LEARNING POSE-PIXEL SYNERGY: JOINT EGOCENTRIC VIDEO AND 3D HAND POSE GENERATION FROM A SINGLE IMAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting future actions based on the current state from an egocentric (Ego) perspective is vital for applications such as embodied AI. Previous works typically predict sparse hand poses or trajectories, or generate action videos relying on restrictive conditioning inputs. We observe that the two forms of Ego actions (i.e., hand pose and video) exhibit inherent synergy: poses provide abstract high-level structured dynamics, while videos preserve detailed pixel-level visual contexts. Inspired by this, we introduce an Egocentric Pose-Pixel Flow Network (EgoP^{2}FN), a unified flow-matching-based framework to jointly generate future Ego 3D hand poses and videos from a single image. To overcome the challenges of insufficient visual cues and the pose-pixel representation gap, we first develop a Geometry-aware Spatial-Temporal Propagation Module (GSTPM). It leverages geometric cues from the estimated first-frame depth map, which are integrated into visual features via spatial enhancement and temporal modulation, enabling explicit scene layout perception and persistent geometric guidance. Then, we propose a Pose-Pixel Synergy Module (PPSM), which learns pose flow from observation-anchored intermediate features and injects the learned structured pose information back into video latents via adaptive layer normalization, thereby enhancing the consistency between the generated poses and videos. Extensive experiments on our newly curated benchmark demonstrate that our EgoP^{2}FN outperforms many state-of-the-art video generation and hand pose prediction approaches by a large margin. Code will be released.

1 INTRODUCTION

When a person picks up a knife from a cutting board, humans can mentally simulate what will happen from their own perspective: the hand will grasp the knife and cut vegetables. This fundamental yet remarkable capability has inspired growing efforts to enable AI systems to predict fine-grained future actions based on current egocentric (Ego) observations. It not only provides profound insights into decoding human-centric cognitive intentions Kosch et al. (2023); McClelland (2022); Shi et al. (2024; 2026b), but also holds great potential for cutting-edge applications such as decision and execution in embodied AI Duan et al. (2022); Wang et al. (2024); Shi et al. (2025) and humanoid robots Dong et al. (2023); Li et al. (2023); Lee et al. (2024).

Ego actions are typically characterized by hand-object interactions and can be represented in two forms: hand poses and video frames. As a compact and structured representation, hand pose modeling has attracted extensive interest. Compared to earlier works Prakash et al. (2024); Moon et al. (2020); Simon et al. (2017) that aim to estimate current hand states, hand pose prediction takes a step forward by forecasting future motions. For example, OCT Liu et al. (2022) is a pioneering work that predicts 2D hand hotspots. USST Bao et al. (2023) and MADiff Ma et al. (2025) forecast future 3D hand trajectories via uncertainty-aware and diffusion-based strategies. Furthermore, SFhand Liu et al. (2025) and Exo2EgoPose Shi et al. (2026a) use multimodal conditions to predict joint-level 3D hand poses in the future. Despite these impressive advances, the hand pose representations remain sparse and abstract, lacking intuitive visual details and scene contexts. With the development of diffusion- and flow-matching-based frameworks, researchers have increasingly turned to the video generation paradigm. Despite the high-fidelity results achieved by general image/video-conditioned

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.

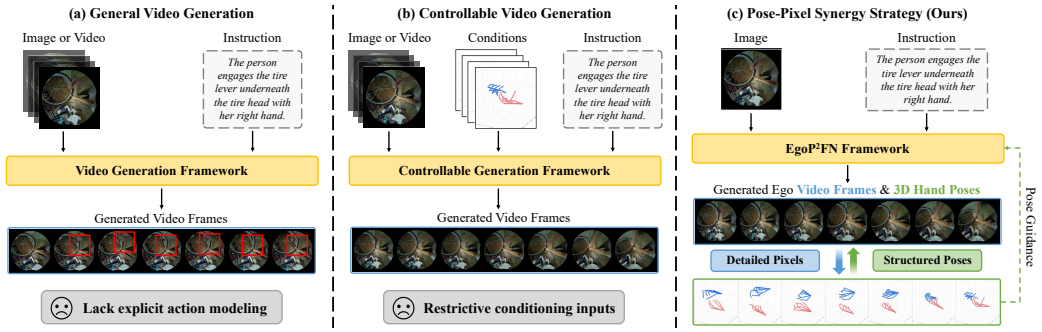

Figure 1: Schematic of general and controllable video generation, and pose-pixel synergy strategy.

video generation models Xing et al. (2024); Zhang et al. (2023); Yang et al. (2025) shown in Fig. 1 (a), these methods struggle to generate plausible Ego hand-object interactions due to the lack of explicit action modeling. In recent years, many controllable methods for Ego video generation have been developed: Ego-X Kang et al. (2026) performs Ego-view generation conditioned on given paired exocentric (Exo) videos. Ego-InBetween Ge et al. (2026) models object state transitions requiring the target frames. In addition, EgoControl Pallotta et al. (2026), HandWorld Sun et al. (2026), and Zhang et al. (2026) predict fine-grained Ego actions conditioned on precise current or future hand pose states, as shown in Fig. 1 (b). Although effective, these methods typically rely on restrictive conditioning inputs (i.e., external viewpoint, future target frames, or proprioceptive signals). For example, pose-conditioned generative methods use ground-truth hand poses as inputs in simulation but may require an additional pose predictor in realistic deployments, whose prediction errors can be further amplified by complex Ego hand joint dynamics.

We observe that Ego hand poses and videos are inherently complementary: poses encode high-level structured joint dynamics, whereas videos provide pixel-level detailed visual contexts. This insight motivates us to learn the pose-pixel synergy (shown in Fig. 1 (c)), which provides a promising way to predict both physically and visually plausible Ego actions. By requiring only a single Ego image with the necessary language instruction, this paradigm avoids relying on restrictive conditions, such as additional visual inputs or externally provided poses in existing methods. Nevertheless, this task remains non-trivial due to two key challenges: On the one hand, complex Ego occlusions and hand-object interactions make it difficult for a single image to infer the 3D spatial layouts and geometric relations between hands and objects for predicting upcoming actions. On the other hand, dense pixels and structured poses present substantially different representation forms, which hinder the transfer and interaction between them, leading to pose-video inconsistency.

To tackle the above problems, we propose a novel Egocentric Pose-Pixel Flow Network (EgoP^{2}FN), a unified flow-matching-based framework that learns pose-pixel synergy by jointly generating future Ego 3D hand poses and videos from a single image. We develop a Geometry-aware Spatial-Temporal Propagation Module (GSTPM) to incorporate geometric cues into the subsequent reasoning process. Specifically, it first uses an off-the-shelf estimator Piccinelli et al. (2025) to obtain the first-frame depth map, which is progressively integrated into visual features through spatial enhancement and temporal modulation, facilitating explicit scene layout perception and persistent geometric guidance. Then, we propose a Pose-Pixel Synergy Module (PPSM), which introduces an auxiliary pose regression head to anchor intermediate features with observation-aware pose information, preventing shortcut denoising from low-noise poses. Moreover, it feeds the structured poses back into video latents through adaptive layer normalization (AdaLN), thereby mitigating the representation gap and enhancing pose-video consistency. For evaluation, we construct a new benchmark from multiple datasets (i.e., *AssemblyHands* Ohkawa et al. (2023), *Ego-Exo4D* Grauman et al. (2024), and *EgoMe-pose* Shi et al. (2026a)) with Ego videos and the corresponding 3D hand poses. Extensive experiments demonstrate that EgoP^{2}FN substantially outperforms many state-of-the-art video generation and hand pose prediction approaches on diverse metrics.

Our major contributions can be summarized as follows:

- We explore the pose-pixel synergy in the Ego view and introduce a novel Egocentric Pose-Pixel Flow Network (EgoP^{2}FN), a unified flow-matching-based framework that jointly generates Ego 3D hand poses and videos of future actions from a single frame.

- We first propose a Geometry-aware Spatial-Temporal Propagation Module (GSTPM), which incorporates geometric cues from the observed frame across spatial and temporal dimensions to facilitate perception and guidance. Then, we propose a Pose-Pixel Synergy Module (PPSM), which learns pose flow from observation-anchored features and feeds the structured poses back into video latents to enhance pose-video consistency.
- We curated a novel benchmark based on *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose* datasets. Extensive experiments show that EgoP^{2}FN achieves superior performance over many state-of-the-art video generation and hand pose prediction methods, demonstrating the effectiveness of jointly learning the pose-pixel synergy.

2 Related Work

2.1 Egocentric Video Generation

With the rapid advances in diffusion-based Ho et al. (2020); Nichol & Dhariwal (2021) and flow-matching-based Lipman et al. (2022) methods for video generation Peebles & Xie (2023); Ali et al. (2025), many works have explored generating action videos from an Ego perspective. EgoGen Li et al. (2024) is the pioneering work on generative Ego data synthesis. Following this work, many studies Huang et al. (2026); Zhang et al. (2026) have focused on controllable Ego generation, which typically incorporates various forms of conditions. Ego-InBetween Ge et al. (2026) models object state transitions conditioned on the initial and target frames. ViHOI Cai et al. (2026) synthesizes human–object interaction sequences with visual and geometric priors. Moreover, many cross-view methods like Exo2Ego-V Liu et al. (2024a), EgoExo-Gen Xu et al. (2025), EgoWorld Park et al. (2025) and EgoX Kang et al. (2026) exploit knowledge from Exo observations for Ego generation. Most recently, some methods introduce structured poses for fine-grained action control. EgoControl Pallotta et al. (2026) and HandWorld Sun et al. (2026) generate future Ego actions conditioned on precise current hand pose states or future whole-body poses. Despite these achievements, they still rely on restrictive conditioning inputs, which may be difficult to obtain in practical scenarios, whereas our method generates future poses and videos from a single image.

2.2 Egocentric Pose Prediction

Human pose naturally provides a structured representation for modeling fine-grained actions. In recent years, Ego pose has gained increasing attention, with numerous datasets Garcia-Hernando et al. (2018); Kwon et al. (2021) being introduced. AssemblyHands Ohkawa et al. (2023) is built upon Assembly101 Sener et al. (2022) with precise 3D hand pose annotations. Ego-Exo4D Grauman et al. (2024) is one of the largest datasets with over 4.2M 3D hand pose annotations. EgoMe-pose Shi et al. (2026a) provides extensive annotations for cross-view daily activities. Some works Wang et al. (2023); Ohkawa et al. (2023) focus on estimating current poses. For instance, WildHands Prakash et al. (2024) estimates 3D hand poses in everyday Ego images, S2DHand Liu et al. (2024b) performs dual-view pose adaptation, and HaWoR Zhang et al. (2025) reconstructs temporally coherent hand motion in world coordinates. Other works Liu et al. (2025; 2022) attempt to forecast future actions. USST Bao et al. (2023) and MADiff Ma et al. (2025) adopt state-space-based and diffusion-based models to predict 3D Ego hand trajectories. Exo2EgoPose Shi et al. (2026a) utilizes Exo videos as guidance for fine-grained pose forecasting with multimodal conditions. Unlike these works that only predict sparse pose representations, our work fully exploits pose-pixel synergy to generate more detailed future actions from a single Ego image.

3 Method

To learn the pose-pixel synergy, we propose a novel EgoP^{2}FN, a unified framework built upon a pre-trained DiT model Ali et al. (2025). As shown in Fig. 2, given the first-frame image I_0 and language instruction l, EgoP^{2}FN jointly generates future Ego 3D hand poses $\hat{\mathbf{P}} = \{\hat{P}_t\}_{t=1}^{T}$ and video frames $\hat{\mathbf{V}} = \{\hat{I}_t\}_{t=1}^{T}$. To address insufficient visual cues, we propose a Geometry-aware Spatial-Temporal Propagation Module (GSTPM), which integrates the geometric cues in spatial and temporal dimensions. Then, we develop a Pose-Pixel Synergy Module (PPSM), which learns pose flow from observation-anchored features and feeds the poses back to enhance pose-video consistency.

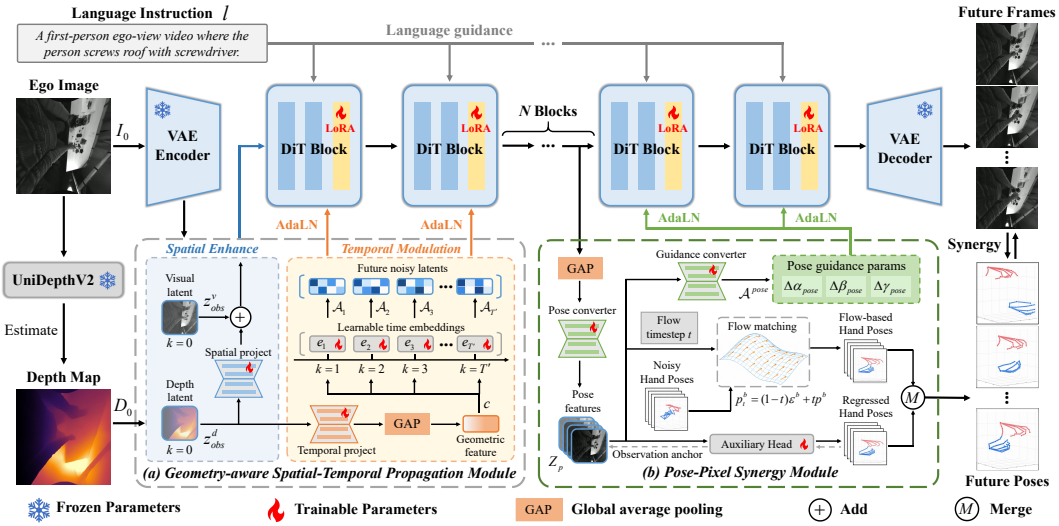

Figure 2: Overview of our EgoP^{2}FN framework. Given the Ego image I_0, we first adopt a frozen VAE encoder to extract the visual latent z^v_{obs}. Then, we estimate the depth map via UniDepthV2 and propose the GSTPM, which progressively integrates the geometric cues via spatial enhancement and temporal modulation. Next, we develop a PPSM, which effectively learns the pose flow based on the observation-anchored features and injects the structured pose information into video latents via AdaLN. Finally, EgoP^{2}FN achieves pose-pixel synergy for consistent video and pose generation.

3.1 Preliminaries

Our EgoP^{2}FN is built upon a flow-matching-based model, which learns a continuous flow between noise and data distributions. Let $\epsilon \sim \mathcal{N}(0, I)$ be the sampled noise distribution, and x_1 denotes a clean latent sampled from the data distribution, flow matching constructs an intermediate state:

$$x_t = (1-t)\epsilon + tx_1, \tag{1}$$

where $t \in [0, 1]$ denotes the flow time, and the time derivative is constant and denoted as the target velocity: $\frac{dx_t}{dt} = x_1 - \epsilon$. Then, a time-conditioned DiT-based network $v_\theta(x_t, t, c)$ is optimized to predict the target velocity field, where c means conditional signals (e.g., language instruction and reference image). The training objective is formulated as:

$$\mathcal{L}_{FM} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\epsilon,x_1,c}\left[\|v_\theta(x_t,t,c) - (x_1-\epsilon)\|_2^2\right]. \tag{2}$$

During inference, the generation starts from random noise ϵ to obtain the clean latent x_1 iteratively by solving the ordinary differential equation: $\frac{dx_t}{dt} = v_\theta(x_t, t, c)$.

3.2 Geometry-aware Spatial-Temporal Propagation Module

Given a single observed Ego image I_0, the model may suffer from insufficient visual cues and struggle to capture spatial layouts and geometric relations. To address this issue, we apply an off-the-shelf model Piccinelli et al. (2025) to estimate its depth map D_0 as geometric cues. Then, we employ a frozen VAE encoder $\mathcal{E}_{VAE}$ to extract the visual latent z^v_{obs} from I_0 and the depth latent z^d_{obs} from the depth map D_0, respectively. Finally, as shown in Fig. 2 (a), we progressively integrate the geometric cues into the visual features via spatial enhancement and temporal modulation.

3.2.1 Spatial Enhancement

The depth map explicitly encodes structural information, serving as a complement to the appearance information in visual observation. Therefore, we perform spatial enhancement on the first-frame visual observation to resolve occlusions and hand-object interactions in Ego scenes, facilitating the perception of spatial layout. In detail, the visual latent z^v_{obs} and depth latent z^d_{obs} are first projected

into a unified token space through the shared patch embedding layer $\mathcal{P}$ to obtain visual tokens $V_{obs} \in \mathbb{R}^{H\times W\times C}$ and geometry tokens $G_{obs} \in \mathbb{R}^{H\times W\times C}$, respectively. Then, we perform spatial projection on G_{obs} and integrate the learned features with the visual tokens V_{obs}, as follows:

$$\Delta V^{spa} = W_{spa}\left(\mathrm{SiLU}\left(W_{in}\mathrm{LN}(G_{obs})\right)\right) \tag{3}$$

$$\widetilde{V}_{obs} = V_{obs} + \Delta V^{spa} \tag{4}$$

where W_{in}, W_{spa} are learnable parameters. By learning the spatial correspondence between visual and geometric tokens, the enhanced visual features $\widetilde{V}_{obs}$ contain explicit geometric cues, enabling more accurate perception of the scene layout for the input observation image.

3.2.2 Temporal Modulation

The geometric information may degrade during successive interactions between the visual and noisy future latents in the DiT blocks. Therefore, we perform temporal modulation to provide persistent geometric guidance for future reasoning and generation. At flow matching timestep t, we denote the noisy latents at different future temporal positions as $\{x_t^{(k)}\}_{k=1}^{T'}$, where T' is the number of noisy latents. Moreover, their corresponding features after patch embedding are denoted as $\{V_t^{(k)}\}_{k=1}^{T'}$.

First, we project and aggregate the geometry tokens G_{obs} into a geometric feature c with global structural information of the observation frame. To enable the modulation to respect the temporal progression of future latent states, we introduce learnable time embeddings $\{e_k\}_{k=1}^{T'}$, where each e_k identifies the temporal latent position in the future sequence. We then integrate the time embeddings $\{e_k\}_{k=1}^{T'}$ with the geometric feature c to construct position-aware geometric features $\{g_k\}_{k=1}^{T'}$:

$$g_k = c \odot (1 + 0.1\tanh(e_k)) \tag{5}$$

where $\odot$ denotes element-wise multiplication. In this manner, the same first-frame geometric prior can be adaptively modulated based on the temporal position of each noisy latent. Next, we convert the position-aware geometric features $\{g_k\}_{k=1}^{T'}$ into temporal modulation parameters $\{\mathcal{A}_k\}_{k=1}^{T'}$:

$$\mathcal{A}_k = W_{ada}g_k = [\Delta\alpha_k^{geo}, \Delta\beta_k^{geo}, \Delta\gamma_k^{geo}] \tag{6}$$

where $\Delta\alpha_k^{geo}$, $\Delta\beta_k^{geo}$, and $\Delta\gamma_k^{geo}$ denote the modulation parameters. Finally, for the l-th DiT block, these temporal modulation parameters are injected into the future visual features $V_t^{(k)}$ via AdaLN:

$$\bar{V}_t^{(k)} = \mathrm{LN}\left(V_t^{(k)}\right) \odot (1 + \Delta\beta_k^{geo}) + \Delta\alpha_k^{geo} \tag{7}$$

$$\widetilde{V}_t^{(k)} = V_t^{(k)} + \Delta\gamma_k^{geo} \odot \mathrm{MLP}\left(\bar{V}_t^{(k)}\right) \tag{8}$$

where $\widetilde{V}_t^{(k)}$ represents the k-th temporally modulated visual feature. In this way, the geometric cues are propagated to future noisy latents, which provide persistent geometric guidance.

3.3 Pose-Pixel Synergy Module

The complementarity between Ego hand poses and video frames motivates us to explore the pose-pixel synergy. However, their substantially different representations (i.e., structured hand dynamics and detailed visual pixels) hinder effective alignment and cause inconsistency. To bridge this gap, we propose a Pose-Pixel Synergy Module (PPSM) as shown in Fig. 2 (b), which learns accurate pose flow from observation-anchored intermediate features and injects the structured pose information back into video latents to enhance consistency between the generated pose and video.

Specifically, we obtain the intermediate feature $F^{l_m} \in \mathbb{R}^{(T'+1)\times H\times W\times C}$ with the visual, depth, and language observation contexts. We spatially aggregate and project the future features via a two-layer Transformer Vaswani et al. (2017) to extract the pose features $Z_p \in \mathbb{R}^{T\times D_p}$ for subsequent pose flow learning. We decompose the absolute Ego 3D hand poses into wrist roots and relative joints. For each hand $h \in \{L, R\}$, we extract the hand wrist as the root $r_f^h = P_f^{h,w}$, and calculate the root-relative joints as $q_f^{h,j} = P_f^{h,j} - r_f^h$, where $P_f^{h,j}$ denotes the absolute coordinate of the j-th joint of hand h at future frame f, and w indicates the wrist joint. Let $b \in \{r, q\}$ denote the root and

relative joint branches, where p^b represents the corresponding clean pose state and ϵ^b denotes the sampled noise. At flow timestep t, the noisy pose state and the target velocity field are as follows:

$$p_t^b = (1-t)\epsilon^b + tp^b \qquad u_t^{b,*} = \frac{\partial p_t^b}{\partial t} = p^b - \epsilon^b \tag{9}$$

Conditioned on the pose features Z_p, the pose flow prediction and optimization are as follows:

$$\hat{u}_t^b = \mathcal{F}_b(p_t^b, t, Z_p) \qquad \mathcal{L}_b = \mathbb{E}\left[\rho\left(\hat{u}_t^b - u_t^{b,*}\right)\right] \qquad \mathcal{L}_{pfm} = \lambda_r \mathcal{L}_r + \lambda_q \mathcal{L}_q \tag{10}$$

where $\mathcal{F}_b(\cdot)$ is the pose flow predictor, $\rho(\cdot)$ is the pose flow loss, λ_r and λ_q are balance coefficients.

However, directly learning from noisy poses may introduce a denoising shortcut. Unlike pixel-level generation, sparse poses can preserve structural priors even under noisy perturbations. Particularly under low-noise conditions, the predictor tends to rely on informative noisy poses for local refinement rather than reasoning from observations. Hence, we introduce an auxiliary head $\mathcal{H}_b(\cdot)$, which regresses poses from Z_p to anchor the features with observation-aware pose information:

$$\hat{p}^{b,aux} = \mathcal{H}_b(Z_p) \qquad \mathcal{L}_b^{aux} = \mathrm{SmoothL1}(\hat{p}^{b,aux}, p^b) \qquad \mathcal{L}_{aux} = \lambda_r^{aux}\mathcal{L}_r^{aux} + \lambda_q^{aux}\mathcal{L}_q^{aux} \tag{11}$$

where λ_r^{aux} and λ_q^{aux} are balance coefficients. This observation-aware supervision encourages Z_p to capture information about hand pose rather than rely on noisy pose states. At each flow step, the prediction $\hat{p}^{b,aux}$ is integrated with the flow-based prediction obtained by $\hat{p}^{b,flow} = p_t^b + (1-t)\hat{u}_t^b$:

$$\hat{p}^b = (1-\eta)\hat{p}^{b,flow} + \eta\hat{p}^{b,aux}. \tag{12}$$

Furthermore, we feed the learned features Z_p back into the video latents to inject the structured hand poses into video generation. Specifically, Z_p is temporally projected into pose guidance parameters:

$$\mathcal{A}^{pose} = W_{pose} Z_p = [\Delta\alpha^{pose}, \Delta\beta^{pose}, \Delta\gamma^{pose}] \tag{13}$$

where W_{pose} denotes the learnable parameter, and $\Delta\alpha^{pose}$, $\Delta\beta^{pose}$, and $\Delta\gamma^{pose}$ denote the shift, scale, and gate terms, respectively. These pose guidance parameters are injected into the later DiT blocks through AdaLN (similar to Eq. 7 and Eq. 8) to modulate the generation process with structured hand dynamics. In this way, the PPSM first learns accurate pose flow from observation-anchored intermediate features and then feeds the structured pose information back into video latents, which effectively mitigates the representation gap and enhances the pose-video consistency.

The proposed EgoP^{2}FN is optimized with the video generation loss $\mathcal{L}_{video}$ and the pose-related objectives, including the pose flow-matching loss $\mathcal{L}_{pfm}$ and auxiliary pose regression loss $\mathcal{L}_{aux}$.

4 Experiments

4.1 Experimental Settings

4.1.1 Dataset Construction

We construct a new benchmark from three Ego datasets: ***AssemblyHands*** Ohkawa et al. (2023) provides high-quality 3D hand pose annotations, from which we correlate Ego clips with textual action descriptions in *Assembly101* Sener et al. (2022) and obtain 719 samples. ***Ego-Exo4D*** Grauman et al. (2024) provides large-scale Ego 3D hand pose annotations, and we align its annotated clips with atomic action descriptions to obtain 14001 samples. ***EgoMe-pose*** Shi et al. (2026a) is based on the *EgoMe* Qiu et al. (2025) dataset with detailed annotations. Since its poses are automatically annotated, we re-split the benchmark and manually verify test samples, resulting in 4596 samples.

Note that we transform Ego 3D hand poses into the camera coordinate system. We split 18916 training samples (644 from *AssemblyHands*, 13776 from *Ego-Exo4D*, and 4496 from *EgoMe-pose*), and 400 test samples, including 75, 225, and 100 samples from the three datasets, respectively.

4.1.2 Evaluation Metrics

To evaluate the video generation performance, we report frame-level PSNR, SSIM, and LPIPS metrics, as well as video-level Fréchet Video Distance (FVD) Unterthiner et al. (2019). Following the VBench-I2V protocol Huang et al. (2025), we further evaluate from five aspects, including I2V Subject Consistency (ISC), I2V Background Consistency (IBC), Subject Consistency (SC), Motion Smoothness (MS), and Aesthetic Quality (AQ). For 3D hand pose generation, we report MPJPE, PA-MPJPE, and MPJVE in millimeters (mm) to evaluate the predicted Ego 3D hand poses.

Table 1: Quantitative comparisons on Ego video generation. * denotes methods that are further fine-tuned. † denotes methods that use the ground-truth Ego hand poses as conditioning inputs.

Methods	*Image Fidelity*			*Video Quality*	*VBench-I2V Metrics*				
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	ISC ↑	IBC ↑	SC ↑	MS ↑	AQ ↑
Cosmos-Predict2.5	11.47	0.332	0.615	782.74	0.888	0.899	0.782	0.961	0.378
Wan2.2-TI2V	12.55	0.476	0.484	524.23	0.936	0.946	0.843	0.980	0.400
CogVideoX-I2V	<u>16.26</u>	**0.579**	**0.384**	258.89	0.939	0.944	0.901	0.978	0.396
AnimateAnyone†	16.22	0.451	0.423	264.57	0.933	0.936	0.903	0.974	0.390
Zhang *et al.*†	13.28	0.416	0.550	224.03	0.921	0.924	0.881	0.983	0.373
Cosmos-Predict2.5*	15.29	0.550	0.418	246.18	0.966	0.953	0.914	0.990	0.407
EgoControl*†	14.57	0.479	0.478	<u>174.60</u>	0.962	**0.968**	0.905	0.990	<u>0.410</u>
EgoP^{2}FN (Ours)	**16.37**	<u>0.575</u>	<u>0.397</u>	**162.21**	**0.980**	<u>0.965</u>	**0.931**	**0.994**	**0.413**

4.1.3 Implementation Details

We adopt the `Cosmos-predict2.5-2B` Ali et al. (2025) as the base framework. Following prior work Pallotta et al. (2026), we generate videos at 480×480 resolution and 16 fps. We fine-tune the pretrained model using LoRA Hu et al. (2022) (rank=512) with a batch size of 3 per GPU. We train the model for 20000 iterations, and the initial learning rate is set to 4e-5. The generation model is conditioned on the first Ego frame and the language instruction to generate the video and 3D hand poses of 45 frames. We select the 16th DiT block for pose-flow prediction. The geometric temporal modulation is applied to the {4,8,12}-th DiT blocks, while the structured pose features are injected into the {18,22,26}-th DiT blocks. The pose feature transformer consists of a linear layer and two Transformer layers with a feature dimension of 512, while the auxiliary regression head uses a feature dimension of 128. For the root and joint branches, the Gaussian noise standard deviations are set to 0.5 and 0.05, with temporal smoothing windows of 15 and 5 frames, respectively. The coefficients λ_r, λ_q, λ_r^{aux}, λ_q^{aux}, and η are set to 2.5, 25.0, 1.0, 1.0, and 0.5, respectively.

4.2 Comparison with State-of-the-Art Methods

For a comprehensive evaluation, we compare our EgoP^{2}FN with state-of-the-art video generation and hand pose prediction methods. For Ego video generation, we compare our EgoP^{2}FN with representative image-to-video (I2V) and Ego video generation methods, i.e., CogVideoX-I2V-5B Yang et al. (2025), Wan2.2-TI2V-5B Wan et al. (2025), AnimateAnyone Hu (2024), Cosmos-Predict2.5-2B Ali et al. (2025), Zhang et al. (2026), and EgoControl Pallotta et al. (2026). For Ego 3D hand pose prediction, we compare with several hand pose prediction methods, i.e., Martinez et al. (2017), USST Bao et al. (2023), HaMeR Pavlakos et al. (2024), EgoH4 Hatano et al. (2025), MADiff Ma et al. (2025), and Exo2EgoPose Shi et al. (2026a).

Table 2: Comparisons on Ego 3D hand pose prediction.

Methods	*Spatial*		*Temporal*
	MPJPE ↓	PA-MPJPE ↓	MPJVE ↓
Random Predict	2043.71	39.87	98.92
Martinez *et al.*	424.04	40.04	42.26
USST	225.95	38.23	59.37
HaMeR	201.54	37.33	55.72
MADiff	<u>179.85</u>	<u>31.54</u>	70.51
EgoH4	184.38	32.84	63.46
Exo2EgoPose	203.30	31.60	<u>26.97</u>
EgoP^{2}FN (Ours)	**139.74**	**21.44**	**26.48**

In Table 1, we compare our EgoP^{2}FN with several general I2V generation methods (upper panel) and Ego video generation methods (bottom panel). For image fidelity, our method achieves the best PSNR, while SSIM and LPIPS metrics are slightly lower than those of CogVideoX-I2V Yang et al. (2025). This is mainly because CogVideoX-I2V tends to preserve the input appearance with relatively limited temporal changes. For video quality, our method achieves the best FVD, indicating that the generated videos exhibit better temporal coherence. For VBench-I2V metrics, EgoP^{2}FN yields competitive scores, indicating its ability to generate high-quality and semantically consistent Ego videos. Notably, although some controllable methods Hu (2024); Zhang et al. (2026); Pallotta et al. (2026) require ground-truth hand poses as additional conditions, EgoP^{2}FN still achieves superior performance without relying on external poses as inputs, validating its effectiveness.

In Table 2, we compare our EgoP^{2}FN with related hand pose prediction methods. Compared with random prediction, all other methods achieve much better results, showing their feasibility for this

Table 3: Ablation studies of key components. "SE" and "TM" denote spatial geometric enhancement and temporal geometric modulation, respectively. "PG" denotes pose guidance via AdaLN, "PF" denotes pose flow learning, and "V" denotes the plain pose flow without auxiliary supervision.

SE	TM	PG	PF	*Visual Metrics*				*Pose Metrics*		
				PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	MPJPE ↓	PA-MPJPE ↓	MPJVE ↓
✓	✓	✓	✓	**16.37**	**0.575**	**0.397**	**162.21**	**139.74**	**21.44**	<u>26.48</u>
✓		✓	✓	<u>16.16</u>	<u>0.569</u>	<u>0.404</u>	<u>163.02</u>	168.81	<u>21.65</u>	28.61
	✓	✓	✓	16.01	0.562	0.413	165.57	<u>156.63</u>	21.82	**25.88**
		✓	✓	15.83	0.559	0.422	197.85	196.19	25.02	27.52
			✓	15.67	0.556	0.409	248.36	239.41	31.79	32.92
			V	15.70	0.553	0.406	230.21	269.08	32.93	35.37
				15.29	0.550	0.418	246.18	-	-	-

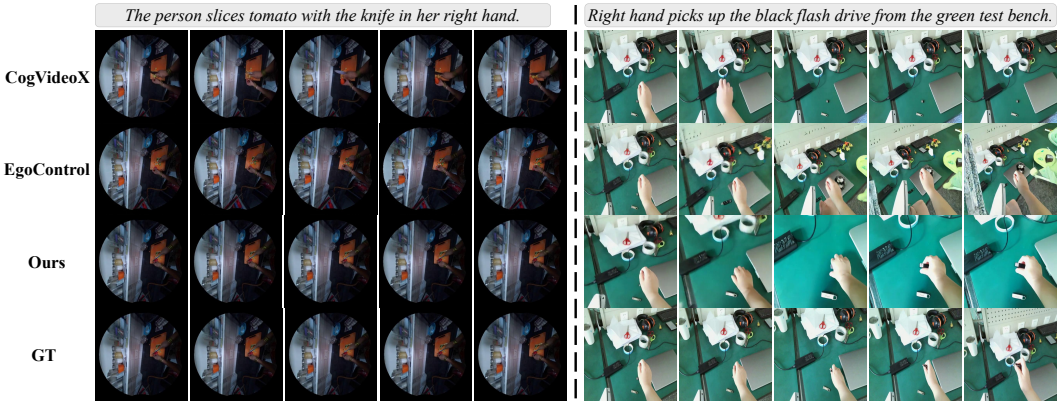

Figure 3: Results of the generated future Ego frames of our EgoP^{2}FN and comparison methods.

task. Compared with related pose prediction approaches, EgoP^{2}FN achieves substantially higher performance in both spatial accuracy and temporal dynamics. In detail, our method significantly reduces MPJPE and PA-MPJPE errors, showing its ability to precisely predict both hand and joint positions. Meanwhile, our EgoP^{2}FN also achieves the lowest MPJVE, indicating that the model can accurately model temporally coherent hand pose dynamics. The above results further demonstrate the effectiveness of EgoP^{2}FN, which learns synergy between structured poses and detailed pixels to jointly generate visually and physically plausible Ego videos and 3D hand poses.

4.3 Ablation Study

In Table 3, we conduct ablation studies. We first remove temporal geometric modulation and spatial geometric enhancement, respectively. Both variants exhibit slight performance degradation, showing the effectiveness of the two geometric guidance branches. Furthermore, removing the whole GSTPM leads to a more substantial performance drop compared with removing either spatial or temporal geometric guidance alone, indicating their complementary roles in future Ego action modeling. In the fifth row, we remove the pose guidance, resulting in a performance decline in both visual and pose metrics, indicating the importance of pose-pixel interaction for consistent video and pose generation. In the sixth row, removing auxiliary supervision from pose flow matching results in inferior pose prediction, showing that observation-aware pose information effectively prevents shortcut denoising. Finally, removing the pose flow branch leads to further degradation, highlighting the effectiveness of learning pose-pixel synergy for joint Ego video and hand pose generation.

4.4 In-depth Analysis

4.4.1 Analysis of Ego Video Generation

In Fig. 3, we show two examples of the generated Ego video frames. In the first row, CogVideoX Yang et al. (2025) generates highly inaccurate knife-grasping and tomato-slicing actions in the left example, while largely preserving the appearance with limited motion in the right example. In the

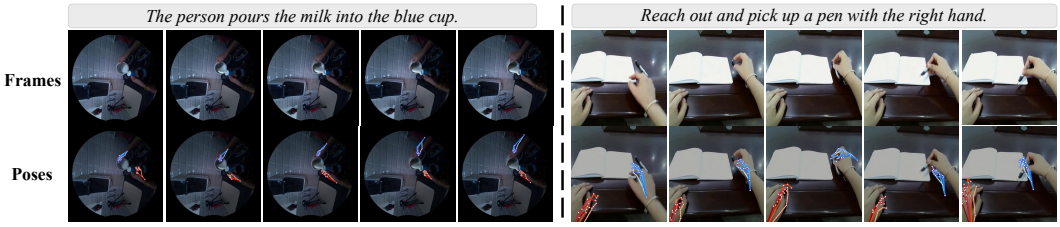

Figure 5: Visualizations of the generated future Ego frames with the corresponding hand poses.

second row, although EgoControl Pallotta et al. (2026) utilizes ground-truth poses as conditions, it still exhibits inaccurate hand-object interactions and scene distortions. In contrast, our EgoP^{2}FN faithfully captures the intended actions of "slicing a tomato with a knife" and "picking up a subtle flash drive" and generates visually plausible future Ego frames from a single image. The results show the effectiveness of introducing hand pose prediction to guide future Ego video generation.

4.4.2 Analysis of Ego 3D Hand Pose Prediction

In Fig. 4, we visualize an example of the predicted future Ego 3D hand poses of the proposed EgoP^{2}FN and two hand pose prediction methods Bao et al. (2023); Shi et al. (2026a). Specifically, we project the 3D hand joints into 2D frames and perform alignment for better visualization. In this example, the subject is going to "hold the toy car with the right hand and place the toy car screw on the table with the left hand". With limited visual cues from a single conditioning frame, the comparison methods tend to produce coarse hand structures with severe errors in hand orientations and joint positions. In contrast, our method exploits geometric cues and explicitly anchors the observation-aware pose information, which facilitates generating more accurate and physically plausible hand joints.

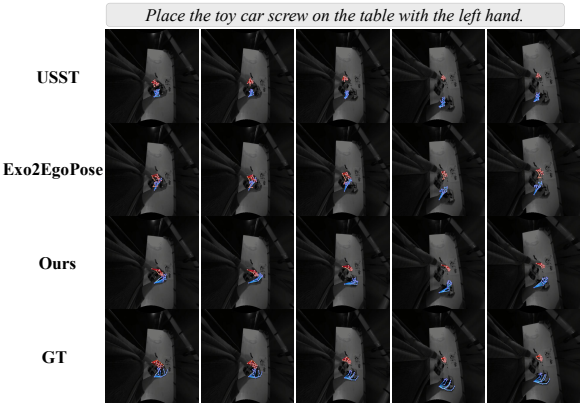

Figure 4: Visualizations of Ego hand pose prediction.

4.4.3 Analysis of Pose-Pixel Synergy

In Fig. 5, we visualize two examples of future Ego video frames with the corresponding 3D hand poses generated by our EgoP^{2}FN. The first row shows the generated video frames, while the second row presents the corresponding projected hand pose predictions. The results show that the generated video frames and hand poses exhibit consistent actions, where the predicted hand joints closely correspond to the hand-object interactions in the generated frames. These results demonstrate that our method learns pose-pixel synergy and enables coherent generation of Ego video and hand pose.

5 Conclusion

In this paper, we investigate the pose-pixel synergy in Ego view and propose a novel Egocentric Pose-Pixel Flow Network (EgoP^{2}FN), a unified flow-matching-based framework that jointly generates future 3D hand poses and video frames from a single image. To address the challenges of insufficient visual cues and the pose-pixel representation gap, we first propose a Geometry-aware Spatial-Temporal Propagation Module (GSTPM), which extracts geometric cues for spatial enhancement and temporal modulation to facilitate scene layout perception and persistent geometric guidance. Then, a Pose-Pixel Synergy Module (PPSM) is developed to learn pose flow from observation-anchored features and feed structured poses back into video latents to improve pose-video consistency. Extensive experiments on our newly curated benchmark demonstrate that our method surpasses many state-of-the-art video generation and hand pose prediction approaches.

AI USE STATEMENT

We used generative AI tools to assist with language polishing and improve the readability of this manuscript. All AI-assisted edits were carefully reviewed by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

We provide implementation details and evaluation protocols in the paper to facilitate reproducibility. We will publicly release the source code of our EgoP^{2}FN to support reproducibility and future research.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Wentao Bao, Lele Chen, Libing Zeng, Zhong Li, Yi Xu, Junsong Yuan, and Yu Kong. Uncertainty-aware state space transformer for egocentric 3d hand trajectory forecasting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 13702–13711, 2023.

Songjin Cai, Linjie Zhong, Ling Guo, and Changxing Ding. Vihoi: Human-object interaction synthesis with visual priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30686–30695, 2026.

Jiale Dong, Weiyong Si, and Chenguang Yang. A novel human-robot skill transfer method for contact-rich manipulation task. *Robotic Intelligence and Automation*, 43(3):327–337, 2023.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 6(2):230–244, 2022.

Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023.

Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 409–419, 2018.

Mengmeng Ge, Takashi Isobe, Xu Jia, Yanan Sun, Zetong Yang, Weinong Wang, Dong Zhou, Dong Li, Huchuan Lu, and Emad Barsoum. Ego-inbetween: Generating object state transitions in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 43145–43154, 2026.

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

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Masashi Hatano, Zhifan Zhu, Hideo Saito, and Dima Damen. The invisible egohand: 3d hand forecasting through egobody pose estimation. *arXiv preprint arXiv:2504.08654*, 2025.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8153–8163, 2024.

Xuan Huang, Mochu Xiang, Zhelun Shen, Jinbo Wu, Chenming Wu, Chen Zhao, Kaisiyuan Wang, Hang Zhou, Shanshan Liu, Haocheng Feng, et al. Genhoi: Towards object-consistent hand-object interaction with temporally balanced and spatially selective object injection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23117–23127, 2026.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11116–11126, 2026.

Thomas Kosch, Jakob Karolus, Johannes Zagermann, Harald Reiterer, Albrecht Schmidt, and Paweł W Woźniak. A survey on measuring cognitive workload in human-computer interaction. *ACM Computing Surveys*, 55(13s):1–39, 2023.

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*, pp. 10138–10148, 2021.

Regina Kyung-Jin Lee, Hao Zheng, and Yuqian Lu. Human-robot shared assembly taxonomy: A step toward seamless human-robot knowledge transfer. *Robotics and Computer-Integrated Manufacturing*, 86:102686, 2024.

Gen Li, Kaifeng Zhao, Siwei Zhang, Xiaozhong Lyu, Mihai Dusmanu, Yan Zhang, Marc Pollefeys, and Siyu Tang. Egogen: An egocentric synthetic data generator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14497–14509, 2024.

Jiehao Li, Junzheng Wang, Shoukun Wang, and Chenguang Yang. Human–robot skill transmission for mobile robot via learning by demonstration. *Neural Computing and Applications*, 35(32): 23441–23451, 2023.

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

Jia-Wei Liu, Weijia Mao, Zhongcong Xu, Jussi Keppo, and Mike Z Shou. Exocentric-to-egocentric video generation. *Advances in Neural Information Processing Systems*, 37:136149–136172, 2024a.

Ruicong Liu, Takehiko Ohkawa, Mingfang Zhang, and Yoichi Sato. Single-to-dual-view adaptation for egocentric 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 677–686, 2024b.

Ruicong Liu, Yifei Huang, Liangyang Ouyang, Caixin Kang, and Yoichi Sato. Sfhand: A streaming framework for language-guided 3d hand forecasting and embodied manipulation. *arXiv preprint arXiv:2511.18127*, 2025.

Shaowei Liu, Subarna Tripathi, Somdeb Majumdar, and Xiaolong Wang. Joint hand motion and interaction hotspots prediction from egocentric videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3282–3292, 2022.

Junyi Ma, Xieyuanli Chen, Wentao Bao, Jingyi Xu, and Hesheng Wang. Madiff: Motion-aware mamba diffusion models for hand trajectory prediction on egocentric videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *2017 IEEE Conference on computer vision and pattern recognition (CVPR)*, pp. 4674–4683. IEEE, 2017.

James L McClelland. Capturing advanced human cognitive abilities with deep neural networks. *Trends in Cognitive Sciences*, 26(12):1047–1050, 2022.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pp. 548–564. Springer, 2020.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pp. 8162–8171. PMLR, 2021.

Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12999–13008, 2023.

Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. Egocontrol: Controllable egocentric video generation via 3d full-body poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4269–4279, 2026.

Junho Park, Andrew Sangwoo Ye, and Taein Kwon. Egoworld: Translating exocentric view to egocentric view using rich exocentric observations. *arXiv preprint arXiv:2506.17896*, 2025.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Aditya Prakash, Ruisen Tu, Matthew Chang, and Saurabh Gupta. 3d hand pose estimation in everyday egocentric images. In *European Conference on Computer Vision*, pp. 183–202. Springer, 2024.

Heqian Qiu, Zhaofeng Shi, Lanxiao Wang, Huiyu Xiong, Xiang Li, and Hongliang Li. Egome: Follow me via egocentric view in real world. *arXiv preprint arXiv:2501.19061*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21096–21106, 2022.

Zhaofeng Shi, Heqian Qiu, Lanxiao Wang, Fanman Meng, Qingbo Wu, and Hongliang Li. Cognition transferring and decoupling for text-supervised egocentric semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4144–4157, 2024.

Zhaofeng Shi, Heqian Qiu, Lanxiao Wang, Qingbo Wu, Fanman Meng, and Hongliang Li. Unsupervised ego-and exo-centric dense procedural activity captioning via gaze consensus adaptation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 3731–3740, 2025.

Zhaofeng Shi, Heqian Qiu, Lanxiao Wang, Xiang Li, and Hongliang Li. Exo2egopose: Leveraging exocentric demonstrations for vision-language guided egocentric 3d hand pose forecasting. *arXiv preprint arXiv:2607.15890*, 2026a.

Zhaofeng Shi, Heqian Qiu, Lanxiao Wang, Qingbo Wu, Fanman Meng, Lili Pan, and Hongliang Li. Test-time ego-exo-centric adaptation for action anticipation via multi-label prototype growing and dual-clue consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16988–16999, 2026b.

Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1145–1153, 2017.

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, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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.

Jian Wang, Diogo Luvizon, Weipeng Xu, Lingjie Liu, Kripasindhu Sarkar, and Christian Theobalt. Scene-aware egocentric 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13031–13040, 2023.

Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19757–19767, 2024.

Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Wangbo Yu, Hanyuan Liu, Gongye Liu, Xintao Wang, Ying Shan, and Tien-Tsin Wong. Dynamicrafter: Animating open-domain images with video diffusion priors. In *European Conference on Computer Vision*, pp. 399–417. Springer, 2024.

Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 76961–76978, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/bf78fc727cf882df66e6dbc826161e86-Paper-Conference.pdf.

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. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025.

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. *arXiv preprint arXiv:2603.11755*, 2026.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 1805–1815, 2025.

Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

A Details of Comparison Methods

In this section, we introduce some technical and implementation details of the state-of-the-art comparison methods, i.e., CogVideoX-I2V-5B Yang et al. (2025), Wan2.2-TI2V-5B Wan et al. (2025), AnimateAnyone Hu (2024), Cosmos-Predict2.5-2B Ali et al. (2025), Zhang *et al.*'s method Zhang et al. (2026), and EgoControl Pallotta et al. (2026) for video generation, while Martinez *et al.*'s method Martinez et al. (2017), USST Bao et al. (2023), HaMeR Pavlakos et al. (2024), EgoH4 Hatano et al. (2025), MADiff Ma et al. (2025), and Exo2EgoPose Shi et al. (2026a) for Ego 3D hand pose prediction.

A.1 Video Generation Methods

A.1.1 CogVideoX-I2V

CogVideoX-I2V Yang et al. (2025) is a representative image-to-video generation method with textual inputs based on a video Diffusion Transformer (DiT). It employs a 3D causal VAE to compress videos into spatiotemporal latents and an expert Transformer to facilitate cross-modal fusion. Specifically, the reference image is encoded into visual latents, which are concatenated with the noisy video latents along the channel dimension. Meanwhile, the language instruction is encoded by a T5 text encoder Raffel et al. (2020) to provide semantic guidance during generation. In this paper, we evaluate CogVideoX-I2V under our experimental setting. For each sample, the model takes the first Ego frame and the corresponding language instruction as inputs. The model generates 49 frames at a resolution of 720×480, and we retain the first 45 consecutive frames to match our evaluation setting. Finally, the generated video frames are resized to 480×480 for evaluation.

A.1.2 Wan2.2-TI2V

Wan2.2-TI2V Wan et al. (2025) is a unified framework for text-to-video (T2V) and image-to-video (I2V) generation based on DiT. It employs a highly compressed video VAE and a Transformer-based denoising architecture to generate high-quality videos. In detail, the language instruction is encoded into text tokens and injected into the DiT through cross-attention, while the reference image is incorporated as visual conditions for the video generation process. In this paper, we evaluate Wan2.2-TI2V-5B under our experimental setting. For each sample, the model takes the first Ego frame and the corresponding language instruction as inputs for future video generation. The model generates 45 frames at its default resolution of 1280×704, and the generated results are resized to 480×480 to align with our evaluation protocol.

A.1.3 AnimateAnyone

AnimateAnyone Hu (2024) is a pose-guided image-to-video generation method based on the U-Net architecture that aims to synthesize consistent animations from an image and a target pose sequence. It employs a ReferenceNet to extract appearance features from the reference image and a Pose Guider to encode the pose sequence, while a temporal layer is introduced to maintain the temporal consistency of generated frames. In this paper, we evaluate AnimateAnyone under our setting. For each sample, the model takes the first Ego frame as the reference image and the future ground-truth 2D hand pose sequence as pose conditions to generate future Ego videos. Specifically, we project the future 3D hand poses into 2D keypoints as the input pose conditions following the original paradigm. The model generates future frames at its default resolution of 512×512, and the generated results are combined with the first frame and resized to 480×480 for evaluation.

A.1.4 Cosmos-Predict2.5

Cosmos-Predict2.5 Ali et al. (2025) is a flow-matching-based world foundation model designed for physical AI applications. It uses a latent DiT architecture with a video VAE to generate videos. Specifically, the input video is first compressed into spatiotemporal latents by the video VAE. The denoising process is guided by image conditions, and language instructions are injected into the DiT through cross-attention. In this paper, we evaluate Cosmos-Predict2.5 under our setting with both pretrained and fine-tuned versions. For the pretrained version, we directly use the original Cosmos-Predict2.5 and provide the first Ego frame and language instruction as inputs without additional

pose or depth conditions. The model generates 45 frames at its default resolution of 640×480. In addition, we also fine-tune the model under our Ego setting. We also adapt the model to our Ego action video generation task by fine-tuning LoRA Hu et al. (2022) parameters, and the learning rate after warmup is set to 1e-5. The fine-tuned model generates videos at a resolution of 480×480. Both versions are evaluated under the same protocol, where the video resolution is set to 480×480.

A.1.5 ZHANG *et al.*'S METHOD

Zhang et al. (2026) propose a controllable Ego video generation method that introduces hand pose conditions for fine-grained action control. It is built upon a pretrained video generation model and incorporates an additional pose control branch to inject structured hand motion information into the video generation process. Specifically, the 3D hand joints are encoded as pose conditions and fused with visual features to guide the generation of future Ego videos with consistent hand-object interactions. In this paper, we evaluate this method on our benchmark. We load the released control weights based on the Wan2.1-I2V backbone. For each sample, the model takes the first Ego frame, the corresponding language instruction, and the future 3D hand pose sequence as inputs. Following the official setting, the future ground-truth hand poses are provided as additional conditions. The model generates 45 frames at its default resolution of 832×480, and the generated videos are resized to 480×480 for evaluation.

A.1.6 EGOCONTROL

EgoControl Pallotta et al. (2026) is a controllable Ego video generation method that leverages 3D full-body poses to guide future Ego action generation. It is built upon a pretrained video generation model and introduces pose-aware conditioning mechanisms to inject structured motion information through cross-attention and feature modulation. Since the original method is designed for full-body pose control, directly applying it to 3D hand poses suffers from a substantial condition signal gap. Therefore, in this paper, we fine-tune the EgoControl model with LoRA parameters and a learnable hand pose adaptation module to bridge the gap between full-body and hand pose conditions, using a learning rate of 2e-5. For each sample, the model takes the first Ego frame, the corresponding language instruction, and future Ego 3D hand pose sequences as inputs. Following the original setting, the ground-truth future hand poses are provided as additional conditions rather than predicted by the model. Finally, the generated results are resized to 480×480 for evaluation.

A.2 HAND POSE PREDICTION METHODS

A.2.1 MARTINEZ *et al.*'S METHOD

Martinez et al. (2017) is a pioneering work for human motion prediction, which forecasts future human poses from observed motion sequences. It adopts a simple yet effective residual-based recurrent framework to predict the future motion as residual velocity rather than absolute pose coordinates. In this paper, we adapt Martinez *et al.*'s method to our future Ego 3D hand pose prediction setting. Since the original method relies on human pose sequences, we replace the poses with the first Ego frame and the corresponding language instruction. The model retains the residual GRU prediction paradigm and predicts future 3D hand joints in an autoregressive manner. The hidden dimension is set to 192 for the GRU-based network, and the model is trained for 10 epochs using the Smooth L1 pose loss with a learning rate of 1e-5.

A.2.2 USST

USST Bao et al. (2023) is a representative work for Ego 3D hand trajectory forecasting, which predicts future hand trajectories from Ego frames and historical hand motions. It introduces a state space Transformer with an uncertainty-aware prediction strategy, consisting of a state transition module and an emission module for future trajectory forecasting. In this paper, we adapt USST to our future Ego 3D hand pose prediction setting. We remove the trajectory inputs and take the first Ego frame and the corresponding language instruction as conditions for the model. We modify the forecasting head to predict future 3D hand joints while retaining the uncertainty estimation and uncertainty-aware loss designs. The hidden dimension is set to 192, and the model is trained for 10 epochs using the Smooth L1 pose loss and uncertainty-aware loss with a learning rate of 1e-5.

A.2.3 HAMER

HaMeR Pavlakos et al. (2024) aims to reconstruct 3D hand joints from RGB images. It adopts an architecture based on a Vision Transformer (ViT) Dosovitskiy (2020) to extract image features and a decoder to predict hand poses. Specifically, the model employs a query-based module to aggregate visual features and regress hand poses. In this paper, we train HaMeR in our setting. We replace the original hand reconstruction pipeline with a future pose prediction framework conditioned on the first Ego frame and the language instruction. The model retains the ViT-based visual encoding and Transformer decoder paradigm, whose hidden dimension is set to 256. The model is trained for 10 epochs using the Smooth L1 pose loss with a learning rate of 1e-5.

A.2.4 EGOH4

EgoH4 Hatano et al. (2025) is a representative work for Ego 3D pose forecasting, which predicts future human motions from historical Ego observations. It leverages body priors and a Transformer-based diffusion framework to model future hand dynamics, especially for cases where hands move out of view. Specifically, the model jointly predicts body and hand motions and incorporates visibility and reprojection constraints to improve the accuracy of future hand estimation. In this paper, we adapt EgoH4 to our future Ego 3D hand pose prediction setting. The model retains the Transformer diffusion prediction paradigm and predicts future 3D hand joints conditioned on the observed Ego frame and language instruction. The hidden dimension is set to 256, and the model is trained for 10 epochs using the Smooth L1 pose loss and diffusion noise loss with a learning rate of 1e-4.

A.2.5 MADIFF

MADiff Ma et al. (2025) aims to forecast future Ego hand trajectories from the historical Ego frames and hand states. MADiff is based on a motion-aware Mamba Gu & Dao (2023) diffusion framework, which leverages visual-language features and motion information to guide the diffusion process. Meanwhile, a Mamba-based denoising network captures long-range motion dynamics. In this paper, we train the MADiff network to adapt to our Ego 3D hand pose prediction task. The model retains the state-space diffusion prediction paradigm and predicts future 3D hand joints based on visual and textual conditioning inputs. The hidden dimension is set to 256, and the model is trained for 10 epochs using the Smooth L1 pose loss and diffusion noise loss with a learning rate of 1e-4.

A.2.6 EXO2EGOPOSE

Exo2EgoPose Shi et al. (2026a) is a recent work for the Vision-Language guided Ego 3D Hand Pose Forecasting (VL-EHPF) task, which aims to predict future Ego 3D hand poses based on the current frame, language instruction, and pose state inputs. In detail, Exo2EgoPose feeds the encoded multi-modal tokens into a multimodal causal Transformer architecture for action reasoning with auxiliary tasks such as Ego and Exo reconstruction. Then, the decoded tokens are refined in a global-to-local manner for accurate 3D hand pose prediction. In this paper, we retrain the Exo2EgoPose framework to adapt to our task. The model is conditioned on a single Ego frame and the corresponding language instruction. We preserve its multimodal tokenization and causal Transformer reasoning framework while modifying the output head to predict future 3D hand joints. The hidden dimension is set to 384, and the model is trained for 10 epochs with a learning rate of 5e-5.

B DETAILS OF BENCHMARK CONSTRUCTION

In this section, we present the detailed step-by-step dataset construction and pre-processing pipelines for *AssemblyHands* Ohkawa et al. (2023), *Ego-Exo4D* Grauman et al. (2024), and *EgoMe-pose* Qiu et al. (2025); Shi et al. (2026a) datasets, respectively.

B.1 ASSEMBLYHANDS DATASET

The *AssemblyHands* dataset Ohkawa et al. (2023) provides Ego videos with accurate 3D hand pose annotations collected from toy assembling activities. To make this dataset suitable for our Ego action prediction task, we perform several preprocessing steps to construct training samples, as illustrated in Fig. 6. The detailed pipeline is described as follows:



Figure 6: The data construction pipeline for the *AssemblyHands* dataset.

(1) Filter unreliable samples: Some frames in the original *AssemblyHands* dataset contain incomplete hand joint annotations. We remove unreliable samples by discarding frames with insufficient valid hand joints to ensure reliable pose supervision.

(2) Transform pose coordinates: The original Ego 3D hand poses are labeled in the world coordinate system. We transform each hand joint into the Ego camera coordinate system using the provided camera extrinsic parameters, including the rotation matrix R and translation vector T.

(3) Align action descriptions: Since the *AssemblyHands* dataset does not directly provide language instruction annotations, we utilize the fine-grained action descriptions from *Assembly101* Sener et al. (2022). We associate Ego video frames, 3D hand poses, and textual descriptions based on unified frame indices between the two datasets.

(4) Construct temporal clips: We organize the synchronized Ego videos and hand pose sequences into temporal clips for future Ego action modeling. Each clip contains a single observation frame and the corresponding future frames with aligned 3D hand poses. To preserve complete action executions, we discard clips shorter than 45 frames or longer than 90 frames. For the remaining clips, we select the first 45 frames as the final training samples.

(5) Generate training samples: Finally, we convert all processed clips into a unified format that contains Ego videos, language instructions, and 3D hand pose annotations for the following generative model training.

B.2 EGO-EXO4D DATASET

Ego-Exo4D Grauman et al. (2024) is one of the largest Ego datasets, which contains various daily activities. It contains diverse annotations, such as manually and automatically labeled Ego 3D hand poses and atomic action descriptions. The detailed dataset construction pipeline is illustrated in Fig. 7 and described as follows:

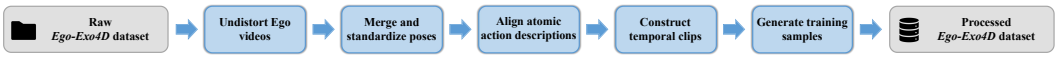

Figure 7: The data construction pipeline for the *Ego-Exo4D* dataset.

(1) Undistort Ego videos: The original *Ego-Exo4D* Ego videos are captured by the Aria glasses platform Engel et al. (2023) with camera distortion. Therefore, we leverage the officially released camera calibration parameters for Ego video undistortion to obtain geometrically corrected video frames.

(2) Merge and standardize poses: Pose annotations in *Ego-Exo4D* include both automatic and manually annotated Ego 3D hand poses. We merge these pose annotations according to the unified video UID and frame indices, where manually annotated poses are prioritized when available. Then, we convert the hand poses into the universal joint representation form Moon et al. (2020). Moreover, we also filter out unreliable annotations where either hand has fewer than 10 valid joints or an invalid root joint.

(3) Align atomic action descriptions: Unlike datasets with predefined temporal action segments, the atomic action descriptions in *Ego-Exo4D* are provided as individual timestamps. We associate each textual description with the corresponding Ego videos and 3D hand poses according to the annotated timestamps and frame index ranges.

(4) Construct temporal clips: Based on each atomic action timestamp, we employ a fixed-length temporal window to select 45 consecutive Ego frames starting from 0.5 seconds before the annotated timestamp. The same frame indices are used to retrieve RGB frames and 3D hand poses, ensuring temporal alignment.

(5) Generate training samples: Finally, we convert all processed clips into a unified format containing Ego video clips, language instructions, and 3D hand pose annotations for subsequent training of generation models.

B.3 EGOME-POSE DATASET

EgoMe-pose Shi et al. (2026a) is a recent dataset built upon *EgoMe* Qiu et al. (2025). It contains paired videos from the Exo observer and the Ego follower with diverse annotations, such as eye gaze, timestamps, and language descriptions. In this paper, we utilize the Ego-view videos and their annotations. The detailed data preprocessing pipeline is illustrated in Fig. 8, and described as follows:



Figure 8: The data construction pipeline for the *EgoMe-pose* dataset.

(1) Filter unreliable samples: Since the *EgoMe* dataset is built upon real-world imitation activities, some samples contain incorrect action executions or extremely short durations. We filter out these unreliable samples and retain valid Ego video clips with meaningful actions.

(2) Standardize poses: The Ego 3D hand poses in *EgoMe-pose* are automatically annotated in the camera coordinate system. We convert the original hand poses into the unified joint representation and remove unreliable samples where either hand has fewer than 10 valid joints or has an invalid root joint, ensuring reliable pose supervision.

(3) Align action descriptions: According to the officially provided fine-level action descriptions and timestamp annotations, we align Ego video frames, 3D hand poses, and language instructions based on the annotated time boundaries to ensure temporal consistency of the multimodal data.

(4) Construct temporal clips: Based on the aligned annotations, we first filter video clips according to their durations. Specifically, Ego video clips with insufficient frames (below 45 frames) for generating the required sequence or excessively long durations (beyond 90 frames) are removed. For the remaining video clips, we organize Ego videos and hand poses for future Ego action modeling. Each clip contains a single observation frame and the corresponding future frames with aligned 3D hand poses and the corresponding language instruction, resulting in 45 frames in total.

(5) Verify test samples: Since hand poses in *EgoMe-pose* are automatically annotated, we further manually verify the test samples to ensure reliable evaluation. Specifically, we develop a human evaluation interface to validate the Ego videos and corresponding 3D hand poses, and select samples with accurate hand pose annotations as the final test samples. An example of the evaluation interface is shown in Fig. 9.

C MORE EXPERIMENTAL RESULTS

In this section, we present more quantitative experimental results to further evaluate the effectiveness of the proposed method. In detail, we first analyze the model complexity of the key components in the EgoP^{2}FN framework to report its computational overhead. Then, we show the per-dataset Ego video generation and 3D hand pose prediction performance on the *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose* subsets, respectively.

C.1 ANALYSIS OF MODEL COMPLEXITY

In Table 4, we conduct experiments to evaluate the computational complexity and parameter overhead of the baseline framework and some components of our EgoP^{2}FN method. By removing all newly introduced parts, the baseline framework contains 2059.18 M parameters and requires 63772.96 GFLOPs during inference. In contrast, the new components only introduce marginal parameter and computational overhead. Specifically, the depth estimation model (i.e., UniDepthV2 Piccinelli et al. (2025)) introduces 353.83 M parameters and 1181.33 GFLOPs, while the depth VAE encoding only requires additional inference computation without introducing extra trainable

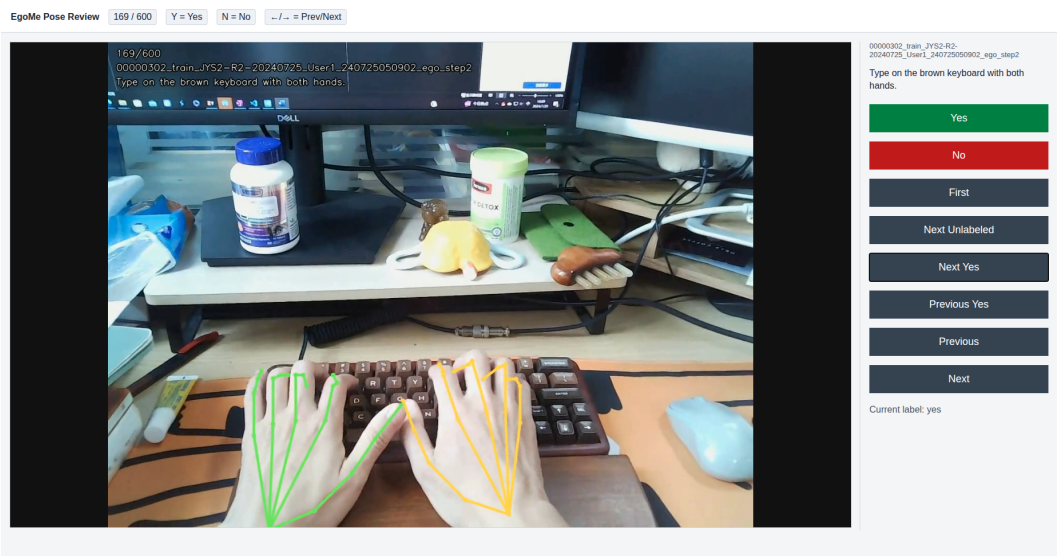

Figure 9: The human evaluation interface for manually verifying the accuracy of *EgoMe-pose* hand pose annotations on test samples.

Table 4: The number of parameters and inference FLOPs of the baseline framework and key components of our EgoP^{2}FN method.

Component	#Params (M)	FLOPs (G)
Baseline framework	2059.18	63772.96
Depth estimation	353.83	1181.33
Depth VAE encoding	-	1553.65
GSTPM	2.625	2.456
PPSM	13.125	1.137

parameters since the model shares the same VAE encoder with the input Ego frame. Moreover, the proposed Geometry-aware Spatial-Temporal Propagation Module (GSTPM) and Pose-Pixel Synergy Module (PPSM) introduce only 2.625 M and 13.125 M parameters, with 2.456 and 1.137 GFLOPs, respectively. Considering the remarkable performance improvements achieved by our EgoP^{2}FN, the additional parameters and computational overhead are tolerable.

C.2 Per-dataset Ego Video Generation

To further evaluate the proposed EgoP^{2}FN method for Ego video generation, we conduct per-dataset experiments on three subsets of our benchmark test set, including *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose*. The detailed results are shown in Table 5, Table 6, and Table 7, respectively.

In Table 5 and Table 7, we compare our method with baselines and show the quantitative results on the *AssemblyHands* and *EgoMe-pose* datasets, respectively. The results show that our EgoP^{2}FN achieves competitive image fidelity while obtaining overall superior performance in video quality and VBench-I2V metrics, including FVD, ISC, MS, and so on. Although some baseline approaches achieve better image fidelity metrics (i.e., PSNR, SSIM, LPIPS), this may be attributed to the fact that they tend to preserve the appearance of the conditioning Ego frame and generate conservative future frames. As a result, they may struggle to model the temporal dynamics in the future, resulting in less plausible action generation. In Table 6, we show the quantitative performance on the *Ego-Exo4D* dataset. The results demonstrate that EgoP^{2}FN achieves competitive image fidelity while obtaining superior performance on video-level quality and several consistency metrics, validating its effectiveness in generating temporally coherent and visually plausible Ego videos. These results demonstrate the effectiveness of jointly modeling poses and videos for future Ego action generation.

Table 5: Quantitative comparisons on Ego video generation on the *AssemblyHands* dataset.

Methods	*Image Fidelity*			*Video Quality*	*VBench-I2V Metrics*				
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	ISC ↑	IBC ↑	SC ↑	MS ↑	AQ ↑
Cosmos-Predict2.5	11.65	0.414	0.555	639.39	0.871	0.903	0.764	0.964	<u>0.334</u>
Wan2.2-TI2V	13.04	0.539	0.459	567.04	0.910	0.938	0.809	0.983	0.310
CogVideoX-I2V	**16.96**	**0.622**	**0.313**	466.68	0.927	0.948	<u>0.917</u>	0.985	0.326
AnimateAnyone†	<u>15.51</u>	<u>0.551</u>	<u>0.370</u>	<u>389.92</u>	0.911	0.936	0.897	0.978	0.312
Zhang *et al.*†	13.11	0.503	0.467	512.68	0.916	0.927	0.878	0.986	0.282
Cosmos-Predict2.5*	12.68	0.450	0.477	412.69	<u>0.944</u>	0.929	0.875	0.983	0.321
EgoControl*†	10.16	0.365	0.609	599.95	0.920	<u>0.950</u>	0.836	<u>0.987</u>	0.330
EgoP^{2}FN (Ours)	13.40	0.540	0.430	**322.93**	**0.976**	**0.951**	**0.926**	**0.995**	**0.334**

Table 6: Quantitative comparisons on Ego video generation on the *Ego-Exo4D* dataset.

Methods	*Image Fidelity*			*Video Quality*	*VBench-I2V Metrics*				
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	ISC ↑	IBC ↑	SC ↑	MS ↑	AQ ↑
Cosmos-Predict2.5	11.22	0.269	0.644	1179.35	0.879	0.880	0.759	0.956	0.392
Wan2.2-TI2V	12.68	0.468	0.452	956.73	0.944	0.943	0.836	0.977	0.439
CogVideoX-I2V	17.53	0.617	0.337	288.66	0.944	0.939	0.894	0.978	0.430
AnimateAnyone†	17.73	0.411	0.385	343.59	0.950	0.940	0.914	0.976	0.425
Zhang *et al.*†	14.41	0.387	0.561	386.26	0.919	0.917	0.881	0.986	0.407
Cosmos-Predict2.5*	<u>18.01</u>	<u>0.634</u>	<u>0.306</u>	245.30	0.977	0.969	<u>0.941</u>	<u>0.996</u>	<u>0.448</u>
EgoControl*†	17.00	0.535	0.389	<u>182.34</u>	<u>0.979</u>	<u>0.971</u>	0.921	0.993	0.445
EgoP^{2}FN (Ours)	**19.40**	**0.644**	**0.298**	**139.32**	**0.983**	**0.974**	**0.941**	**0.996**	**0.450**

Table 7: Quantitative comparisons on Ego video generation on the *EgoMe-pose* dataset.

Methods	*Image Fidelity*			*Video Quality*	*VBench-I2V Metrics*				
	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	ISC ↑	IBC ↑	SC ↑	MS ↑	AQ ↑
Cosmos-Predict2.5	11.90	0.413	0.594	972.05	0.922	0.939	0.846	0.971	0.378
Wan2.2-TI2V	11.88	0.446	0.576	910.64	0.940	<u>0.958</u>	0.886	0.983	0.380
CogVideoX-I2V	<u>12.86</u>	<u>0.461</u>	**0.545**	729.87	0.936	0.949	0.906	0.971	0.371
AnimateAnyone†	**13.34**	**0.465**	<u>0.549</u>	704.00	0.912	0.927	0.884	0.965	0.371
Zhang *et al.*†	10.86	0.415	0.589	770.49	0.928	0.936	0.884	0.977	0.367
Cosmos-Predict2.5*	11.14	0.435	0.625	753.78	<u>0.956</u>	0.938	0.884	0.982	0.382
EgoControl*†	12.40	0.439	0.582	<u>673.68</u>	0.954	**0.972**	**0.922**	<u>0.986</u>	**0.391**
EgoP^{2}FN (Ours)	11.78	0.445	0.596	**668.49**	**0.977**	0.954	<u>0.911</u>	**0.989**	<u>0.390</u>

C.3 Per-dataset Ego 3D Hand Pose Prediction

To further evaluate the performance of EgoP^{2}FN across different data subsets for the Ego 3D hand pose prediction task, we conduct per-dataset experiments on three subsets of our benchmark test set, including the *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose*. The detailed results are shown in Table 8, Table 9, and Table 10. As shown in the Tables, our EgoP^{2}FN generally achieves superior performance across different datasets. In detail, our method obtains the lowest MPJPE and PA-MPJPE errors on all three datasets, demonstrating its capability to accurately predict positions of hand joints under diverse Ego activities and scenarios. Moreover, the proposed EgoP^{2}FN achieves the best MPJVE performance on *AssemblyHands* and *Ego-Exo4D*, while maintaining competitive performance on *EgoMe-pose*. These results indicate that our method can not only effectively capture spatial hand joint positions and structures but also accurately model temporal motion dynamics across various daily activity scenarios, which further demonstrates the robustness of our method and the effectiveness of learning the synergy between poses and pixels in a unified framework.

D More Visualization Results

In this section, we show more visualization examples of the generated Ego video frames and predicted Ego 3D hand poses of our EgoP^{2}FN method and several comparison methods Yang et al. (2025); Pallotta et al. (2026); Bao et al. (2023); Shi et al. (2026a). We provide visualization results

Table 8: Comparisons of Ego 3D hand pose prediction methods on the *AssemblyHands* dataset.

Methods	*Spatial*		*Temporal*
	MPJPE ↓	PA-MPJPE ↓	MPJVE ↓
Random Predict	2004.99	46.83	95.47
Martinez *et al.*	434.58	47.29	34.29
USST	278.57	44.44	52.47
HaMeR	229.33	44.97	45.74
MADiff	<u>208.01</u>	36.50	64.32
EgoH4	223.85	<u>36.29</u>	55.91
Exo2EgoPose	218.22	43.12	<u>18.81</u>
EgoP^{2}FN (Ours)	**136.47**	**22.85**	**10.86**

Table 9: Comparisons of Ego 3D hand pose prediction methods on the *Ego-Exo4D* dataset.

Methods	*Spatial*		*Temporal*
	MPJPE ↓	PA-MPJPE ↓	MPJVE ↓
Random Predict	2033.61	40.37	100.13
Martinez *et al.*	409.84	40.54	44.59
USST	189.74	38.59	61.37
HaMeR	182.10	36.60	55.13
MADiff	170.44	30.67	72.05
EgoH4	170.64	32.90	65.56
Exo2EgoPose	<u>168.16</u>	<u>30.29</u>	<u>29.88</u>
EgoP^{2}FN (Ours)	**138.64**	**20.16**	**29.66**

Table 10: Comparisons of Ego 3D hand pose prediction methods on the *EgoMe-pose* dataset.

Methods	*Spatial*		*Temporal*
	MPJPE ↓	PA-MPJPE ↓	MPJVE ↓
Random Predict	2095.47	33.52	98.81
Martinez *et al.*	448.09	33.47	43.03
USST	267.97	32.76	60.10
HaMeR	224.45	33.26	64.53
MADiff	<u>179.93</u>	29.78	71.74
EgoH4	185.71	30.14	64.46
Exo2EgoPose	271.17	<u>25.89</u>	**26.61**
EgoP^{2}FN (Ours)	**144.66**	**23.27**	<u>31.13</u>

on three subsets of our benchmark, i.e., *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose*, to further demonstrate the effectiveness of our method in diverse Ego scenarios.

D.1 Qualitative Results of Ego Video Generation

We provide more visualization results of Ego video generation across the three subsets of our benchmark test set. As shown in Fig. 10, Fig. 11, and Fig. 12, we visualize the generated Ego action videos of our EgoP^{2}FN method and two representative comparison methods Yang et al. (2025); Pallotta et al. (2026) on samples from *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose*, respectively.

From the visualization results, our EgoP^{2}FN produces more accurate future action videos than the two comparison methods. Specifically, the comparison methods tend to generate Ego videos with limited motion, background distortions, or inaccurate action transitions. In contrast, although EgoP^{2}FN only takes a single conditioning Ego frame and the corresponding language instruction as inputs, it can generate plausible future actions with coherent hand-object interactions. These results demonstrate the effectiveness of our method for future Ego video generation under diverse activity scenarios.

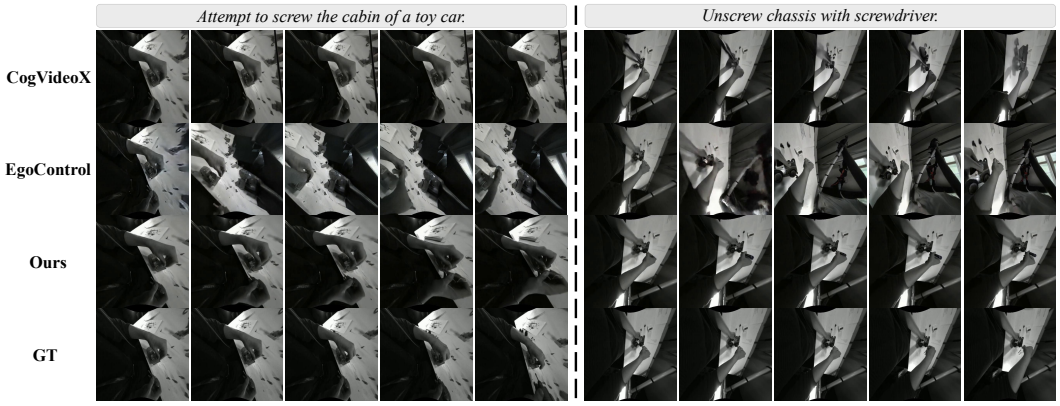

Figure 10: Visualization results of the generated future Ego frames of our EgoP^{2}FN and comparison methods on the *AssemblyHands* dataset.

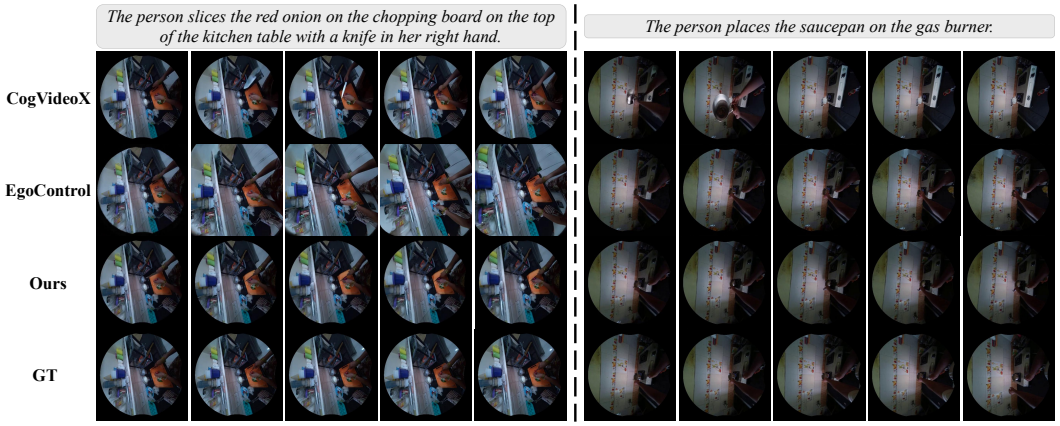

Figure 11: Visualization results of the generated future Ego frames of our EgoP^{2}FN and comparison methods on the *Ego-Exo4D* dataset.

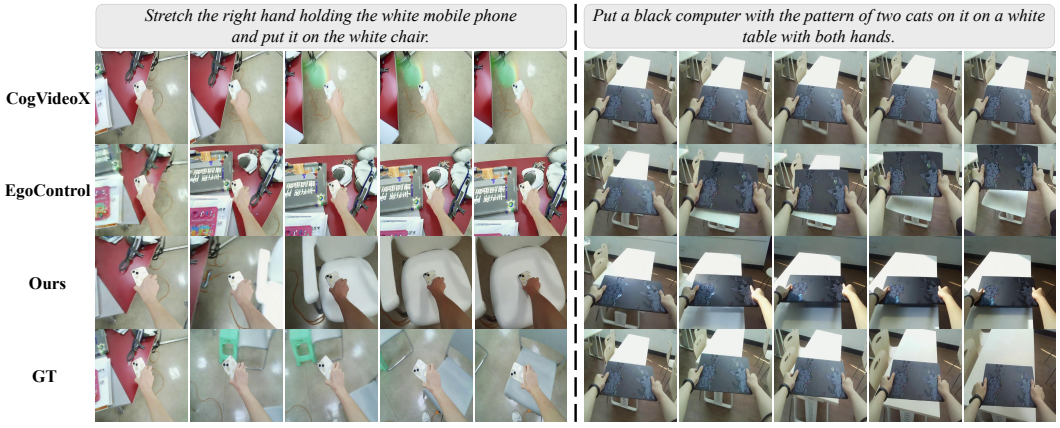

Figure 12: Visualization results of the generated future Ego frames of our EgoP^{2}FN and comparison methods on the *EgoMe-pose* dataset.

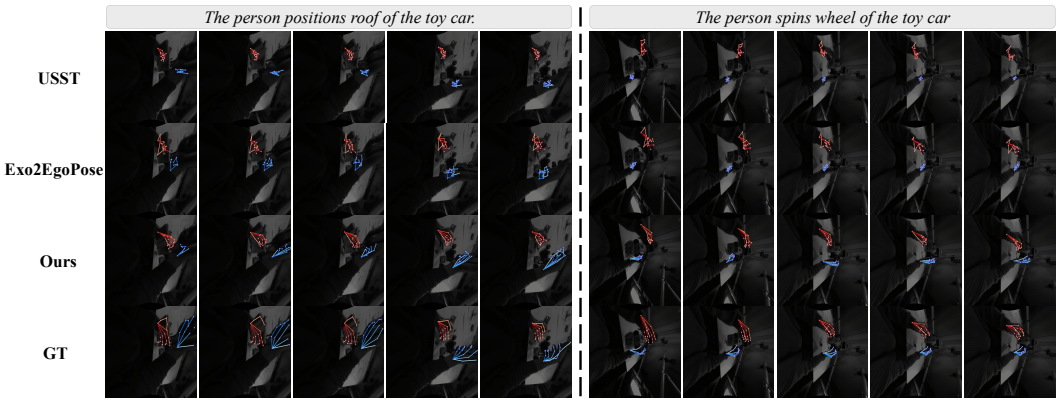

Figure 13: Visualization results of the predicted future Ego 3D hand poses of our EgoP^{2}FN and comparison methods on the *AssemblyHands* dataset.

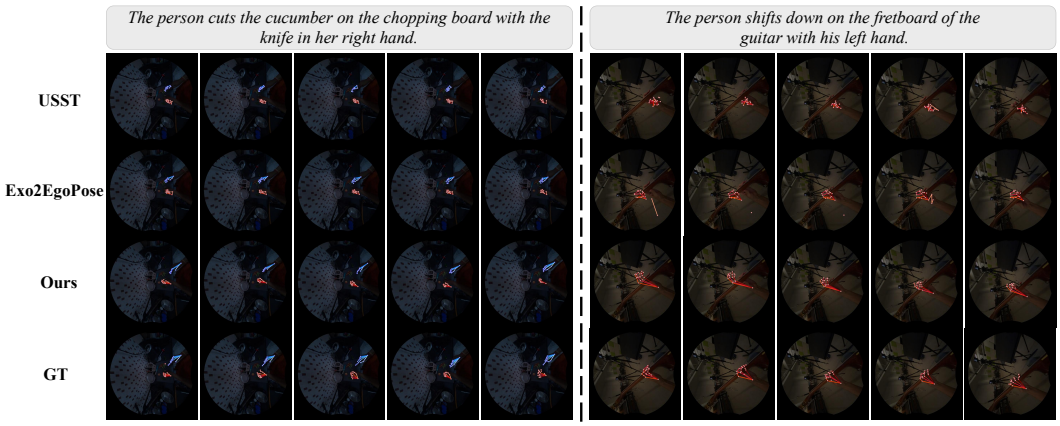

Figure 14: Visualization results of the predicted future Ego 3D hand poses of our EgoP^{2}FN and comparison methods on the *Ego-Exo4D* dataset.

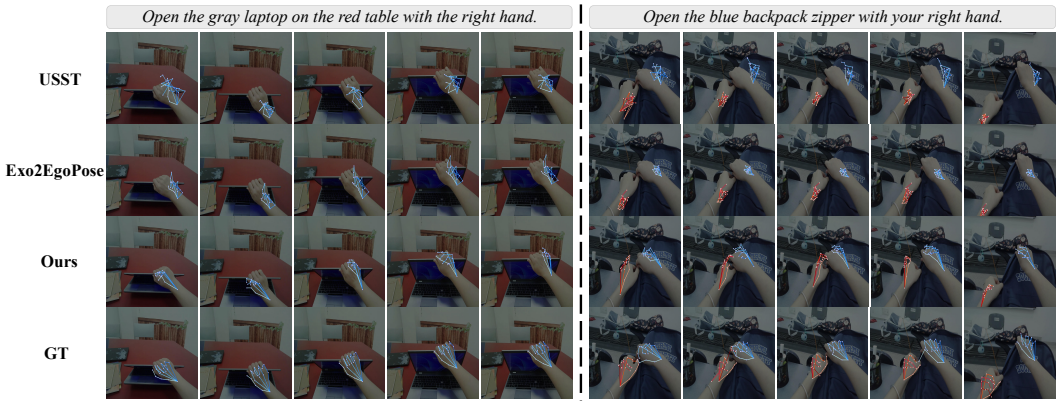

Figure 15: Visualization results of the predicted future Ego 3D hand poses of our EgoP^{2}FN and comparison methods on the *EgoMe-pose* dataset.

D.2 Qualitative Results of Ego 3D Hand Pose Prediction

We also show more visualization results for Ego 3D hand pose prediction across the three subsets of our benchmark test set. As shown in Fig. 13, Fig. 14, and Fig. 15, we project the 3D hand joints

onto 2D frames to visualize the predicted Ego 3D hand poses of the proposed EgoP^{2}FN method and two comparison methods Bao et al. (2023); Shi et al. (2026a), using samples from *AssemblyHands*, *Ego-Exo4D*, and *EgoMe-pose* datasets, respectively.

From the visualization results, our EgoP^{2}FN generally predicts more accurate and plausible future 3D hand poses than the comparison methods. In detail, the comparison methods often predict low-quality hand poses, such as incorrect hand orientations, distorted hand structures, and noticeable joint position errors. In contrast, our method comprehensively leverages geometric cues and observation-aware pose information to predict more accurate joint positions and coherent temporal dynamics. These results further demonstrate the effectiveness of our EgoP^{2}FN method for future Ego 3D hand pose prediction across diverse activity scenarios.