DYNAMIC PROJECTIVE GUIDANCE FOR UNIFIED VIDEO AND 3D HAND MOTION GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating paired egocentric video and 3D hand motion can provide supervision for dexterous robot learning. Recent video–action models jointly generate both modalities, avoiding error propagation from sequential generation. However, they learn cross-modal correspondence *implicitly*, so the generated video and hand motion are often individually plausible yet spatio-temporally inconsistent. For example, given the same instruction, the 2D video hand trajectory may not align with the reprojected 3D hand motion trajectory on the image plane. To address this limitation, we introduce **Uni-HaVi**, a **uni**fied model for 3D **ha**nd motion and **vi**deo generation. Compared with 3D hand motion generation, which gradually stabilizes in the later flow-matching steps, we observe that the hand trajectory in the generated video is largely determined within the first few steps. Exploiting this asymmetry, we propose **Dynamic Projective Guidance (DPG)**, which uses a video hand locator and a learned adapter to progressively guide 3D hand motion generation toward the video hand trajectory throughout the flow-matching process. As a result, DPG enforces stronger spatio-temporal consistency between the video and 3D hand motion, while allowing their appearance and articulation to be independently refined during generative process. On EgoDex, DPG reduces wrist alignment error from 42.3 to 19.0 px and improves whole-hand IoU from 0.245 to 0.313, as measured by wrist and hand evaluators. By treating either modality as given or generated, Uni-HaVi supports video-to-hand-motion, hand-motion-to-video, and joint generation with a single model, and achieves strong video–hand spatio-temporal alignment and improved 3D hand-motion quality while maintaining competitive video quality.

1 INTRODUCTION

Generating egocentric human demonstrations with paired video and 3D hand motion provides cost-efficient supervision for dexterous robot learning (Paliwal et al., 2026). However, reliable paired supervision requires spatio-temporal consistency between the two modalities. For example, a hand may appear to grasp a cup in the video while its 3D trajectory places it beside the cup, producing conflicting visual and motion supervision.

To address the above limitation, a common strategy is to generate one modality and infer the other. For example, one can generate a video and recover 3D hand motion from it (Duran et al., 2024; Zhang et al., 2025a; Yu et al., 2025; Zhang et al., 2025b; Ismayilzada et al., 2026), or specify a hand trajectory and generate a video from it (Li et al., 2026; Akkerman et al., 2025; Chen et al., 2026a; Kim et al., 2026). When both modalities need to be generated and the two stages are chained, however, errors from the first stage can carry over to the second. This motivates a unified model that generates video and hand motion together, and can also generate one modality from the other. Recent world action models take a step toward this goal by jointly generating video and robot actions (Li et al., 2025a; Bi et al., 2026; Ye et al., 2026a), with recent work extending this idea to egocentric video and bimanual hand trajectories (Sun et al., 2026; Wang et al., 2026a). However, joint generation alone does not ensure spatio-temporal consistency: the generated 3D hand motion may not align with the hand seen in the video. These models have to learn spatio-temporal consistency implicitly from paired video and hand-motion data, which remain scarce. As a result, the generated video and hand motion may each look plausible but spatio-temporally inconsistent. What

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.

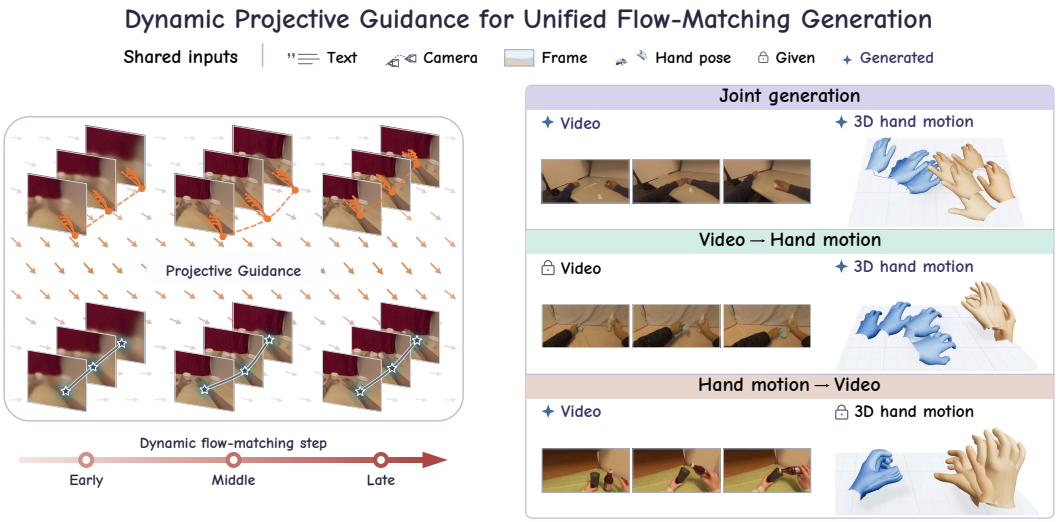


Figure 1: **Dynamic Projective Guidance.** DPG progressively guides 3D hand motion toward the video hand trajectory, improving spatio-temporal consistency. **Unified video and 3D hand-motion generation.** One model supports joint generation and both conditional directions.

remains missing is an explicit geometric constraint that aligns the generated 3D hand with its 2D visual counterpart, e.g., the wrist location, during flow-matching.

Ideally, such a geometric constraint should be self-aware for the model, rather than from an external instruction. Interestingly, we observe that the 2D video hand trajectory is determined by video flow matching within the first few flow-matching steps, leaving the rest steps to refine the hand appearance, whereas the trajectory of 3D hand motion stabilizes only in the last few steps. The video therefore can better specify where the hand location should be in the early stage of generative process, while the 3D hand motion is still uncertain, and video hand location can guide the 3D hand motion generation during flow-matching.

Based on this observation, we introduce **Dynamic Projective Guidance (DPG)** for **Uni-HaVi**, a **uni**fied diffusion model for 3D **ha**nd motion and **vi**deo generation. Our proposed model denoises video latents and hand motion with a shared Transformer, but a separate noise level for each modality. To enforce spatio-temporal consistency, we represent hand motion as explicit 3D articulations, allowing the current 3D hand location to be projected onto the image and compared with its visual counterpart for correction, progressively at each flow-matching step. Specifically, we first locate 2D hand position in video, e.g., hand wrist. DPG then calculates the discrepancy between 2D hand position and that from the 3D hand projection, and uses it as the guidance to correct the 3D hand motion. Used alone, however, this projective guidance yields limited improvement because it does not maintain alignment through subsequent flow-matching. We therefore introduce a lightweight adapter that conditions the noise level to guide the correction of the above discrepancy. The resulting hand location corrections guide the evolving 3D hand trajectory, while the shared Transformer blocks continues to refine finger articulation to match the video. As a result, Uni-HaVi supports video-to-hand-motion generation, hand-motion-to-video generation, and joint generation with a single model while maintaining spatio-temporal consistency (Figure 1).

To evaluate cross-modal consistency independently of our Locator, we use RGB-based measurements at two spatial scales: a DINOv2-based wrist evaluator for wrist-level alignment and SAM 2 segmentation for whole-hand overlap. On EgoDex, Uni-HaVi with DPG reduces wrist alignment error from 42.3 to 19.0 px, and improves Hand IoU from 0.245 to 0.313 compared with the unguided model. Beyond alignment, the same model outperforms specialized hand-motion estimators in trajectory-level comparisons and generates hand-conditioned video with quality comparable to specialized video generators. Ablations show that learned guidance during late flow-matching recovers most of the alignment gain, whereas direct analytic correction provides little improvement.

Our contributions are summarized as follows:

- **Dynamic Projective Guidance.** In order to improve the spatio-temporal hand alignment between video and 3D and motion, we propose DPG to combine geometrical constraint with learned adapter to guide the generation process.
- **A unified model with flexible generation capability.** We develop a unified video–hand flow-matching model that supports video-to-hand-motion, hand-motion-to-video, and joint generation with a single model.
- **Improved Cross-Modal Alignment and Generation Quality.** Uni-HaVi improves 3D hand-motion quality via better spatio-temporal both hand and wrist alignment, while maintaining comparable video quality.

2 Related Work

Hand-hand-motion-conditioned interaction video generation. Hand-centric video synthesis has improved the visual fidelity and task coverage of generated interactions (Pang et al., 2025; Zhao et al., 2025). More directly, hand-motion-conditioned methods control video diffusion with hand masks, trajectories, poses, or structured 3D interaction signals (Li et al., 2026; Akkerman et al., 2025; Chen et al., 2026a; Kim et al., 2026). These controls improve the fidelity of generated interactions, but the structured hand signal remains an input rather than a jointly generated output. Consequently, obtaining a paired video and hand trajectory still requires a separate hand-motion source or estimator.

Hand-motion recovery from video. Video-based hand reconstruction estimates temporally coherent articulated hand motion by combining image evidence with learned hand priors, temporal models, or physical constraints (Duran et al., 2024; Zhang et al., 2025a; Yu et al., 2025; Zhang et al., 2025b; Ismayilzada et al., 2026). These methods provide structured 3D trajectories from an observed video and are therefore complementary to hand-motion-conditioned video generation. However, the visual sequence is treated as a fixed observation rather than a generative variable, so consistency cannot be optimized jointly when both modalities must be synthesized.

Joint multimodal generation. General multimodal diffusion models use shared generative backbones and modality-dependent noise levels to support joint and conditional distributions with one set of parameters (Bao et al., 2023; Kim et al., 2024; Li et al., 2025b). Video–action models extend this principle to visual dynamics and structured actions (Li et al., 2025a; Bi et al., 2026; Sun et al., 2026). These models learn cross-modal correspondence implicitly, without an explicit geometric constraint between dense visual locations and compact 3D hand states. Uni-HaVi retains the unified interface while introducing dynamic projective guidance for this missing spatio-temporal constraint.

Geometric guidance for diffusion and motion generation. Guided motion diffusion incorporates differentiable objectives, joint constraints, or physical projections during sampling to improve spatial control and plausibility (Karunratanakul et al., 2023; Xie et al., 2024; Yuan et al., 2023). Image-conditioned 3D recovery further uses keypoint likelihoods or reprojection objectives to align generated body and hand motion with visual observations (Stathopoulos et al., 2024; Yi et al., 2025; Ye et al., 2026b). Classical image-based visual servoing maps an image-plane discrepancy to a minimum-norm spatial update through the camera Jacobian (Hutchinson et al., 1996). DPG brings this geometric principle into joint video–hand flow matching by using an independently learned observation to derive an analytic direction that redirects hand-motion generation.

3 Proposed Method

Generating paired video and 3D hand motion from a reference image, text prompt, and initial hand state can provide interaction data for robot pretraining. Existing conditional approaches face a circular dependency, as generating either modality requires the other as input. To resolve this dependency, we propose Uni-HaVi, a unified flow-matching model supporting video-to-hand-motion generation, hand-motion-to-video generation and joint generation.

Task formulation. Formally, let $\mathcal{V} \in \mathbb{R}^{T \times D_v}$ denote an egocentric RGB video of T frames with D_v-dimensional frames and $\mathcal{H} \in \mathbb{R}^{T \times 2 \times D_h}$ its paired bimanual hand motion (Romero et al., 2017), with D_h the hand dimensions. The shared condition $\mathcal{C} = \{I^0, c, \mathcal{H}^0\}$ contains the reference image I^0, text prompt c, and initial hand state $\mathcal{H}^0$. Calibrated cameras define the DPG projection and

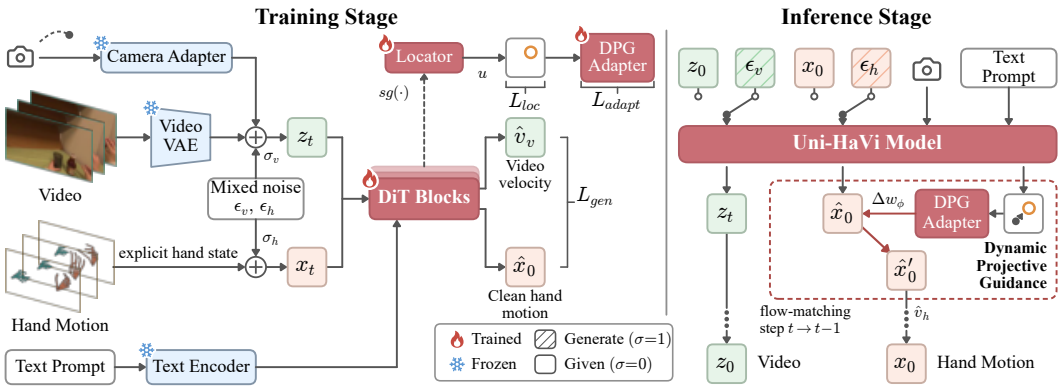


Figure 2: **Dynamic Projective Guidance for Uni-HaVi overview.** Modality-specific noise levels determine whether video and hand motion are given or generated, enabling joint generation, video-to-hand-motion generation, and hand-motion-to-video generation within a single flow-matching pipeline. The video hand locator provides a 2D wrist location, which the DPG Adapter uses to progressively guide the 3D hand trajectory at each flow-matching step.

Jacobian during both training and inference. Let $\mathbf{X} = (\mathcal{V}, \mathcal{H})$, and let $\mathcal{O} \subseteq \{v, h\}$ index its observed components. With $\overline{\mathcal{O}} = \{v, h\} \setminus \mathcal{O}$ denoting the complementary set of modalities to be generated, we build one model with parameter θ to handle both joint generation and cross-model generation:

$$\widehat{\mathbf{X}}_{\overline{\mathcal{O}}} \sim p_\theta(\mathbf{X}_{\overline{\mathcal{O}}} \mid \mathbf{X}_{\mathcal{O}}, \mathcal{C}), \qquad \mathcal{O} \in \{v, h\}. \tag{1}$$

Here $\mathcal{O}$ indicates whether the model performs joint generation, video-to-hand-motion, or hand-motion-to-video. For example, $\mathcal{O} = \varnothing$ gives $\overline{\mathcal{O}} = \{v, h\}$, corresponding to joint generation of video and hand motion, conditioned on reference image, text prompt, and initial hand state. These three generation modes are realized through modality-specific noise levels, while DPG progressively improves spatio-temporal consistency between video and 3D hand motion, as illustrated in Figure 2.

3.1 VIDEO AND HAND-MOTION GENERATION

Joint generation with flow matching. To jointly generate video and hand motion, we define a linear flow-matching path for each (Lipman et al., 2022). Video follows a flow-matching path in its learned VAE latent space, whereas hand motion follows a flow-matching path directly in its explicit parametric space. Let $\mathbf{z}_0 = \mathcal{E}(\mathcal{V})$ denote the clean video latent and $\mathbf{x}_0 := \mathcal{H} \in \mathbb{R}^{T \times 2 \times D_h}$ the clean hand trajectory. Given modality-specific noise levels $\sigma_v, \sigma_h \in [0, 1]$, their flow-matching paths are

$$\mathbf{z}_t = (1 - \sigma_v)\mathbf{z}_0 + \sigma_v \boldsymbol{\epsilon}_v, \qquad \mathbf{x}_t = (1 - \sigma_h)\mathbf{x}_0 + \sigma_h \boldsymbol{\epsilon}_h, \qquad \boldsymbol{\epsilon}_v, \boldsymbol{\epsilon}_h \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

The video head predicts the flow velocity $\boldsymbol{\epsilon}_v - \mathbf{z}_0$, while the hand head directly predicts the clean hand motion $\mathbf{x}_0$.

Conditional generation with modality-specific noise levels. The same flow-matching formulation also supports conditional generation. For example, setting $\sigma_v = 0$ treats the video as observed, while generating the corresponding 3D hand motion. Conversely, setting $\sigma_h = 0$ fixes the 3D hand motion and generates the corresponding video, as shown in the inference stage of Figure 2. Across all three generation modes, a shared Transformer backbone (Peebles & Xie, 2023) jointly processes video and hand-motion tokens, enabling cross-modal interaction through joint attention. However, such a Transformer does not have geometric constraint between the generated video and 3D hand motion.

3.2 DYNAMIC PROJECTIVE GUIDANCE

To address this limitation, we introduce DPG (Figure 3) to progressively guide 3D hand-motion generation using 2D wrist locations that emerge early during video generation. For each frame f and hand $k \in \{L, R\}$, let $\mathbf{x}_{0,f,k} \in \mathbb{R}^{D_h}$ denote the corresponding clean hand state. For readability, we omit the frame and hand indices f, k below when unambiguous. Accordingly, the 2D wrist

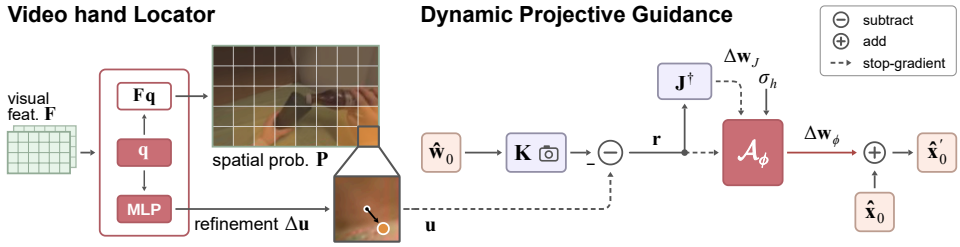


Figure 3: **Dynamic Projective Guidance.** The video hand Locator predicts a 2D wrist location from video features using a spatial probability map with local refinement. DPG computes the 2D reprojection discrepancy $\mathbf{r}$ between this location and the projected 3D wrist, and maps it to a Jacobian-based 3D update $\Delta\mathbf{w}_J$. A lightweight Adapter adjusts $\Delta\mathbf{w}_J$ based on the current noise level, guiding the 3D hand trajectory at each flow-matching step.

location $\mathbf{u} \in \mathbb{R}^2$ constrains the corresponding 3D wrist position $\mathbf{w} \in \mathbb{R}^3$, expressed in the current camera coordinate frame, to

$$\mathcal{S}(\mathbf{u}) = \left\{ \mathbf{w} \in \mathbb{R}^3 \;\middle|\; \tilde{\mathbf{u}} \sim \mathbf{K}\mathbf{w},\ [\mathbf{w}]_z > 0 \right\}, \tag{3}$$

where $\tilde{\mathbf{u}} = [\mathbf{u}^\top, 1]^\top \in \mathbb{R}^3$ is the homogeneous image coordinate, $\mathbf{K} \in \mathbb{R}^{3\times3}$ is the camera intrinsic matrix, and $[\mathbf{w}]_z > 0$ enforces positive wrist depth in the current camera frame.

Flow-matching guidance. At each flow-matching step, DPG produces a 3D wrist correction $\Delta\mathbf{w}_\phi \in \mathbb{R}^3$ to guide the 3D hand generation moving toward $\mathcal{S}(\mathbf{u})$. Note this correction only updates the 3D location of wrist:

$$\hat{\mathbf{x}}'_0 = \hat{\mathbf{x}}_0 + \begin{bmatrix} \Delta\mathbf{w}_\phi \\ \mathbf{0} \end{bmatrix}, \qquad \hat{\mathbf{v}}_h = \frac{\hat{\mathbf{x}}_t - \hat{\mathbf{x}}'_0}{\sigma_h}, \tag{4}$$

where $\mathbf{0} \in \mathbb{R}^{D_h-3}$ is a zero vector, indicating the remaining hand parameters are unchanged; $\hat{\mathbf{v}}_h$ is the hand-motion flow velocity corresponding to the corrected clean prediction at its noise level σ_h.

Reprojection constraint. We first compute a 2D reprojection discrepency $\mathbf{r} \in \mathbb{R}^2$ between wrist location and the current 3D wrist prediction $\hat{\mathbf{w}}_0 \in \mathbb{R}^3$ expressed in the current camera frame, and map it through the camera Jacobian $\mathbf{J} \in \mathbb{R}^{2\times3}$ to a Jacobian-based 3D update $\Delta\mathbf{w}_J \in \mathbb{R}^3$,

$$\mathbf{r} = \mathbf{u} - \frac{(\mathbf{K}\hat{\mathbf{w}}_0)_{1:2}}{(\mathbf{K}\hat{\mathbf{w}}_0)_3}, \qquad \Delta\mathbf{w}_J = \mathbf{J}^\dagger\mathbf{r}. \tag{5}$$

where $\mathbf{J}^\dagger$ denotes the pseudoinverse of the camera Jacobian $\mathbf{J}$. To obtain a better correction $\Delta\mathbf{w}_\phi$ that adapts the flow matching process, a DPG Adapter $\mathcal{A}_\phi$ learns to adjust $\Delta\mathbf{w}_J$ to the current hand flow-matching state with discrepency $\mathbf{r}$ at noise level σ_h,

$$\Delta\mathbf{w}_\phi = \mathcal{A}_\phi(\Delta\mathbf{w}_J, \mathbf{r}, \sigma_h). \tag{6}$$

Coarse-to-fine visual localization. To find the wrist location $\mathbf{u}$ more accurately, we further design a Locator to predict a spatial probability map $\mathbf{P} \in \mathbb{R}^N$ from the video features $\mathbf{F} \in \mathbb{R}^{N\times D_F}$ and a learned hand-side query $\mathbf{q} \in \mathbb{R}^{D_F}$, where N is the number of visual tokens per frame and D_F is the visual feature dimension. The peak of $\mathbf{P}$ gives the coarse wrist location, which is then locally refined to obtain the continuous 2D wrist location $\mathbf{u}$.

$$\mathbf{P} = \mathrm{softmax}\left(\frac{\mathbf{F}\mathbf{q}}{\tau\sqrt{D_F}}\right), \qquad \hat{s} = \arg\max_s P_s, \qquad \mathbf{u} = \mathbf{c}_{\hat{s}} + \Delta\mathbf{u}_{\hat{s}}. \tag{7}$$

where $\hat{s}$ denotes the index of the peak in the spatial probability map, $\tau > 0$ is a temperature controlling its sharpness, $\mathbf{c}_{\hat{s}} \in \mathbb{R}^2$ is the corresponding visual-token center, and $\Delta\mathbf{u}_{\hat{s}} \in \mathbb{R}^2$ is a learned local refinement from the corresponding visual feature $F_{\hat{s}}$ and hand-side query q (Figure 3).

3.3 Training Objectives

During base-model training, we uniformly sample among the three generation settings by choosing the corresponding modality-specific noise levels. The major training loss functions are listed below.

Video Hand Locator. To provide accurate 2D wrist location $\mathbf{u}$ for DPG, we supervise the spatial probability map $\mathbf{P}$ and local refinement $\Delta\mathbf{u}$, using the ground-truth 2D wrist location $\mathbf{u}_{\mathrm{gt}}$ and its corresponding token index $s_{\mathrm{gt}} \in \{1, \ldots, N\}$,

$$\mathcal{L}_{\mathrm{loc}} = \mathbb{E}\left[\mathrm{CE}(\mathbf{P}, s_{\mathrm{gt}}) + \frac{\lambda_u}{2}\left\|\Delta\mathbf{u} - (\mathbf{u}_{\mathrm{gt}} - \mathbf{c}_{s_{\mathrm{gt}}})\right\|_2^2\right]. \tag{8}$$

where CE denotes cross-entropy over the spatial probability map. The Locator predicts wrist locations from the visual features, with localization gradients blocked from the base model to avoid altering the representation.

DPG Adapter. Using the video hand locator, we further train the Adapter to learn an 3D wrist correction $\Delta\mathbf{w}_\phi$ for updating the 3D wrist location $\mathbf{w}$, via minimizing its 2D reprojection discrepancy and maintaining temporal consistency of the 3D wrist correction,

$$\begin{aligned}\mathcal{L}_{\mathrm{adapt}} = \ & \mathrm{SL1}\big(\Delta\mathbf{w}_\phi, \Delta\mathbf{w}_J\big) + 1 - \cos\left(\Delta\mathbf{w}_\phi, \Delta\mathbf{w}_J\right) \\ & + \lambda_{\mathrm{p}}\,\mathrm{SL1}\left(\pi_{\mathbf{K}}(\widehat{\mathbf{w}}_0 + \Delta\mathbf{w}_\phi), \mathbf{u}\right) + \lambda_{\mathrm{t}}\left\|\nabla_f^2\Delta\mathbf{w}_\phi - \nabla_f^2\Delta\mathbf{w}_J\right\|_1.\end{aligned} \tag{9}$$

Here, $\Delta\mathbf{w}_J$ is the learning objective from 2D wrist location; SL1 denotes the Smooth-L1 loss, cos denotes cosine similarity, and ∇_f^2 denotes the second-order temporal difference across frames.

Flow matching. The shared generative model is trained using flow matching:

$$\mathcal{L}_{\mathrm{gen}} = \mathbb{E}\Big[\|\widehat{\mathbf{v}}_v - (\boldsymbol{\epsilon}_v - \mathbf{z}_0)\|_2^2 + \lambda_h\|\widehat{\mathbf{x}}_0 - \mathbf{x}_0\|_2^2\Big] + \lambda_{\mathrm{smth}}\left\|\nabla_f^2\widehat{\mathbf{x}}_0 - \nabla_f^2\mathbf{x}_0\right\|_1. \tag{10}$$

Beyond per-frame reconstruction, the second-order temporal term aligns the predicted hand-motion dynamics with the ground truth. Together, we optimize $\mathcal{L}_1 = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{loc}}\mathcal{L}_{\mathrm{loc}}$ and $\mathcal{L}_2 = \mathcal{L}_{\mathrm{adapt}}$ in two stages. These objectives enable a single model to support joint generation and both conditional directions, while DPG improves spatio-temporal consistency between video and 3D hand motion.

4 Experiments

4.1 Experimental Setup

Dataset. HOI4D (Liu et al., 2022) provides a relatively small, high-quality dataset of short, action-segmented manipulation clips with accurate MANO annotations. In contrast, EgoDex (Hoque et al., 2026) contains longer recordings across a broader range of everyday manipulation tasks, with synchronized hand and camera tracking, but its hand annotations require retargeting to MANO (Appendix D). We follow the official splits, with approximately 9k and 36k training clips for HOI4D and EgoDex, respectively. For both datasets, each sample comprises an RGB clip, paired 3D hand motion, a text description, and a camera trajectory.

Metrics. For joint generation, where no paired ground truth exists, we use distribution-level metrics against real test data: FVD (Unterthiner et al., 2018) for video quality and Motion-FID from an adapted HandX evaluator (Zhang et al., 2026b) for hand-motion quality. CLIP-T (Radford et al., 2021), motion diversity, and acceleration magnitude measure text alignment, motion variation, and temporal dynamics. When paired references are available, absolute MPJPE (Ionescu et al., 2014), wrist translation error, PA-MPJPE, and acceleration error (Zhang et al., 2025a; Yu et al., 2025) evaluate 3D hand reconstruction, while LPIPS (Zhang et al., 2018) and PSNR evaluate video fidelity. For joint generation, we additionally report hand-mask IoU, which measures whole-hand spatial overlap between the RGB hand segmentation obtained with SAM2 Ravi et al. (2025) and the projected 3D hand mask. Across all settings, wrist alignment error, PCK, and wrist-heatmap soft IoU measure cross-modal consistency using a DINOv2-based RGB wrist evaluator (Oquab et al., 2023).

Evaluator protocol. For fair comparison, both RGB evaluators operate independently of UniHaVi and have no access to Locator predictions or hand-motion parameters. The DINOv2-based wrist evaluator uses the same frozen checkpoint, preprocessing, and scoring parameters for all methods. For whole-hand evaluation, we apply the same SAM2 model and segmentation protocol to all generated videos, without method-specific calibration. The complementary wrist-level and whole-hand measurements provide independent checks of cross-modal spatial consistency. A held-out manual audit on 100 real images yields a localization error of 7.63 px and a miss rate of 3.58%; additional evaluator validation on generated videos is provided in Appendix E.

Table 1: **Joint video–hand generation on both datasets**. Results cover cross-modal consistency, hand-motion quality, and video quality. † Adapted for 3D hand-motion generation.

	Video–hand consistency				Hand-motion quality / dynamics			Video quality	
Method	Align. (px)↓	PCK (%)↑	Soft IoU↑	Hand IoU↑	Motion-FID↓	Diversity	Accel. mag. (mm/f^2)	FVD↓	CLIP-T↑
(a) EgoDex									
Paired test data (reference)	9.68	90.87	0.464	0.395	–	6.18	5.79	–	0.267
DreamZero† (Ye et al., 2026a)	62.42	9.60	0.055	0.193	3.51	6.13	16.24	196.94	0.269
Donk (Wang et al., 2026a)	126.74	1.65	0.012	0.058	11.99	5.34	165.24	379.54	0.262
HandWorld (Sun et al., 2026)	74.97	6.73	0.040	0.147	7.49	5.80	34.57	153.72	0.268
Uni-HaVi (Base Transformer)	42.32	25.67	0.124	0.245	**3.40**	5.53	5.24	231.13	0.270
Uni-HaVi (DPG)	**19.02**	**67.41**	**0.301**	**0.313**	3.54	5.96	6.20	**132.22**	**0.272**
(b) HOI4D									
Paired test data (reference)	4.30	97.19	0.650	0.409	–	4.76	13.25	–	0.240
DreamZero† (Ye et al., 2026a)	36.85	28.82	0.127	0.268	2.53	4.76	16.00	410.19	0.246
Donk (Wang et al., 2026a)	89.85	4.03	0.023	0.001	7.48	4.28	147.53	709.93	**0.248**
HandWorld (Sun et al., 2026)	43.97	17.55	0.089	0.001	4.15	4.64	27.75	466.03	0.247
Uni-HaVi (Base Transformer)	33.92	27.65	0.126	0.284	2.90	4.08	7.25	365.00	0.247
Uni-HaVi (DPG)	**21.04**	**48.89**	**0.238**	**0.392**	**2.36**	4.63	11.85	**337.86**	0.247

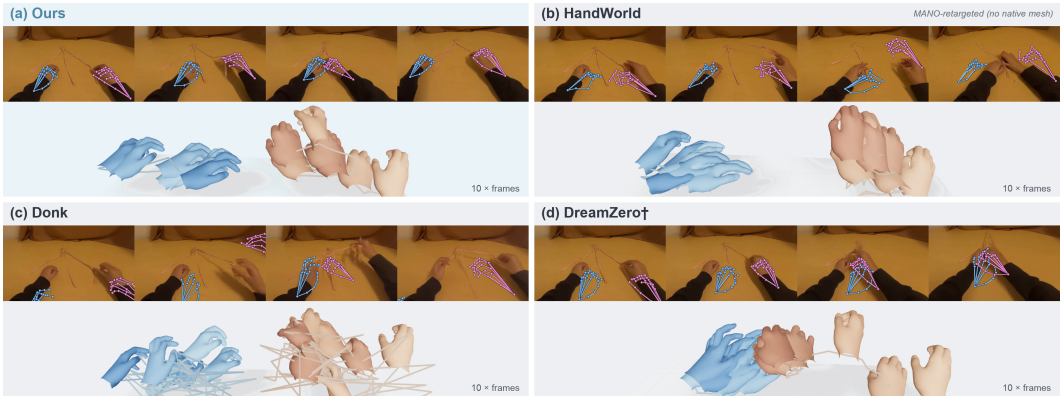


Figure 4: **Joint generation of video and hand motion.** Uni-HaVi maintains comparable visual quality while its projected 3D hand skeletons better match the visible wrist locations in the video, and produces more coherent and plausible 3D hand-motion trajectories with smoother wrist motion consistent with the generated interaction.

4.2 Unified Model for Joint and Conditional Generation

Joint generation. Uni-HaVi achieves better overall video and hand-motion quality together with substantially stronger cross-modal consistency on both datasets, as shown in Table 1. DPG consistently improves all four cross-modal metrics, with particularly large gains in wrist alignment and PCK. The improvement also extends beyond wrist localization to the hand IoU, indicating better spatial consistency over the whole hand. These improvements are achieved without sacrificing generation quality, while the generated hand motion exhibits temporal dynamics closer to real-data statistics. Figure 4 further illustrates the improved consistency between generated video and 3D hand motion, not only at the wrist but also across the whole hand.

Conditional generation. All models are provided with the same camera trajectory, initial state, and conditioning inputs. In the video-to-hand-motion setting, existing reconstruction methods are included as reference points rather than strictly matched baselines, since not all methods can be retrained under the same EgoDex setting. Under this reference comparison, Uni-HaVi shows lower errors in wrist position, articulation, and temporal accuracy, together with stronger video–hand consistency, as shown in Table 2 and Figure 5. For hand-motion-to-video generation, where the compared models are trained or finetuned under the same setting, Uni-HaVi achieves competitive video

Table 2: **Conditional generation.** Top: video-to-hand-motion generation, with existing reconstruction methods reported as reference points. Bottom: hand-motion-to-video generation under matched training or finetuning settings.

Video → Hand motion	Video–hand consistency			Hand-motion accuracy			
Method	Align. (px)↓	PCK (%)↑	Soft IoU ↑	Abs-MPJPE (mm)↓	Wrist err. (mm)↓	Accel. err. (mm/f^2)↓	PA-MPJPE (mm)↓
HaMeR (Pavlakos et al., 2024)	99.02	2.40	0.028	136.09	170.72	54.18	12.10
HaWoR (Zhang et al., 2025a)	34.94	31.97	0.164	113.08	102.83	20.08	9.54
Dyn-HaMR (Yu et al., 2025)	50.20	3.63	0.037	100.97	101.28	9.30	10.33
HandFlow (Xu et al., 2026)	40.50	15.84	0.089	92.48	88.03	8.41	10.19
Uni-HaVi (Ours)	20.46	56.68	0.251	27.52	23.19	5.61	4.07

Hand motion → Video	Condition alignment			Video quality			
Method	Align. (px)↓	PCK (%)↑	Soft IoU ↑	FVD ↓	CLIP-T ↑	LPIPS ↓	PSNR (dB)↑
Wan2.2-5B (FT) (Wan et al., 2025)	58.39	3.34	0.034	338.83	0.263	0.321	18.74
InterDyn (Akkerman et al., 2025)	34.29	42.94	0.215	202.85	0.268	0.255	22.25
JointControlVideo (Zhang et al., 2026a)	**28.09**	43.54	0.225	**190.01**	**0.271**	0.210	22.95
EgoSim (Hao et al., 2026)	29.42	38.46	0.188	224.79	0.268	0.213	23.93
Uni-HaVi (Ours)	33.83	**46.56**	**0.228**	198.52	0.269	**0.180**	**24.23**

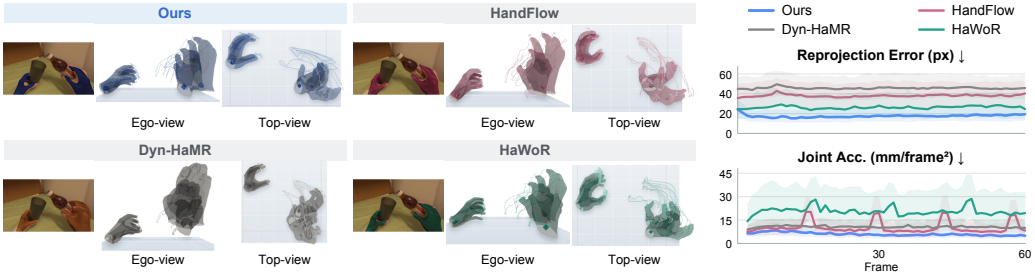


Figure 5: **Video-to-hand-motion generation.** Uni-HaVi recovers 3D hand motion that more closely follows the observed video, with lower wrist reprojection error and more temporally coherent trajectories across frames. The quantitative results in Table 2 further show lower errors in wrist position, articulation, and temporal accuracy under the reference comparison.

quality while maintaining strong alignment with the supplied hand motion. Together, these results show that a single model supports both precise 3D hand-motion generation and hand-conditioned video synthesis. Additional qualitative comparisons are provided in Appendix C.

4.3 Ablation Study

Table 3: **Dynamic guidance ablation.** Late and Full DPG substantially improve wrist alignment, while Early DPG and direct analytic guidance provide little gain. The 2D wrist location stabilizes early, whereas the projected 3D wrist aligns mainly during later flow-matching steps.

Setting	Guidance steps	Align. (px)↓	PCK (%)↑	Soft IoU↑
Off	–	42.32	25.67	0.124
Analytic	Full	34.30	29.57	0.144
Early DPG	0–9	34.34	29.67	0.144
Late DPG	40–49	19.14	67.27	0.299
Full DPG	Full	**19.02**	**67.41**	**0.301**



We ablate the guidance timing and the learned Adapter in DPG on Egodex(Table 3). All five settings use the same model weights, camera inputs, and initial noise over 50 flow-matching steps. Analytic guidance directly applies the Jacobian-based update, whereas DPG adapts this update using the

reprojection residual and noise level. The middle panel shows that the 2D wrist location stabilizes early and remains stable, while the projected 3D wrist error decreases primarily during the later flow-matching steps. Late DPG recovers most of the alignment improvement of Full DPG, whereas Early DPG and direct analytic guidance provide only limited gains. This pattern suggests that early guidance may be difficult to accumulate or preserve over subsequent flow-matching steps, while direct analytic updates may be less adaptive to the evolving hand state.

4.4 Physical Interaction in Simulation

As an exploratory robotics study, we investigate whether jointly generated hand motion can be directly used for downstream robotic learning and deployment. We consider a pouring example using a shared reinforcement-learning (RL) pipeline in simulation, and compare the hand motion jointly generated by our model with trajectories reconstructed from the same generated video using HaMeR (Pavlakos et al., 2024) and HaWoR (Zhang et al., 2025a). All variants use the same generated video, object meshes reconstructed with SAM 3D Objects (Chen et al., 2026b), and object poses estimated with FoundationPose (Wen et al., 2024). In this example, our jointly generated hand motion supports successful pouring in simulation while remaining consistent with the generated video. HaMeR also completes the task but produces a less human-like hand pose, whereas HaWoR does not complete the task. We further transfer the resulting policy to a real robot, demonstrating the feasibility of using our generated hand motion for downstream robotic deployment (Figure 6).

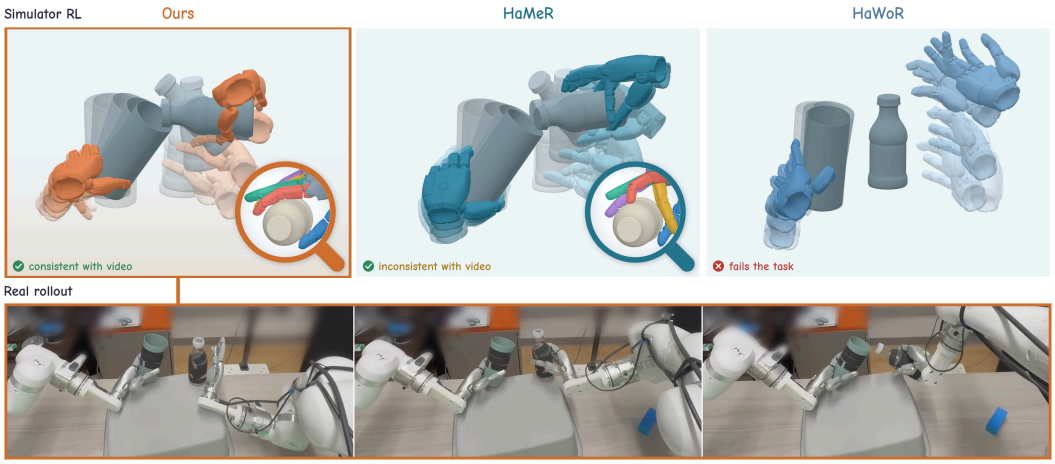


Figure 6: **Downstream robotics demonstration.** Top: qualitative example in simulation under the same reconstruction and RL pipeline. Bottom: real-robot deployment of the resulting policy.

5 Conclusion

We introduced Uni-HaVi, a unified flow-matching model that supports joint generation of egocentric video and 3D hand motion and both conditional directions within a single model. Leveraging the observation that the 2D wrist trajectory stabilizes earlier than the 3D hand trajectory, DPG uses predicted 2D wrist locations to progressively guide 3D hand motion during flow matching. Across joint and conditional generation, Uni-HaVi achieves substantially stronger video–hand alignment and improved hand-motion quality while maintaining competitive video quality. An exploratory simulation study and real-robot transfer further demonstrate the feasibility of jointly generated hand motion for downstream policy learning and deployment.

Limitations. **(1) Camera calibration.** DPG relies on camera intrinsics and time-varying extrinsics, which may be inaccurate or unavailable in videos; future work could jointly generate camera motion with the video. **(2) Depth ambiguity.** Because DPG derives guidance from 2D wrist locations, it constrains image-plane alignment more directly than depth. **(3) Interaction modeling.** Uni-HaVi does not explicitly model object states or contact dynamics, so the generated hand trajectories serve as visually aligned references for downstream retargeting and policy learning.

AI USE STATEMENT

Generative AI tools were used only for language polishing, and improving the presentation of figures. All research decisions, results, claims, figures, and final content were reviewed and verified by the authors.

REFERENCES

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

Fan Bao, Shen Nie, Kaiwen Xue, Chongxuan Li, Shi Pu, Yaole Wang, Gang Yue, Yue Cao, Hang Su, and Jun Zhu. One transformer fits all distributions in multi-modal diffusion at scale. In *International Conference on Machine Learning*, pp. 1692–1717. PMLR, 2023.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, et al. Motus: A unified latent action world model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 35101–35113, 2026.

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. *arXiv preprint arXiv:2604.03305*, 2026a.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7220–7232, 2026b.

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6341–6351. IEEE, 2024.

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

Ryan Hoque, Peide Huang, David Yoon, Jian Zhang, et al. EgoDex: Learning dexterous manipulation from large-scale egocentric video. In *International Conference on Learning Representations*, volume 2026, pp. 4218–4237, 2026.

Seth Hutchinson, Gregory D Hager, and Peter I Corke. A tutorial on visual servo control. *IEEE transactions on robotics and automation*, 12(5):651–670, 1996.

C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE TPAMI*, 2014. Origin of the MPJPE metric.

Elkhan Ismayilzada, Yufei Zhang, and Zijun Cui. Pad-hand: Physics-aware diffusion for hand motion recovery. *arXiv preprint arXiv:2603.26068*, 2026.

Korrawe Karunratanakul, Konpat Preechakul, Supasorn Suwajanakorn, and Siyu Tang. Guided motion diffusion for controllable human motion synthesis. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2151–2162. IEEE, 2023.

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*, pp. 29663–29673, 2026.

Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, et al. A versatile diffusion transformer with mixture of noise levels for audiovisual generation. *Advances in neural information processing systems*, 37: 11837–11865, 2024.

Gen Li, Bo Zhao, Jianfei Yang, and Laura Sevilla-Lara. Mask2iv: Interaction-centric video generation via mask trajectories. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6091–6099, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified Video Action Model. In *Proceedings of Robotics: Science and Systems*, LosAngeles, CA, USA, June 2025a. doi: 10.15607/RSS.2025.XXI.074.

Shufan Li, Konstantinos Kallidromitis, Akash Gokul, Zichun Liao, Yusuke Kato, Kazuki Kozuka, and Aditya Grover. Omniflow: Any-to-any generation with multi-modal rectified flows. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13178–13188. IEEE, 2025b.

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

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 *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20981–20990. IEEE, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Bhawna Paliwal, Haritheja Etukuru, William Liang, Pieter Abbeel, Nur Muhammad Mahi Shafiullah, and Jitendra Malik. Do as i do: Dexterous manipulation data from everyday human videos. *arXiv preprint arXiv:2606.19333*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12209–12219. IEEE, 2025.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. **HaMeR**.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 906–915. IEEE, 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*, pp. 15976–15985, 2026.

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

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

Dingrui Wang, YuAn Wang, Jinkun Liu, Yue Zhang, Mattia Piccinini, Yu Sun, and Johannes Betz. Unified video-action joint denoising for dexterous action and data generation, 2026a. arXiv:2606.03868. TUM + ByteDance + Tsinghua.

Zun Wang, Jaemin Cho, Jialu Li, Han Lin, Jaehong Yoon, Yue Zhang, and Mohit Bansal. Epic: Efficient video camera control learning with precise anchor-video guidance. In *Forty-third International Conference on Machine Learning*, 2026b.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Yiming Xie, Varun Jampani, Lei Zhong, Deqing Sun, and Huaizu Jiang. Omnicontrol: Control any joint at any time for human motion generation. In *International Conference on Learning Representations*, volume 2024, pp. 28176–28194, 2024.

Mingxi Xu, Bowen Duan, Yi Gu, Zhengyang Shen, Renjing Xu, and Yutao Yue. HandFlow: Fully generative 4D hand recovery with flow matching. *arXiv preprint arXiv:2607.11221*, 2026.

Seonghyeon Ye et al. World action models are zero-shot policies, 2026a. `https://dreamzero0.github.io/`.

Yufei Ye, Jiaman Li, Ryan Rong, and C Karen Liu. Whole: World-grounded hand-object lifted from egocentric videos. *arXiv preprint arXiv:2602.22209*, 2026b.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7072–7084. IEEE, 2025.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27716–27726. IEEE, 2025.

Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15964–15975. IEEE, 2023.

Chenyangguang Zhang, Botao Ye, Boqi Chen, Alexandros Delitzas, Fangjinhua Wang, Marc Pollefeys, and Xi Wang. Controllable egocentric video generation via occlusion-aware sparse 3d hand joints. In *European Conference on Computer Vision*, pp. 668–687. Springer, 2026a.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815. IEEE, 2025a.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Yufei Zhang, Zijun Cui, Jeffrey O Kephart, and Qiang Ji. Diffusion-based 3d hand motion recovery with intuitive physics. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7306–7317. IEEE, 2025b.

Zimu Zhang, Yucheng Zhang, Xiyan Xu, Ziyin Wang, Sirui Xu, Kai Zhou, Bing Zhou, Chuan Guo, Jian Wang, Yu-Xiong Wang, et al. Handx: Scaling bimanual motion and interaction generation. *arXiv preprint arXiv:2603.28766*, 2026b.

Hongxiang Zhao, Xingchen Liu, Mutian Xu, Yiming Hao, Weikai Chen, and Xiaoguang Han. Taste-rob: Advancing video generation of task-oriented hand-object interaction for generalizable robotic manipulation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27683–27693. IEEE, 2025.

A Implementation details

Hand-motion representation. For frame f and hand side $k \in \{L, R\}$, the component $\mathbf{x}_{0,f,k}$ of the clean hand trajectory $\mathbf{x}_0 = \mathcal{H}$ is represented as

$$\mathbf{x}_{0,f,k} = \left[\mathbf{w}_{f,k}^{(0)}; \mathbf{r}_{f,k}^{\text{root}}; \mathbf{r}_{f,k}^{\text{art}}\right] \in \mathbb{R}^{3+6+15\times 6} = \mathbb{R}^{99}.$$

Here $\mathbf{w}_{f,k}^{(0)} \in \mathbb{R}^3$ is the wrist translation in the first-frame camera coordinate frame and is measured in meters. The global wrist orientation is $\mathbf{r}_{f,k}^{\text{root}} \in \mathbb{R}^6$. The 15 articulated-joint rotations form $\mathbf{r}_{f,k}^{\text{art}} \in \mathbb{R}^{15\times 6}$. Both use a continuous 6D rotation representation, in which each vector stores the first two columns of a rotation matrix. The 10-D MANO shape vector $\boldsymbol{\beta}$ is not included in the generated hand-motion state. For visualization, MANO decoding augments the 15 articulated joints with the wrist and five mesh-derived fingertips to form the conventional 21-joint skeleton.

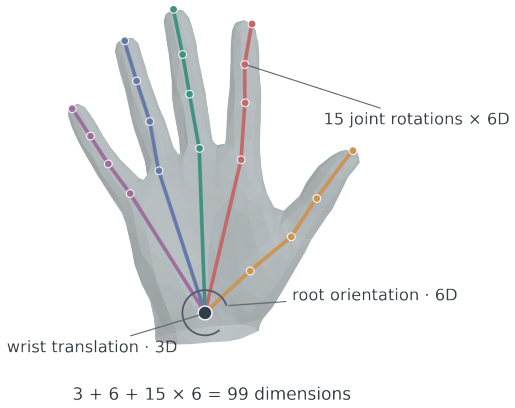


Figure 7: **Per-frame MANO hand-motion representation.** The generated 99-D state concatenates global translation, global orientation, and 15 local joint rotations.

Model configuration. Table 4 summarizes the representation, architecture, and projective-guidance configuration used in our experiments for reproducibility.

The base-model stage in Section 3.3 uses $\lambda_{\text{loc}} = 10^{-4}$ and $\lambda_{\text{smth}} = 0.3$. Adapter training combines response matching, direct alignment of the corrected projection to the detached 2D wrist location, and temporal acceleration matching to the analytic displacement, with weights 1.0, 0.3, and 10.0, respectively. In Equation 10, squared errors are averaged over valid generated elements for video and each hand separately, with the two hand losses summed. For clarity, the equation omits the scheduler's timestep-dependent weighting of the video loss; hand-motion reconstruction uses no additional timestep-dependent multiplier. Clean anchors, absent hands, and fully observed modalities are excluded from generative supervision, while Equation 8 averages valid hand-frame observations. In Equation 9, response matching combines batch-scale-normalized Smooth-L_1 regression with a cosine-direction penalty, and reprojection supervision uses Smooth-L_1 error normalized by the visual-grid dimensions. The temporal L_1 term averages errors over valid frame triplets and translation coordinates for each hand, then sums the two hand losses.

Table 4: **Architecture and implementation configuration.**

Setting	Value
Representation and tokenization	
Backbone	WAN 2.2 TI2V-5B (Wan et al., 2025), 30 DiT blocks
RGB / VAE latent frames	61 / 16
Video resolution	480×832
VAE latent grid ($T \times H \times W$)	$16 \times 30 \times 52$
Video patch size / token grid	$1 \times 2 \times 2$ / $16 \times 15 \times 26$
Number of video tokens	6,240
Hand-motion tokenization	99-D state to one token per side and frame
Number of hand-motion tokens	$2 \times 61 = 122$
Hand-motion token embeddings	side, temporal, and modality
Clean hand-motion anchor	frame 0, excluded from the prediction target
Shared denoiser and conditioning	
Shared hidden width	3072
Joint self-attention	symmetric in all 30 blocks
Camera conditioning	pretrained Plücker adapter (Wang et al., 2026b), added to video tokens
Hand-to-video residual blocks	14 / 19 / 24 / 29
Locator feature width	256
Locator temperature τ	0.25
Prediction, objectives, and projective guidance	
Base smoothness weight λ_{smth}	0.3
DPG Adapter hidden width / output	256 / 3D wrist correction
Adapter loss weights	response 1.0 / UV 0.3 / temporal 10.0
Minimum camera-frame depth	0.05 m

B PROJECTION DETAILS

Camera transform. Each wrist translation is represented in camera-0 coordinates. For frame f, the rigid transform to the instantaneous camera is

$$\mathbf{T}_{f\leftarrow 0} = \mathrm{w2c}_f \, \mathrm{w2c}_0^{-1} \in \mathrm{SE}(3). \tag{11}$$

Let the source image be $H_s \times W_s$ and the training frame be $H \times W$. Aspect-preserving resize uses

$$s = \max\left(\frac{W}{W_s}, \frac{H}{H_s}\right), \qquad \mathbf{o}_{\mathrm{crop}} = \frac{1}{2}\begin{bmatrix} sH_s - H \\ sW_s - W \end{bmatrix}, \tag{12}$$

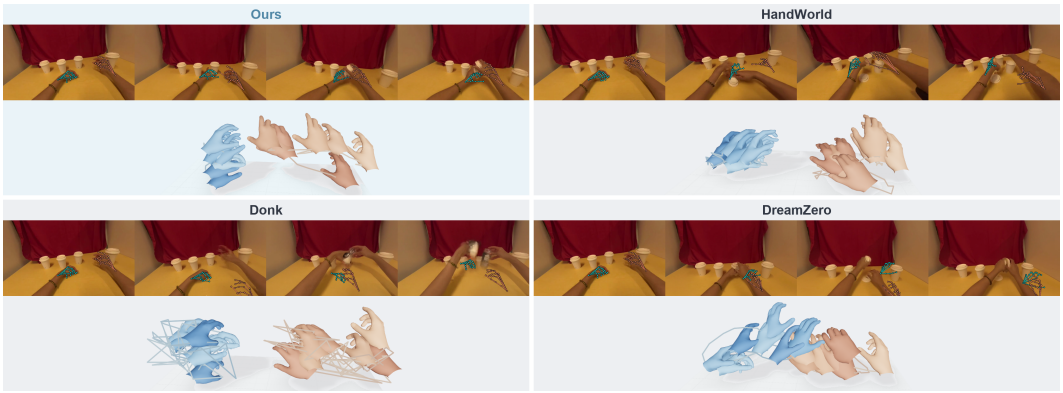
where $\mathbf{o}_{\mathrm{crop}}$ is rounded consistently with RGB preprocessing. After perspective projection, we subtract this crop offset and scale the point to the DiT patch grid. Applying the same operator to annotations and provisional clean-state predictions avoids a deterministic coordinate bias.

Masked hand objective. Let $\Gamma_k \subseteq \{1, \ldots, T-1\}$ denote valid non-anchor frames for side $k \in \{L, R\}$. Translation and rotation losses are averaged within their respective dimensions:

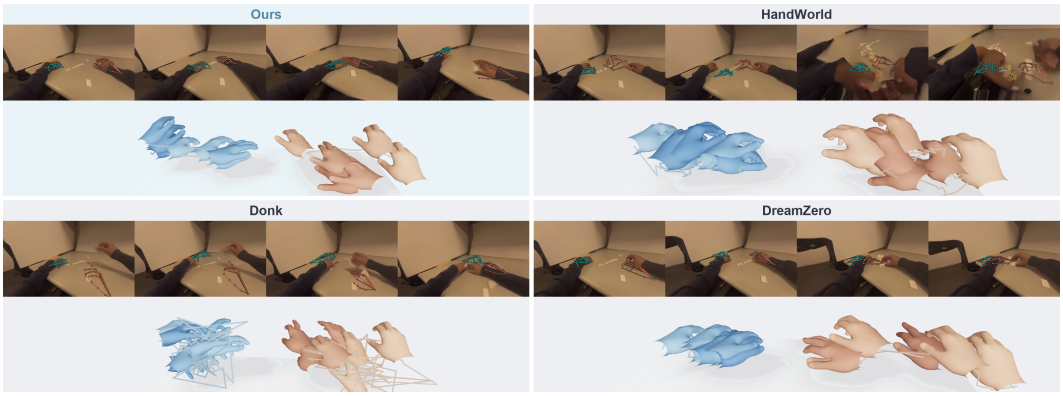
$$\mathcal{L}_h^{(k)} = \frac{1}{|\Gamma_k|} \sum_{f \in \Gamma_k} \left[\frac{1}{3} \left\| \widehat{\mathbf{w}}_{f,k}^{(0)} - \mathbf{w}_{f,k}^{(0)} \right\|_2^2 + \frac{1}{96} \left\| \widehat{\mathbf{r}}_{f,k} - \mathbf{r}_{f,k} \right\|_2^2 \right], \tag{13}$$

where $\mathbf{w}_{f,k}^{(0)} \in \mathbb{R}^3$ is the wrist translation and $\mathbf{r}_{f,k} \in \mathbb{R}^{96}$ concatenates the global and articulated 6D rotations. Frames with an absent hand are excluded rather than represented as zero-valued supervision.

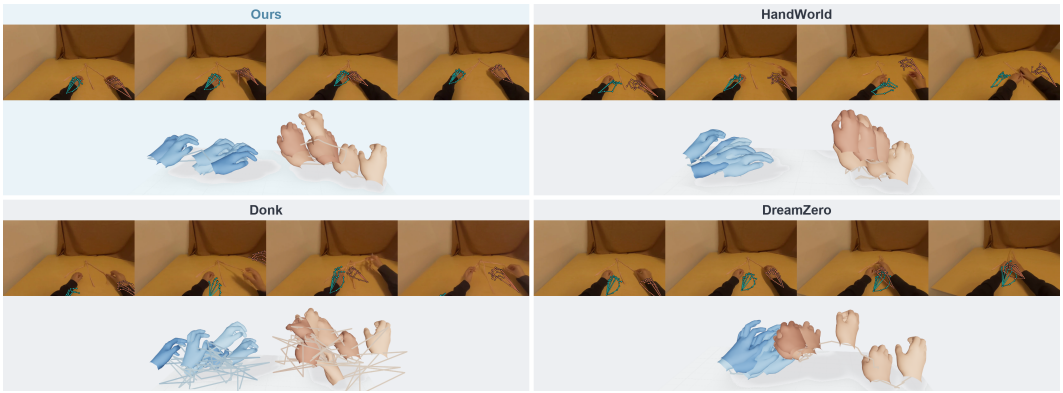
C SUPPLEMENTARY VISUALIZATIONS



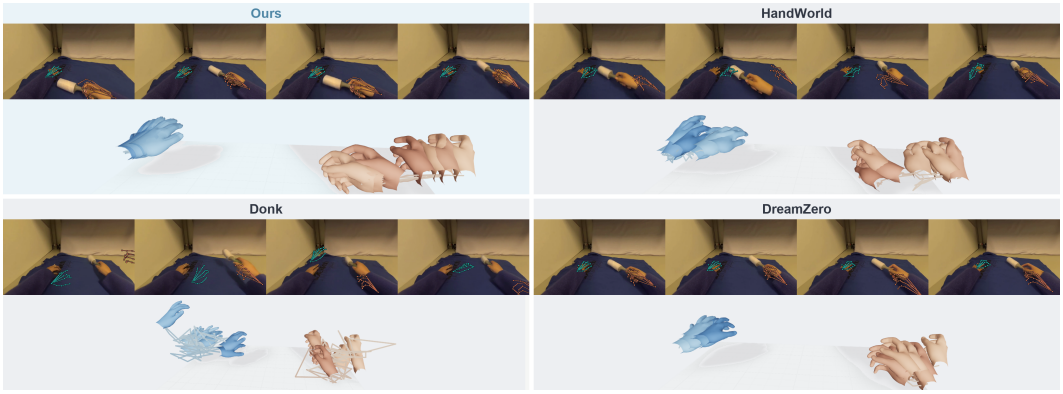
1) Remove lids from four cups placed on a wooden table with a red background.



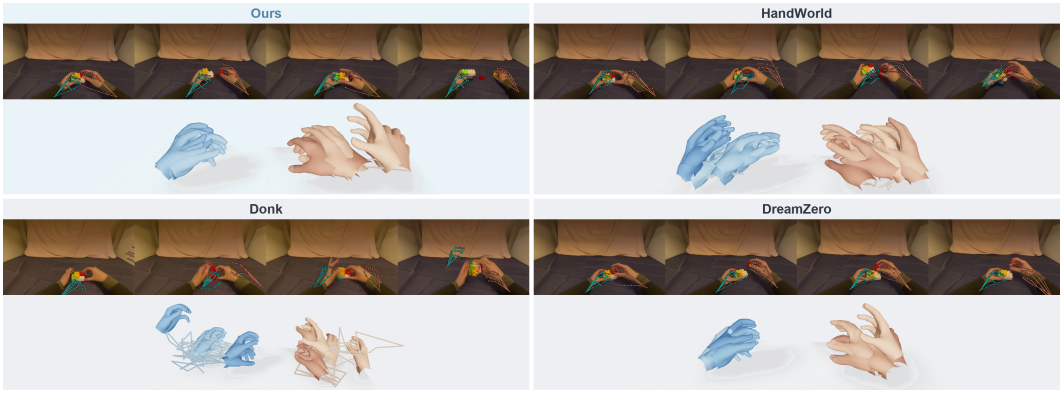
2) Topple the arranged dominoes on the green table.



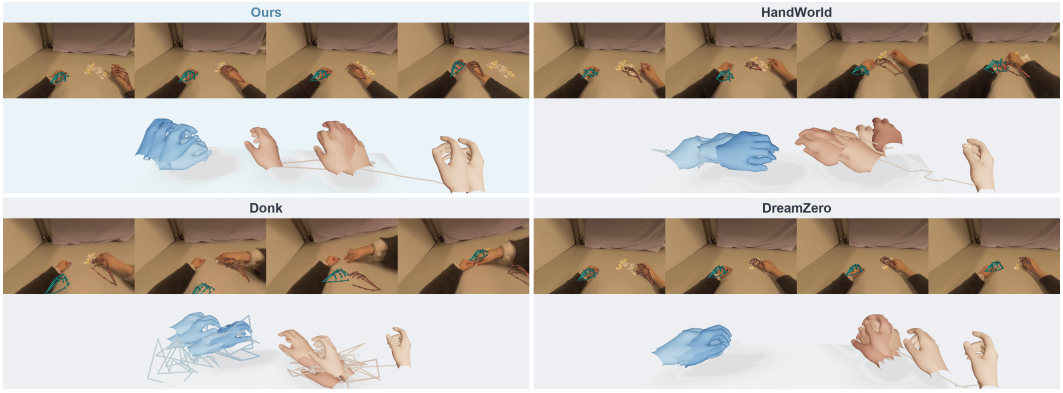
3) Braid the three strands of the striped ribbon.



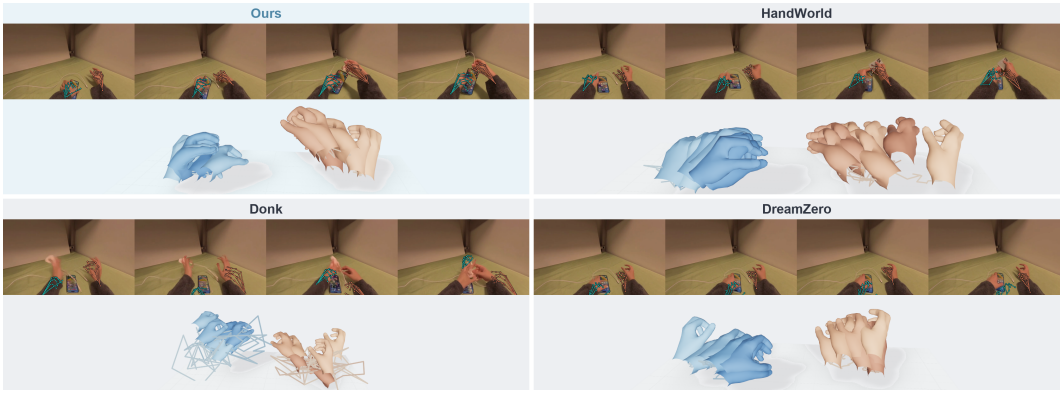
4) Clean a large shirt surface using a roller while sitting.



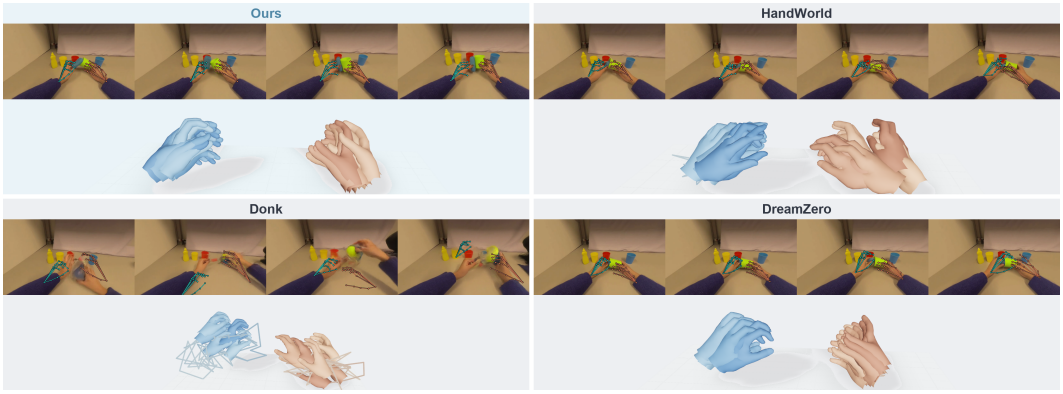
5) Disassemble six legos on a blue tablecloth while sitting against a pink background.



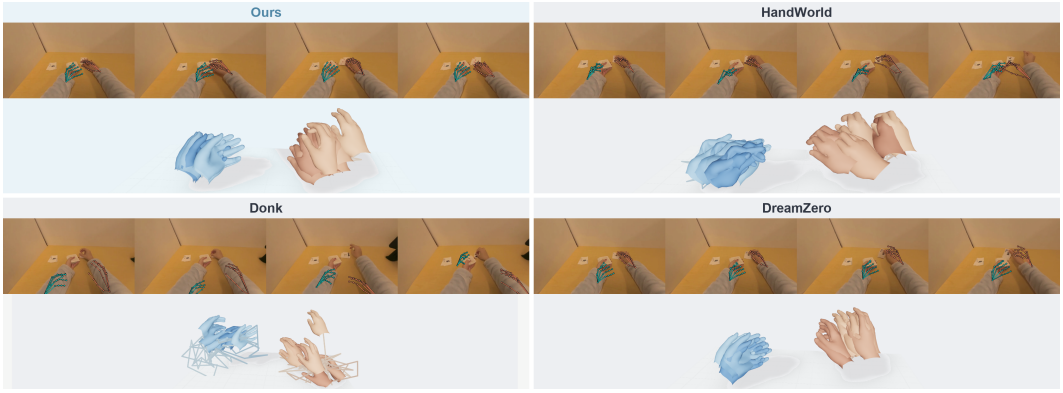
6) Arrange dominoes upright in a pattern on a table with a lavender background.



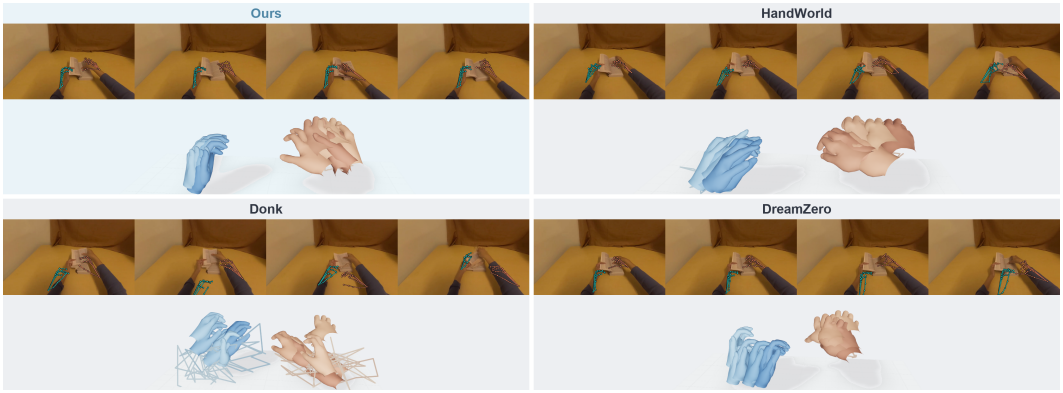
7) Plug the white phone into the lightning cable while sitting at a green tablecloth.



8) Clean four cups using a sponge and soap bottle on a pink tablecloth with a lavender background.



9) Deal cards onto a wooden table for a game of pyramid solitaire while sitting.



10) Flip pages of a paperback book placed on a wooden table while sitting against a brown background.

Figure 8: **More visualizations for the joint video and hand-motion generation task.**



Figure 9: **Qualitative comparison of hand-motion-to-video generation.**

D EgoDex dataset preparation

EgoDex pairs RGB video with tracked joint transforms, whereas our model requires synchronized video, parametric hand motion, and camera trajectories. We therefore convert the tracked skeletons into the hand-motion representation in Appendix A, then apply the same temporal sampling to all three components. This conversion must preserve the annotated wrist trajectory and camera geometry so that fitting and preprocessing do not introduce additional video–hand spatial misalignment.

D.1 Joint correspondence and camera coordinates

The source annotations contain 25 joints per hand, which do not directly match the 21-joint MANO skeleton (Romero et al., 2017). To establish the correspondence, we retain the wrist and four joints along each finger, dropping the additional metacarpal point on each non-thumb finger. The retained joints are ordered as the wrist followed by contiguous thumb, index, middle, ring, and little-finger chains, matching the decoder output shown in Figure 7. This output-joint ordering is distinct from the ordering of MANO's 15 local rotations, which remains unchanged during fitting and hand-motion encoding.

Because the retained joint transforms are expressed in a recording-level world frame, they must first be placed in the camera frame used for fitting. For an episode indexed by e with $N_e \in \mathbb{N}$ source frames, let $t \in \{0, \dots, N_e - 1\}$ index frames and $j \in \{0, \dots, 20\}$ index the retained joints. We denote the camera-to-world transform by $\mathbf{C}_{et} \in \mathrm{SE}(3) \subset \mathbb{R}^{4\times4}$ and the world transform of joint j by $\mathbf{G}_{etj} \in \mathrm{SE}(3) \subset \mathbb{R}^{4\times4}$. The corresponding metric joint position $\mathbf{y}_{etj} \in \mathbb{R}^3$ in the camera frame is

$$\mathbf{y}_{etj} = \left(\mathbf{C}_{et}^{-1}\mathbf{G}_{etj}\right)_{1:3,4}, \qquad j = 0, \dots, 20, \tag{14}$$

where $j = 0$ denotes the wrist and the hand-side index is omitted for clarity. This composition follows the source projection convention without an additional axis flip, keeping the fitting targets and RGB projection in the same camera coordinates.

D.2 Wrist-preserving hand-motion fitting

To preserve the wrist positions in Equation 14, we center the MANO decoder at the wrist and fix its translation to $\mathbf{w}_{et} := \mathbf{y}_{et0} \in \mathbb{R}^3$. Let $\bar{\mathbf{J}}_j \in \mathbb{R}^3$ denote the wrist-centered, zero-shape rest position of joint j, and let $\mathcal{P} = \{0, 5, 9, 13, 17\}$ index the wrist and four non-thumb finger bases. With translation fixed, we estimate the global orientation $\mathbf{R}^*_{et} \in \mathrm{SO}(3) \subset \mathbb{R}^{3\times3}$ by rigidly aligning the template palm to the observed palm,

$$\mathbf{R}^*_{et} = \operatorname*{arg\,min}_{\mathbf{R}\in\mathrm{SO}(3)} \sum_{j\in\mathcal{P}} \left\|\mathbf{R}\bar{\mathbf{J}}_j - (\mathbf{y}_{etj} - \mathbf{y}_{et0})\right\|_2^2, \tag{15}$$

An SVD solution with a positive-determinant correction yields a proper rotation, which is then fixed together with the wrist translation. Fixing both quantities prevents finger fitting from reducing its residual by moving the annotated wrist or rotating the entire hand.

The remaining variables are the finger pose $\boldsymbol{\phi}_{et} \in \mathbb{R}^{45}$ and hand shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$. Because shape is absent from the generated hand-motion state, we estimate one shared shape vector per side instead of assigning an unobserved shape to every episode. Let $\mathcal{J}_j : \mathrm{SO}(3) \times \mathbb{R}^{45} \times \mathbb{R}^{10} \to \mathbb{R}^3$ denote the root-centered MANO decoder for joint j, expressed in meters. The fitted position $\widehat{\mathbf{y}}_{etj} \in \mathbb{R}^3$ and scalar per-episode objective terms $\mathcal{D}_e, \mathcal{S}_e \in \mathbb{R}_{\geq0}$ are

$$\begin{aligned}
\widehat{\mathbf{y}}_{etj} &= \mathcal{J}_j(\mathbf{R}^*_{et}, \boldsymbol{\phi}_{et}, \boldsymbol{\beta}) + \mathbf{w}_{et}, \\
\mathcal{D}_e &= \frac{1}{21N_e} \sum_{t=0}^{N_e-1} \sum_{j=0}^{20} \|\widehat{\mathbf{y}}_{etj} - \mathbf{y}_{etj}\|_2^2, \\
\mathcal{S}_e &= \frac{1}{N_e - 1} \sum_{t=1}^{N_e-1} \|\boldsymbol{\phi}_{et} - \boldsymbol{\phi}_{e,t-1}\|_2^2,
\end{aligned} \tag{16}$$

where $\mathcal{S}_e = 0$ for a single-frame sequence. The data term preserves the tracked joint positions, while the temporal term discourages abrupt changes in finger pose.

Let $\boldsymbol{\phi}_e = (\boldsymbol{\phi}_{e0}, \dots, \boldsymbol{\phi}_{e,N_e-1}) \in \mathbb{R}^{N_e\times45}$ denote the finger-pose sequence for episode e. To estimate a reusable shape, Stage A jointly optimizes $\boldsymbol{\beta}$ and $\{\boldsymbol{\phi}_e\}_{e\in\mathcal{E}_A}$ over a calibration episode set $\mathcal{E}_A$, using a temporal weight $\lambda_A \in \mathbb{R}_{\geq0}$,

$$\min_{\boldsymbol{\beta}, \{\boldsymbol{\phi}_e\}_{e\in\mathcal{E}_A}} \sum_{e\in\mathcal{E}_A} \left(\mathcal{D}_e + \lambda_A \mathcal{S}_e\right). \tag{17}$$

This subset is sampled across task categories.

Stage B fixes the resulting shape and optimizes only the finger poses of each episode. To process variable-length episodes efficiently, the implementation groups them by length and pads each batch to its longest sequence. For a batch episode set $\mathcal{B}$ and temporal weight $\lambda_B \in \mathbb{R}_{\geq0}$, padding is excluded from both objective terms, giving the batch objective

$$\min_{\{\boldsymbol{\phi}_e\}_{e\in\mathcal{B}}} \frac{\sum_{e\in\mathcal{B}} N_e \mathcal{D}_e}{\sum_{e\in\mathcal{B}} N_e} + \lambda_B \frac{\sum_{e\in\mathcal{B}} (N_e - 1)\mathcal{S}_e}{\max\left(1, \sum_{e\in\mathcal{B}} (N_e - 1)\right)}. \tag{18}$$

The two normalizers count real frames and real consecutive-frame pairs, respectively, so padded entries contribute neither fitting error nor artificial temporal transitions. Let $H, W \in \mathbb{N}$ denote the training-frame height and width. Table 5 records the default optimization and clip-construction settings.

Table 5: **EgoDex preprocessing settings.**

Setting	Value
Retained joints per hand	21 of 25
Translation / global orientation	Fixed wrist / fixed palm alignment
Shape / finger-pose dimensions	10 per side / 45 per frame
Stage A optimizer / iterations	Adam / 5,000
Stage A shape / pose learning rate	0.05 / 0.02
Stage A smoothness weight λ_A	0.03
Stage B optimizer / iterations	Adam / 1,500
Stage B pose learning rate / λ_B	0.02 / 0.05
Stage B batch size / frame-slot budget	256 episodes / 80,000
Source / output frame rate	30 / 15 fps
Clip length / source stride	61 frames / 2
Minimum source length	100 frames
Short-clip padding	Repeat the last selected source frame
Hand motion / camera arrays	$61 \times 2 \times 99$ / 61×19
Hand-motion reference frame / translation unit	First-frame camera / meters
RGB training size ($H \times W$)	480×832

Fitting constraints and initialization. Table 6 specifies the nine fitting variants compared in Figure 10. The table separates initialization from optimized variables, since a fitted translation can reduce joint error by moving the hand away from the tracked wrist. The Pose+shape configuration preserves the tracked wrist translation and fixed palm alignment while fitting finger poses and episode-specific shape, reaching a mean joint error of 6.37 mm with zero wrist drift on this sequence. Because initialization, regularization, and optimization budgets also differ, these results are configuration diagnostics rather than a controlled ablation of the cross-episode shared shape used above.

D.3 Synchronized clips and a common reference frame

The fitted trajectories retain the source video's 30 Hz sampling, so clip construction must resample hand motion and RGB together. For an episode with $N \in \mathbb{N}$ source frames and $N \geq 100$, let $f \in \{0, \ldots, 60\}$ index the output frames and $a_f \in \{0, \ldots, N-1\}$ their source-frame indices. We construct one 61-frame clip using

$$a_f = \min\left(2f,\ 2\left\lfloor \frac{\min(N, 121) - 1}{2} \right\rfloor\right), \qquad f = 0, \ldots, 60. \tag{19}$$

RGB frames, fitted hand states, and camera transforms all use the same source index a_f, with the output video encoded at 15 fps. For $N \geq 121$, this selects source frames $0, 2, \ldots, 120$ and spans 4.0 seconds between the first and last frame. Shorter retained episodes repeat the last selected even-indexed frame in all three components, while episodes with fewer than 100 source frames are excluded to limit repeated content. Longer episodes contribute only this initial window rather than multiple sliding windows.

Temporal alignment alone leaves each fitted hand state in a different instantaneous camera frame. Let $\mathbf{T}_{0\leftarrow f} \in \mathrm{SE}(3) \subset \mathbb{R}^{4\times4}$ map points from the sampled frame-f camera to the first-frame camera, with rotation $\mathbf{A}_f \in \mathrm{SO}(3) \subset \mathbb{R}^{3\times3}$ and translation $\mathbf{b}_f \in \mathbb{R}^3$. The sampled wrist translation and global orientation in this common frame, denoted $\mathbf{w}_f^{(0)} \in \mathbb{R}^3$ and $\mathbf{R}_f^{(0)} \in \mathrm{SO}(3) \subset \mathbb{R}^{3\times3}$, are

$$\begin{aligned} \mathbf{T}_{0\leftarrow f} &= \mathbf{C}_{e0}^{-1}\mathbf{C}_{e,a_f} = \begin{bmatrix} \mathbf{A}_f & \mathbf{b}_f \\ \mathbf{0}_3^\top & 1 \end{bmatrix}, \\ \mathbf{w}_f^{(0)} &= \mathbf{A}_f \mathbf{w}_{e,a_f} + \mathbf{b}_f, \qquad \mathbf{R}_f^{(0)} = \mathbf{A}_f \mathbf{R}_{e,a_f}^{*}. \end{aligned} \tag{20}$$

The local finger rotations remain unchanged because they are relative to the hand rather than the camera. Combining these rotations with $\mathbf{w}_f^{(0)}$ and $\mathbf{R}_f^{(0)}$ gives the 99-D state in Appendix A, with the left and right sequences stored together in $\mathbb{R}^{61\times2\times99}$.

Table 6: **Fitting configurations grouped by joint correspondence.** Each row defines one curve in Figure 10. All variants use the same sequence. The four configurations marked † retain an incorrect joint correspondence and are diagnostic controls, not valid alternatives to the corrected fitting pipeline.

Configuration	Initialization	Translation	Global rotation	Finger pose	Shape	Updates
Incorrect joint correspondence						
Tracked init†	Tracked wrist rotation	Optimized	Optimized	Masked	Zero	1,500
Zero init†	Zero pose	Optimized	Optimized	Masked	Zero	2,000
Palm warmup†	Procrustes + palm warmup	Optimized	Optimized	Masked	Zero	2,300
Wrist-fixed†	Procrustes	Fixed	Optimized	Full	Zero	1,500
Corrected joint correspondence						
Rigid fit	Procrustes	Optimized	Optimized	Full	Zero	2,000
Bone init	Procrustes + bone directions	Fixed	Optimized	Full	Zero	2,000
Shape prior	Procrustes	Fixed	Fixed	Full	Regularized	3,000
Rigid+shape	Procrustes	Optimized	Optimized	Full	Free	5,000
Pose+shape	Procrustes	Fixed	Fixed	Full	Free	5,000

Fixed translation retains the tracked wrist, and fixed global rotation retains the Procrustes estimate. Finger poses start from zero except in Bone init, which initializes them from bone directions. Masked pose applies a coordinate mask after each update; full pose optimizes all 45 finger coordinates. Tracked init uses the default MANO origin; all other configurations use wrist-centered outputs. Palm warmup includes 300 palm-only warmup updates followed by 2,000 full-joint updates. Zero shape is fixed; regularized shape is fitted with an ℓ_2 penalty, and free shape is fitted without that penalty. Any fitted shape is shared across frames of this episode, not across episodes.

To project this common-frame hand motion back into RGB, we retain the inverse sampled camera transforms together with the source intrinsics. Each frame stores its index, four normalized intrinsic parameters, two source-image dimensions, and the 12 entries of the world-to-camera matrix, yielding a camera array in $\mathbb{R}^{61\times 19}$. The source dimensions allow the projection to follow the same aspect-preserving resize and center crop as the RGB input, as detailed in Appendix B. This shared image transform avoids introducing a deterministic offset when supervision is expressed on the visual token grid.

D.4 Sample association and validation scope

To preserve the correspondence through training, each metadata row associates a video with its hand trajectory, camera sequence, source annotation file, and text description. The description follows the source annotation's selected language variant, falling back to the primary description and then the task name when necessary. These associations are checked when loading hand motion and cached video features, including duplicate identifiers, missing hand-motion entries, tensor shapes, and finite values. When camera conditioning or projective guidance is enabled, training also requires a camera file for every selected sample.

These checks establish input compatibility but do not establish annotation accuracy.

The fitting residual itself measures agreement with the tracked 3D joints and does not certify anatomical validity, since the objective contains no explicit joint-limit constraint. Nor does it measure agreement with the visible hand boundary. This distinction matters because the EgoDex release notes a perspective mismatch between its tracked poses and synthesized RGB views (Hoque et al., 2026). A small 3D fitting error therefore does not establish pixel-level alignment, motivating separate checks of skeleton reconstruction and video–hand spatial consistency.

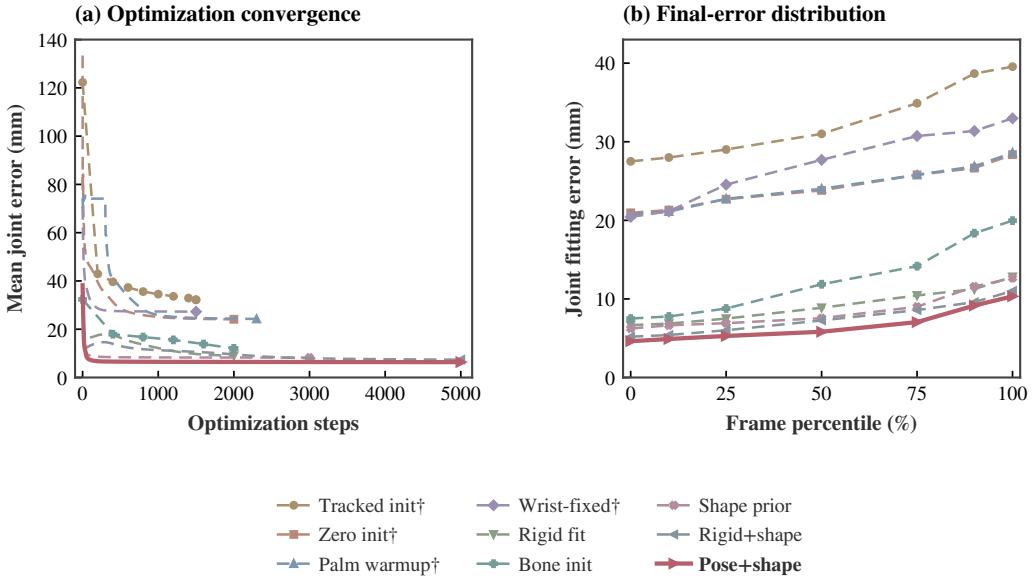


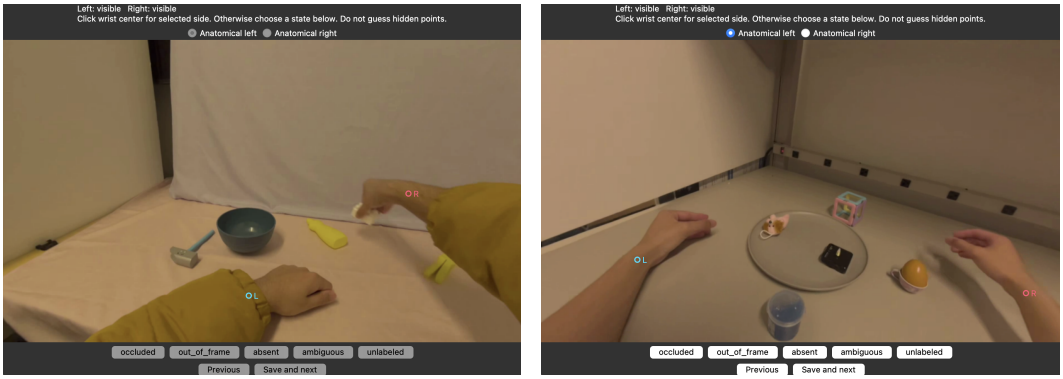
Figure 10: **EgoDex fitting under different initialization and optimization constraints.** Table 6 defines the nine configurations. (a) Mean 21-joint error over all 94 frames of `add_remove_lid/0` (right hand) versus optimization steps. Tracked init and Bone init use sparse original logs and saved final results; the other configurations use complete traces replayed from the original code. (b) Selected empirical quantiles of the final frame-wise mean joint errors. Segments between sparse samples and quantile markers are visual guides, without smoothing. The four configurations marked † retain an incorrect joint correspondence. These curves are configuration diagnostics rather than a controlled ablation.

E Independent RGB consistency evaluation

For cross-modal consistency, a frozen DINOv2-based RGB wrist evaluator detects wrist locations directly from each video frame, without access to Uni-HaVi features, the video hand Locator, MANO parameters, or camera input. The DINOv2 backbone remains frozen, while the wrist heatmap and presence heads are trained separately using projected wrist labels and additional manual annotations. All methods use the same evaluator checkpoint, RGB preprocessing, presence threshold, PCK threshold, and heatmap parameters, with no method-specific calibration. We compare the detected 2D wrist locations with the projected 3D wrists and report wrist alignment error, PCK at 2% of the image diagonal, and wrist-heatmap soft IoU. For video-to-hand-motion evaluation, wrist locations are detected from the observed video and compared with projections of the generated hand trajectory.

Manual audit of wrist detection. We conduct a preliminary manual audit on 100 held-out real images, disjoint from the manual annotations used to train the evaluator. Using the interface in Figure 11, the annotator selects the anatomical hand side, marks the wrist center when visible, and otherwise records it as not visible. Only RGB images are shown during annotation; evaluator predictions and projected hand motion are hidden. Localization error is the mean Euclidean distance between evaluator predictions and manual wrist locations over all manually visible wrists, including predictions below the fixed presence threshold of 0.5. Miss rate is the fraction of these wrists whose predicted presence score falls below this threshold.

To additionally assess evaluator reliability on generated videos, we randomly sample 10 clips from each evaluated model and manually annotate wrist locations using the same protocol. Across these generated clips, the evaluator remains within 10 px of the manual annotations on average. This evaluator error is substantially smaller than the largest cross-modal alignment errors observed in our evaluation.



(a) Clip A (b) Clip B

Figure 11: **Manual wrist annotation on real images.** Two examples from different clips in the annotation interface. Cyan L and pink R mark manually selected anatomical left and right wrists; the controls record visibility when a wrist cannot be localized. No detector predictions or hand-motion projections are shown.