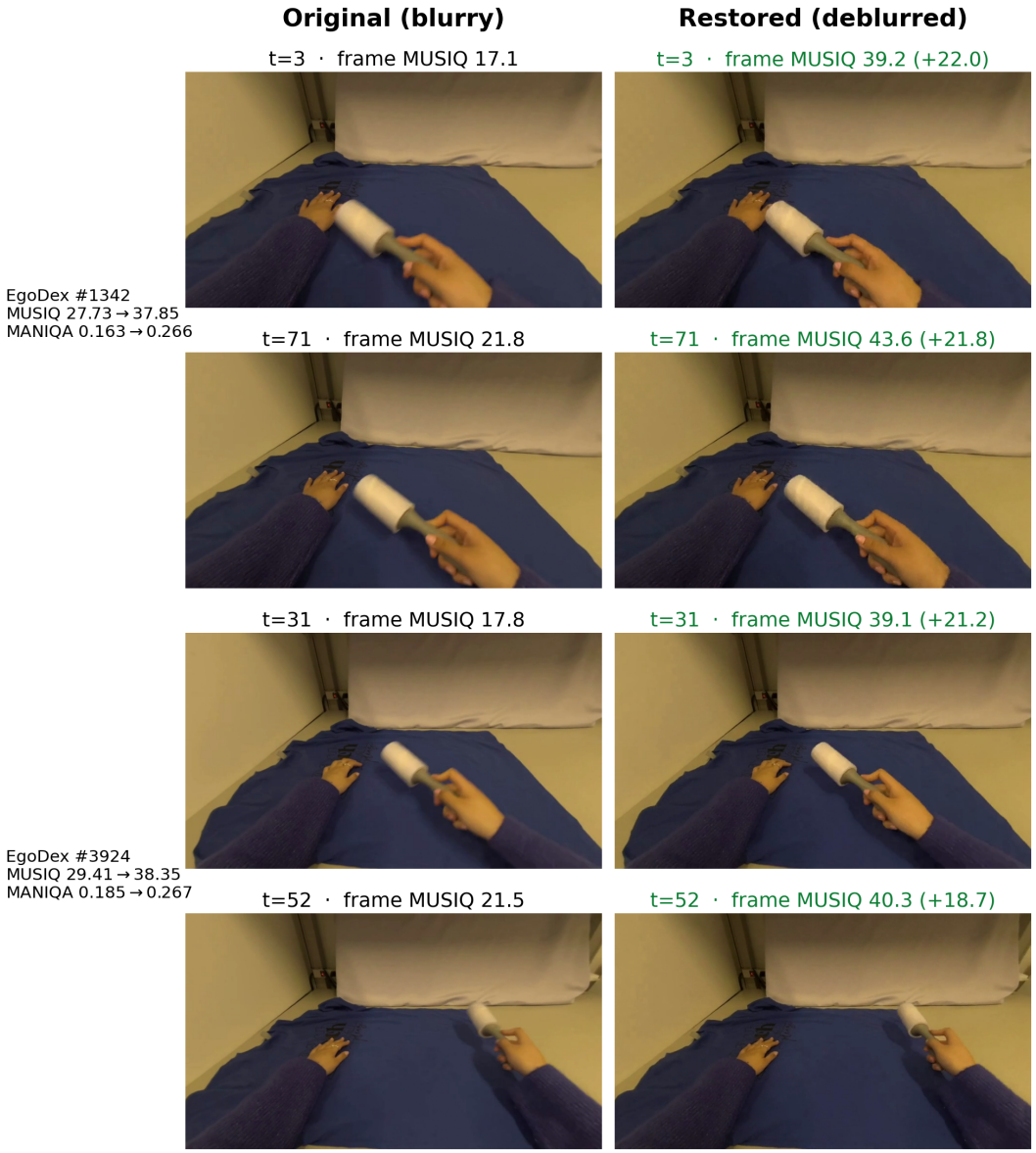

Figure 9: **Restoration examples on EgoDex.** Left: original frames. Right: restored frames. Tile titles report the frame index and the per-frame hand-window MUSIQ with the change in parentheses, and clip-level MUSIQ/MANIQA are reported in the row labels.

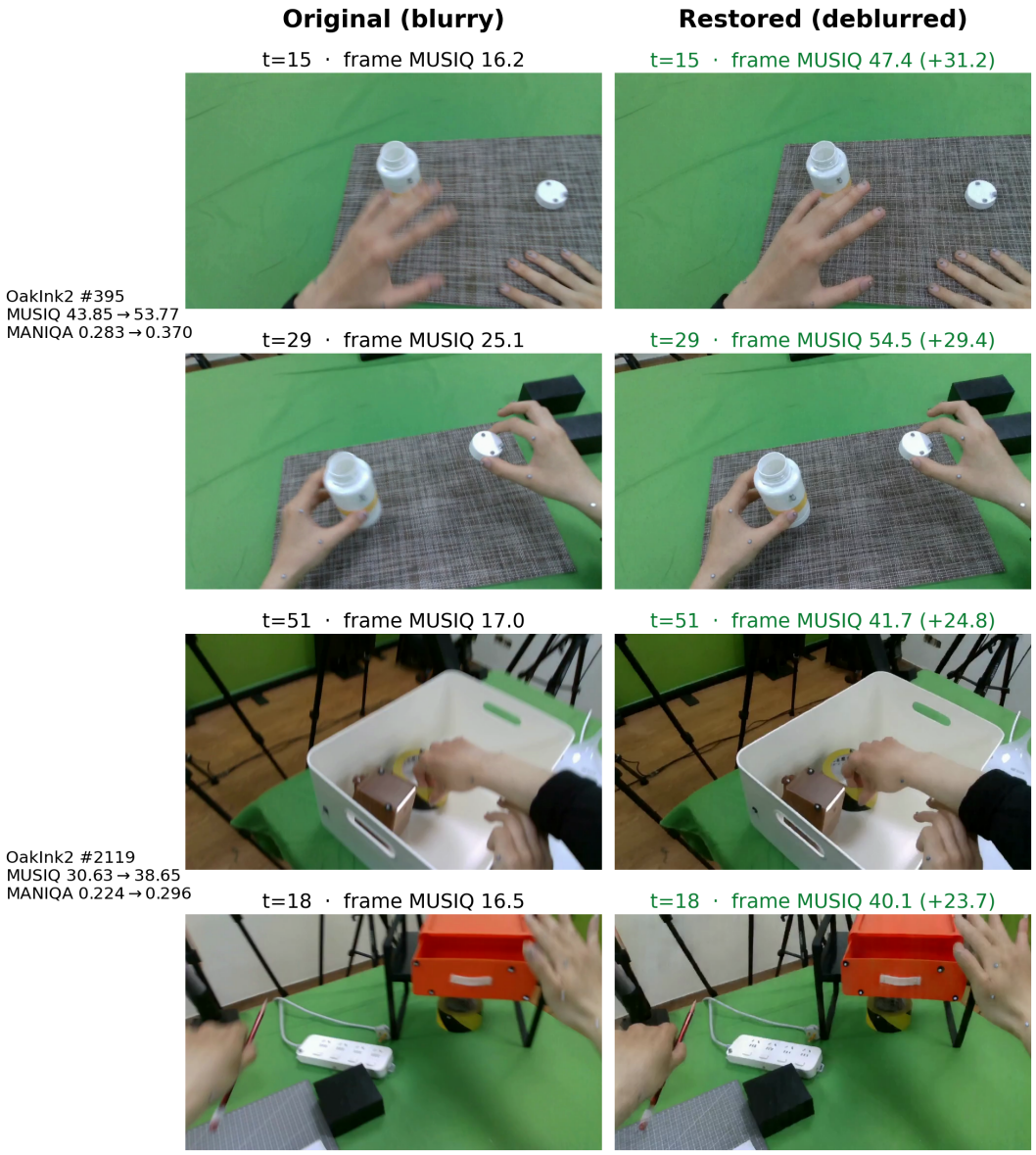

Figure 10: **Restoration examples on OakInk2.** Selection, layout, and annotations follow Figure 9.

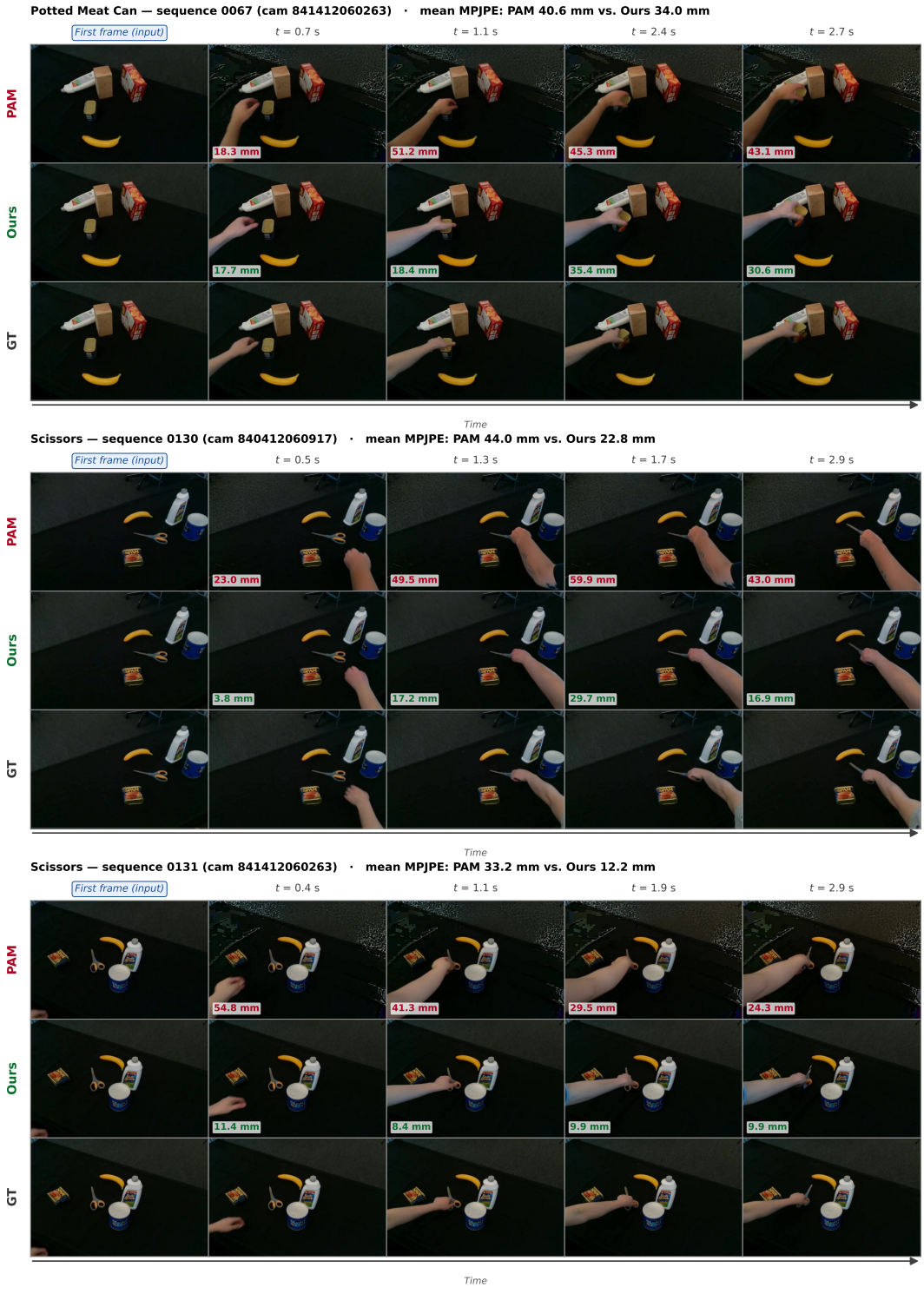

Figure 11: **Additional qualitative comparisons with PAM (1/4).** Sequences 0067 (potted meat can), 0130 (scissors), and 0139 (large marker). Layout and annotations follow Figure 3.



Figure 12: **Additional qualitative comparisons with PAM (2/4).** Sequences 0189 (tomato soup can), 0303 (large marker), and 0352 (mustard bottle). Layout and annotations follow Figure 3.

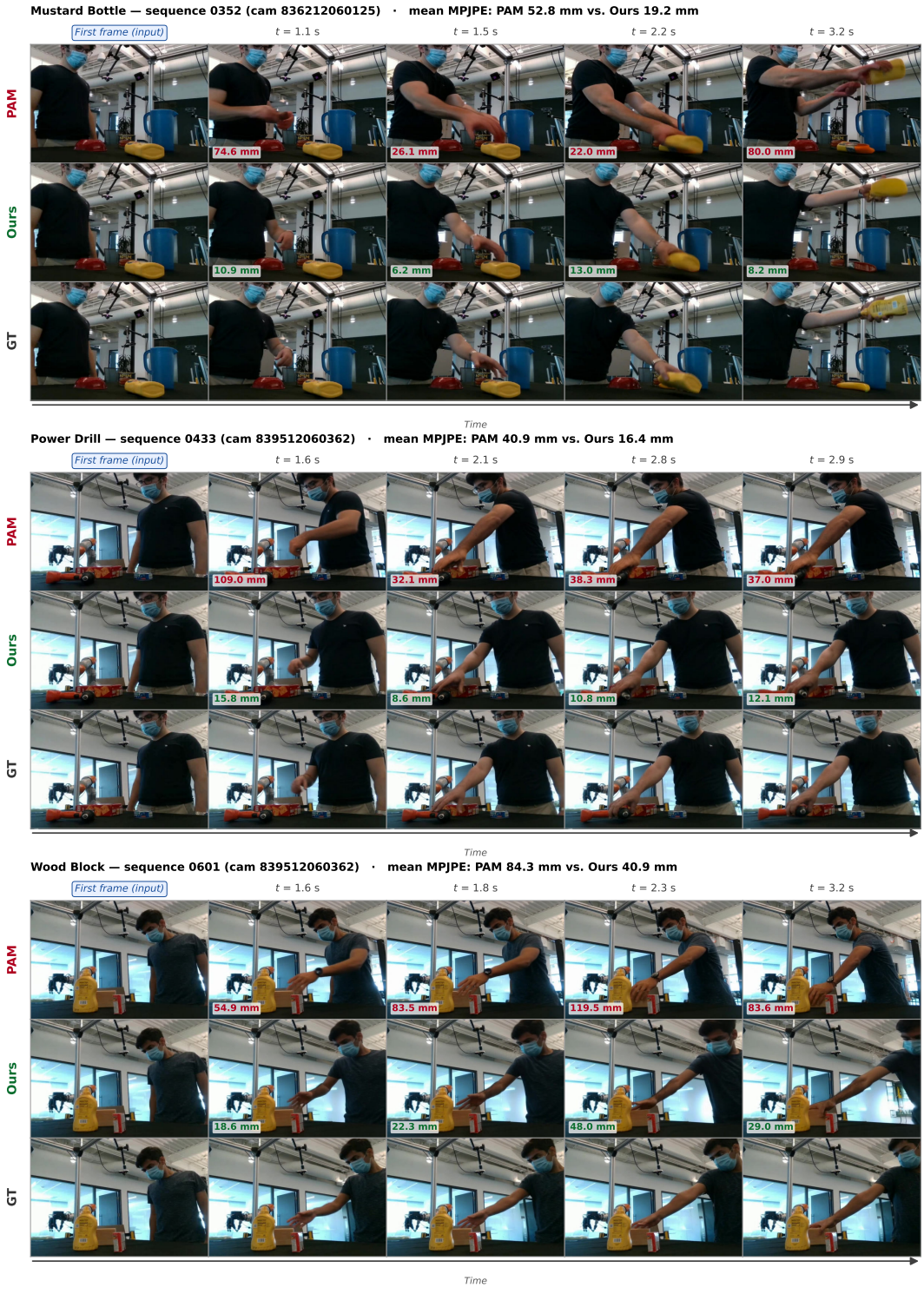

Figure 13: **Additional qualitative comparisons with PAM (3/4).** Sequences 0433 (power drill), 0601 (wood block), and 0720 (pitcher base). Layout and annotations follow Figure 3.

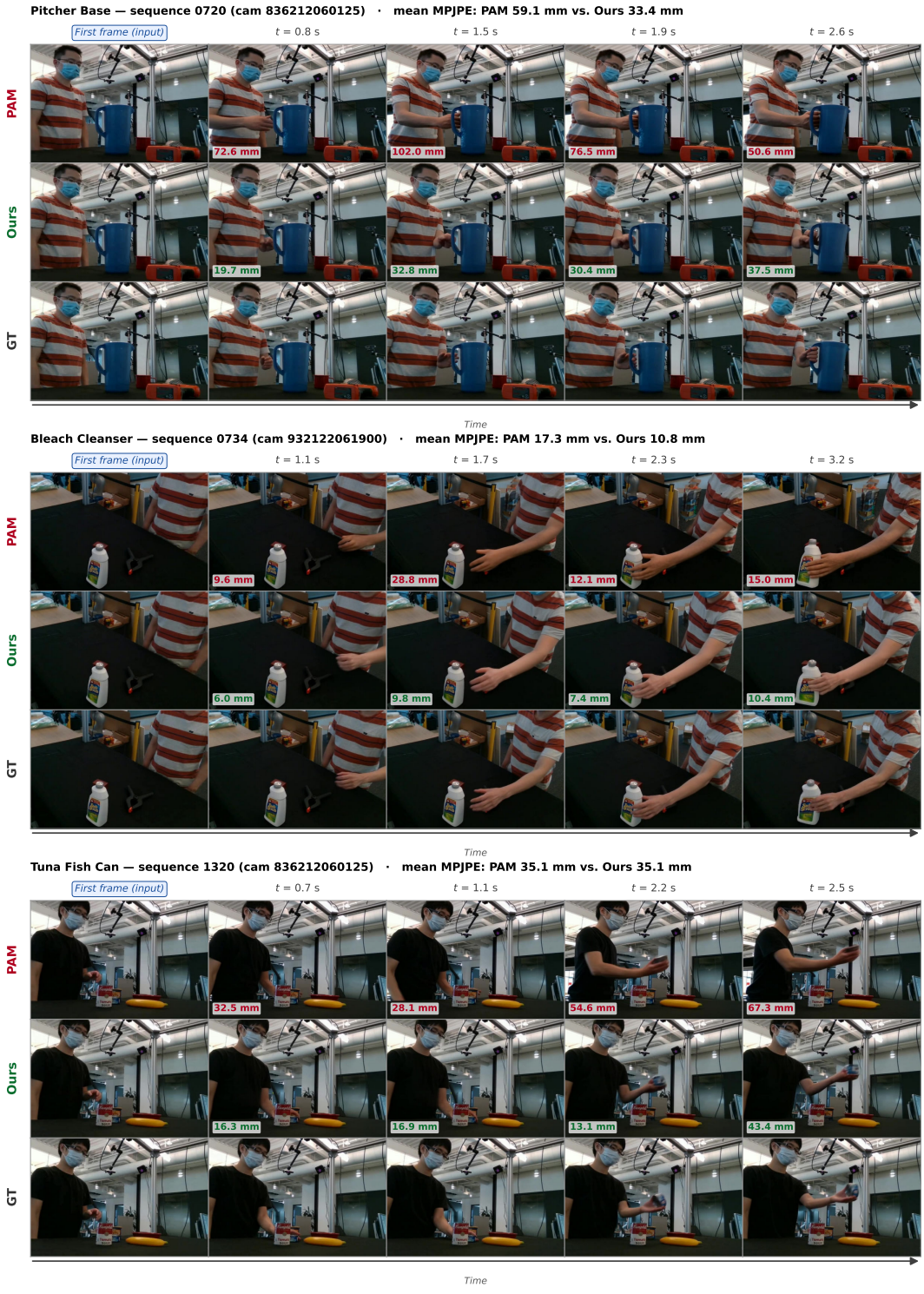

Figure 14: **Additional qualitative comparisons with PAM (4/4).** Sequences 0734 (bleach cleanser), 1320 (tuna fish can), and 1359 (banana). Layout and annotations follow Figure 3.

G FAILURE CASES

Figures 15 and 16 visualize two representative failure cases in which our model does not reproduce the demonstrated interaction. The layout follows Figure 3. Rows show PAM, our model, and the ground truth, columns show the conditioning first frame and four sampled frames, and the bottom three rows magnify the regions marked by the red boxes. Per-frame hand MPJPE (mm) is annotated on the PAM and Ours tiles when the frame falls within the evaluation window.

Wood block (Seq. 0602, Figure 15). The ground truth contains an inverted grasp: the hand passes over the far side of the block and curls over its top edge with the palm facing downward. Our generated video instead depicts a lateral approach from the near side. The hand geometry diverges from the reference as the interaction progresses (per-frame MPJPE reaches 147.3 mm at frame 26, versus 53.0 mm for PAM), and in the second half the generated hand drifts toward the neighboring mustard bottle while the block is displaced without being grasped. This example shows that the observation model can produce a locally plausible but interaction-inconsistent outcome for an atypical hand–object configuration.

Scissors (Seq. 1088, Figure 16). In the ground truth, the hand reaches *behind* the mug, which occludes the scissors, and subsequently pulls the scissors into view. Our generated video instead keeps the hand on the near side of the mug, and the scissors never emerge. The sequence therefore exposes a failure involving occluded object state and hand–object interaction, rather than hand pose alone.

Both cases share a broader pattern: local appearance and hand motion can remain plausible while the resulting hand–object interaction is inconsistent with the reference. They also highlight complementary limitations of the current evaluation: hand-pose error captures substantial geometric drift in the first case but does not fully characterize the occluded object-state failure in the second.



Figure 15: **Failure case: inverted grasp (Seq. 0602, wood block).** Rows show PAM, our model, and the ground truth, and columns show the conditioning first frame and four sampled frames. The bottom three rows magnify the regions marked by red boxes, and per-frame hand MPJPE (mm) is annotated on the PAM and Ours tiles when the frame falls within the evaluation window. The ground truth performs an inverted grasp over the top of the block, while our model approaches from the near side, fails to grasp, and displaces the block.



Figure 16: **Failure case: retrieval behind an occluder (Seq. 1088, scissors).** Conventions follow Figure 15. The ground truth reaches behind the mug and pulls the scissors into view, whereas our model keeps the hand in front of the mug and the scissors never emerge. PAM behaves similarly.