HandWorld: Hand-Centric Unified Video Action Generation

Zhihao Sun[1,2], **Zhiying Du**[1,2], **Xitong Yang**[3], **Zuxuan Wu**[1,2†]
[1]Institute of Trustworthy Embodied AI, Fudan University
[2]Shanghai Key Laboratory of Multimodal Embodied AI
[3]University of Maryland
[†]Correspondence Author
https://sunzhihao18.github.io/HandWorld

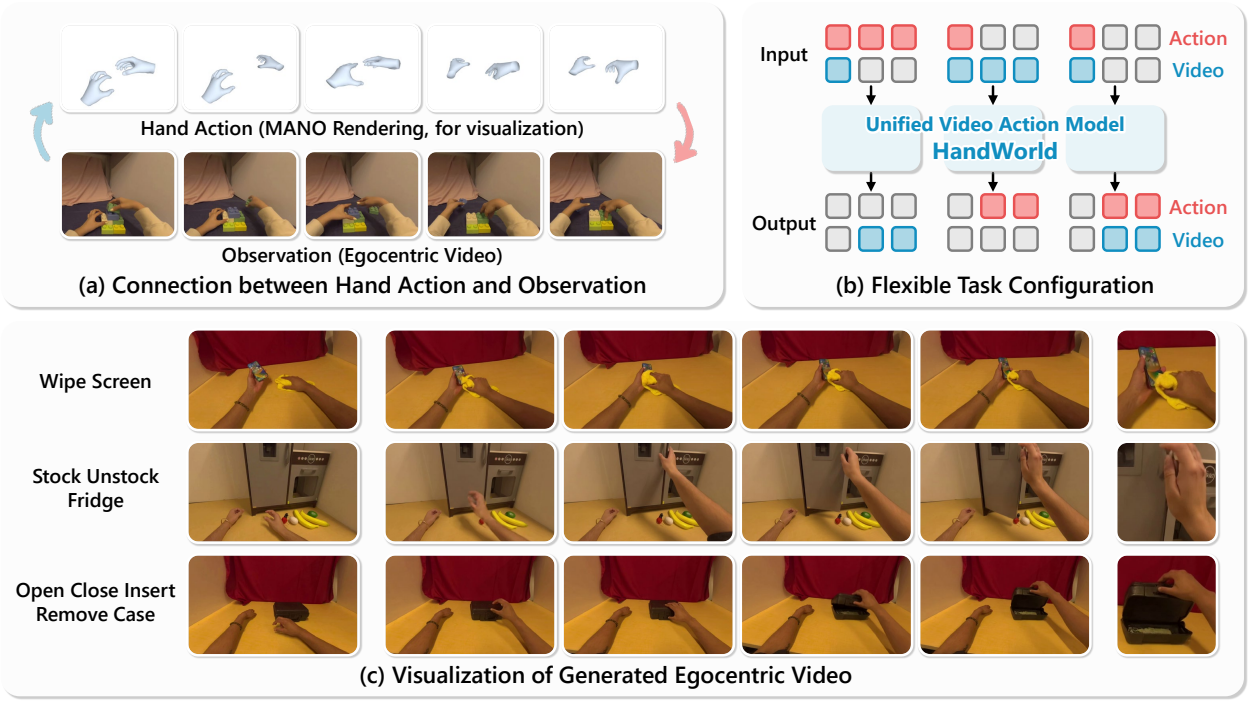

Figure 1. (a) HandWorld models the connection between hand action (MANO-rendered hand only for visualization) and visual observation. (b) The unified framework supports flexible task configurations, enabling action-conditioned video generation, action forecasting, and joint prediction. (c) Examples generated by HandWorld showing accurate hand motion and high-fidelity hand-object interaction.

Abstract

Hand-object interaction forms the foundation of how humans interact with the world. Understanding the connection between hand action and egocentric video is essential for enabling embodied agents to perceive, simulate, and plan like humans. However, it is challenging to learn and predict across hand actions and egocentric videos due to their non-linear relationship. In this work, we introduce HandWorld, a unified generative framework that focuses on hand-object interaction and jointly models egocentric videos and hand actions. HandWorld learns shared cross-domain conditions through a dual-branch condition network that integrates information from both video and action domains. MANO-rendered hand representation is incorporated as an intermediate input to further enhance cross-domain coherence. Conditioned on the shared representation, two decoupled diffusion transformers are trained

to predict in their respective domain. A flexible training strategy enables the model to learn across diverse task configurations, including action forecasting and controllable video generation. Experiments on large-scale egocentric HOI datasets demonstrate that HandWorld achieves high-fidelity video synthesis and accurate action prediction, outperforming existing baselines across diverse scenarios.

1. Introduction

Hand-object interaction (HOI) is one of the most common human behaviors and forms the foundation of how people manipulate and interact with the world. Humans determine hand actions based on their visual observations, and these actions, in turn, alter what will be observed next. In VR/AR, gaming, and embodied intelligence, visual observations typically take the forms of egocentric videos, where hand actions are represented using temporal wrist poses and joint positions to depict hand motion [12, 14, 26, 42]. A comprehensive understanding of the connection between egocentric videos and hand actions is essential for enabling models and embodied agents to perceive, simulate, and plan like humans [6, 13].

Recent studies focus on predicting either action or video while treating the other as a given condition. For example, hand forecasting emphasizes on predicting future hand trajectories or poses from past egocentric observations [5]. In contrast, controllable video generation methods aim to produce realistic videos conditioned on action cues in the forms of masks [29, 34], trajectories [37, 45], human skeletons [16, 22], and high-level action conditions [3, 7, 8, 30, 39], enabling interaction-rich applications in gaming and robotic simulation. These directions highlight the connection between videos and hand actions but remain limited to one-way conditions.

Building a model that can learn and predict across hand actions and egocentric videos presents significant challenges. Hand actions are expressed in structured forms and describe only motion-related dynamics, whereas videos encode the entire scene in pixels, containing richer information about environments and interactions. As a result, the relationship between action and observation is highly nonlinear [4]. The same action can correspond to different observations depending on the manipulated objects and the surrounding scene. A few approaches attempt to couple actions and videos through unified architectures [9, 19, 44]. However, these models primarily optimize for action policy learning and treat video generation as an auxiliary objective, resulting in limited visual fidelity. PEVA [4] further explores the connection between whole-body actions and future egocentric observations, but focuses on large-scale spatial motion such as navigation and locomotion.

Inspired by these unified models, we introduce HandWorld, a unified generative framework that focuses on hand-object interaction and jointly models egocentric videos and hand actions within a single conditional generative process. A central challenge is how to represent the complex, non-linear relationship between hand action and visual observations. We build a shared condition network that learns cross-domain condition signals through two coordinated branches. Each branch encodes its respective domain (video or action), and bidirectional cross-attention enables effective feature fusion between them. To further enhance the coherence between action and video, we introduce MANO-rendered hands [23] as an auxiliary intermediate representation that lies in the visual domain while containing only hand geometry. Built on the shared conditions, HandWorld uses two decoupled diffusion transformers for the video and action generation. Condition sharing enforces cross-domain consistency, whereas decoupling maintains flexibility and improves inference efficiency.

We leverage the large-scale EgoDex dataset [14] and supplement it with MANO-based annotations extracted through our designed reconstruction pipeline. A flexible multi-task training strategy is employed to preserve generation quality in each domain while jointly optimizing the shared condition. Together with the unified generative framework, HandWorld supports diverse task configurations, such as action-conditioned egocentric video generation and action forecasting, as illustrated in Figure 1 (b). Experiments show that HandWorld achieves high-fidelity video synthesis and accurate hand motion prediction across diverse HOI scenarios. Ablation studies further validate the importance of the shared cross-domain condition, with clear improvements over text-based or existing action conditions. In summary, our main contributions are as follows.

- We propose a unified generative framework conditioned on shared cross-domain representation, which is learned by a dual-branch network. MANO-based intermediate input is utilized to enhance cross-domain coherence.
- We utilize two decoupled diffusion transformers to generate video and action separately. They leverage shared conditions to couple both domains, together with a flexible training strategy that supports diverse tasks.
- We evaluate HandWorld on multiple tasks, including action forecasting and action-conditioned video generation, and demonstrate consistent improvements and strong adaptability in egocentric HOI scenarios.

2. Related Work

Controllable Video Generation. Controllable video generation aims to synthesize realistic videos while allowing explicit control over spatial and temporal dynamics. Existing research has explored diverse forms of motion and action control at different levels. Region-based condi-

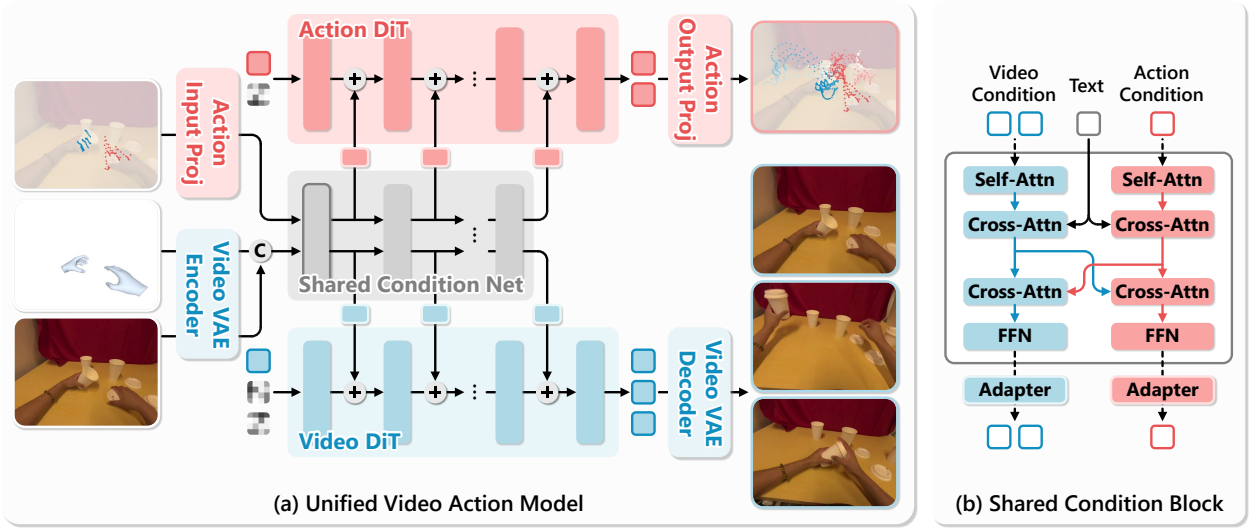

Figure 2. (a) HandWorld integrates a video DiT, an action DiT, and a shared cross-domain condition network, and (b) each condition block in the shared condition network learns across domains through bidirectional cross-attention and outputs with domain-specific adapters.

tions [29, 34] provide motion guidance by indicating specific regions. Trajectory-based methods [37, 45] extend control to temporal dynamics by describing either object translations or camera movement. Character animation techniques achieve higher precision through structured conditions such as skeletons [16, 22] and mesh rendering [46]. Recent studies have begun to explore action-conditioned video generation, where control signals correspond to human or agent behaviors. Text-based control provides a flexible way to express high-level intentions [7, 30, 39], but textual descriptions lack spatial precision and cannot represent detailed dynamics. Several methods predefined discrete action primitives with explicit transition rules and learned latent action spaces from unlabeled videos for more precise control [3, 8]. Although these methods demonstrate the potential of using actions to control visual generation, they remain limited by coarse and abstract representations. In this work, we define actions as hand motions, enabling continuous and fine-grained control that naturally aligns with the dynamics of egocentric hand-object interactions.

Action Video Model. Several works have attempted to leverage video generation models for action prediction. CosHand [31] and InterDyn [2] predict future hand actions using coarse masks generation. However, mask-based representations provide only approximate motion cues, which limit applications requiring highly accurate actions. In the field of embodied intelligence, video models have been employed to facilitate action learning and transfer. For example, [38] generates object flow as an intermediate representation, which is then used to predict actions for transferring manipulation skills across different robotic embodiments and environments. [17] use the latent representations of predicted videos as inputs to a policy network for action prediction. Other studies unify video and action modeling within a shared latent space [19] or a unified vision-language-action model [9, 44], enabling simultaneous prediction of actions and visual outcomes. These methods primarily emphasize improving action generation, while visual prediction often remains an auxiliary objective, resulting in limited visual fidelity. Video-based world models learn internal representations of the environment for future prediction and planning. Works on egocentric pose forecasting [41] and visual foresight [11] demonstrate that predicting future perception supports downstream decision making and planning. PEVA [4] extends this concept to detailed whole-body control, predicting future egocentric videos that integrate realistic motion and visual dynamics. Our model jointly generates accurate hand motion and high-fidelity egocentric videos, maintaining coherence between action and vision.

3. Method

Problem Statement. Given a sequence of observations (*i.e.*, egocentric video frames) $\mathbf{O}_{0:t} = \{\mathbf{O}_0, \ldots, \mathbf{O}_t\}$ and corresponding hand actions $\mathbf{A}_{0:t} = \{\mathbf{A}_0, \ldots, \mathbf{A}_t\}$, our goal is to predict the future video frames $\mathbf{O}_{t+1:T} = \{\mathbf{O}_{t+1}, \ldots, \mathbf{O}_T\}$ and future hand actions $\mathbf{A}_{t+1:T} = \{\mathbf{A}_{t+1}, \ldots, \mathbf{A}_T\}$. Each action $\mathbf{A}_t \in \mathbb{R}^d$ represents a hand pose with dimension d. Unless otherwise specified, we set $d = 138$, which includes the 9D rotation and 3D translation in camera extrinsic parameters, and the 3D coordinates of 21 joints for each hand. Video latent $\mathbf{Z}_t$ is encoded by a pretrained VAE [28] from an egocentric video frame $\mathbf{O}_t \in \mathbb{R}^{H \times W \times 3}$.

Overview. In this section, we introduce our HandWorld framework as illustrated in Figure 2, which models the coupling between hand actions and egocentric videos through a unified generative process. We first introduce the overall action video generation framework, where two decoupled diffusion transformers operate under hierarchical conditions (Section 3.1). We then describe how a shared cross-domain representation is learned through a dual-branch condition network that integrates information from both the vision and action domains (Section 3.2). Finally, we describe the masked multi-objective training strategy that enables optimization across tasks and generalization to diverse scenarios (Section 3.3).

3.1. Action Video Generation Model

Given a hand action sequence $\mathbf{A}_{0:T}$ and video latents $\mathbf{Z}_{0:T}$ encoded by a pretrained VAE encoder [28] from egocentric video frames $\mathbf{O}_{0:T}$, we aim to build a generative model that captures the coupling between action and observation:

$$\begin{aligned} &P(\mathbf{A}_{0:T}, \mathbf{Z}_{0:T}) \\ =&P(\mathbf{A}_0, \mathbf{Z}_0) \prod_{t=0}^{T-1} P(\mathbf{A}_{t+1}, \mathbf{Z}_{t+1} | \mathbf{A}_{0:t}, \mathbf{Z}_{0:t}). \end{aligned} \tag{1}$$

To simplify the model, we make a Markov assumption that the next n states, including video frames and actions, are dependent on the last m states:

$$\begin{aligned} &\prod_{k=1}^{n} P(\mathbf{A}_{t+k}, \mathbf{Z}_{t+k} | \mathbf{A}_{0:t}, \mathbf{Z}_{0:t}) \\ =&P(\mathbf{A}_{t+1:t+n}, \mathbf{Z}_{t+1:t+n} | \mathbf{A}_{0:t}, \mathbf{Z}_{0:t}) \\ =&P(\mathbf{A}_{t+1:t+n}, \mathbf{Z}_{t+1:t+n} | \mathbf{A}_{t-m+1:t}, \mathbf{Z}_{t-m+1:t}). \end{aligned} \tag{2}$$

We train a model parametrized by θ that minimizes the negative log-likelihood:

$$\hat{\theta} = \arg\min_{\theta} \Big[-\log P(\mathbf{A}_{0:T}, \mathbf{Z}_{0:T}) \Big]. \tag{3}$$

Each transition $P(\mathbf{A}_{t+1:t+n}, \mathbf{Z}_{t+1:t+n} | \mathbf{C}_t)$ can be factorized into two components corresponding to the action and video domains, like $P(\mathbf{Z}_{t+1:t+n} | \mathbf{C}_t)$. We leverage the flow matching framework [10, 21] to model each transition $P(\mathbf{S}_{t+1:t+n} | \mathbf{C}_t)$, where s represents either the action or video state. Given the target latent $x_1 = \mathbf{S}_{t+1:t+n}$, a noise sample $x_0 \sim \mathcal{N}(0, I)$, and a timestep $\tau \in [0, 1]$, we define a forward process that obtains an intermediate latent x_τ as the training input. Following Rectified Flows [10], x_τ is defined as a linear interpolation between x_0 and x_1, *i.e.*, $x_\tau = (1-\tau)x_0 + \tau x_1$. The ground-truth velocity is defined as:

$$v_\tau = \frac{dx_\tau}{d\tau} = x_1 - x_0. \tag{4}$$

The model v_θ is trained to predict this velocity field conditioned on the shared cross-domain representation $\mathbf{C}_t$, which encodes the context $\mathbf{A}_{t-m+1:t}$ and $\mathbf{Z}_{t-m+1:t}$ from both domains. The loss function for a transition is formulated as the mean squared error (MSE) between the model output and velocity v_τ:

$$\mathcal{L} = \mathbb{E}_{x_0, x_1, \mathbf{C}_t, \tau} \Big[||v_\theta(x_\tau, \mathbf{C}_t, \tau) - v_\tau||^2 \Big]. \tag{5}$$

We adopt two decoupled diffusion transformers that generate in the action and vision domains, respectively. They shares the hierarchical cross-domain conditions $\mathbf{C}_t$ generated by the condition network. This design allows both models to focus on domain-specific generation while remaining synchronized through a unified conditioning pathway. At each transformer layer, the corresponding level of conditions is fused into the model through residual addition to the hidden states. This hierarchical conditioning enables fine-grained control and maintains consistency between the visual and action outputs. This joint modeling effectively captures the bidirectional relationship between hand actions and egocentric videos, forming the foundation for flexible cross-domain reasoning and generation.

3.2. Shared Cross-Domain Representation

We introduce a dual-branch condition network to establish a shared cross-domain representation $\mathbf{C}_t$ that captures the mutual dependencies between hand actions and egocentric videos. As illustrated in Figure 2 (b), the condition block encodes the two domains separately while allowing information exchange at multiple layers. The video branch processes latent visual features extracted by a pretrained video VAE, while the action branch encodes hand action sequences. At each layer, the two branches interact through bidirectional cross-attention to align their temporal and structural features, producing hierarchical conditions. These signals are further processed by adapters that refine the shared conditions and match the feature scales required for generation.

To facilitate effective interaction between the inherently heterogeneous domains, we introduce an auxiliary intermediate representation based on MANO-rendered hands [23]. This representation belongs to the visual domain but is geometrically aligned with the action domain. It provides a structured, environment-independent signal and enables the condition network to learn a coherent cross-domain representation and stabilize alignment. Since none of the available egocentric datasets provide high-fidelity MANO hand meshes, we design a pipeline to extract 3D hand meshes from raw videos. We first apply multi-hand detection and tracking methods [1, 25] to obtain 2D keypoints and corresponding bounding box sequences for each hand. We then employ the 3D hand reconstruction method HaMeR [24]

to recover per-frame 3D hand meshes from each detected bounding box. The resulting mesh sequences are further refined by removing outliers and smoothing temporally.

3.3. Training with Flexible Objectives

We adopt a flexible training strategy that enables the framework to be jointly optimized across multiple objectives. Each training configuration can be formulated as predicting unknown states in either the action or video domain under partial conditions, corresponding to the unified loss in Equation 5 with different $\mathbf{C}_t$ and x_1. The unused tokens in the condition sequence $\mathbf{C}_t$ are replaced with learnable mask tokens in the corresponding domain.

For instance, when all egocentric video frames and historical actions are provided, the model learns to predict future actions. The training objective $\mathcal{L}_{\text{action}}$ in the action domain is applied to supervise the model. Conversely, when all actions and historical frames are given, it predicts future frames, with supervision applied in the vision domain. The action diffusion transformer and video diffusion transformer are selectively updated depending on the specific task, while the shared cross-domain condition network is optimized across all objectives. As a result, HandWorld learns a coherent generative process that generalizes to multiple task configurations, effectively coupling hand action and egocentric video in hand-object interaction scenarios.

4. Experiment

Dataset. We evaluate our method on the EgoDex dataset [14], which covers a wide range of egocentric hand-object interaction videos and long-term daily activities. EgoDex provides native hand pose annotations and focuses on complex manipulation tasks across 194 different tabletop scenarios, ranging from tying shoelaces to folding laundry. Following the official split, we use approximately 314K samples for training and 3K samples for evaluation. Although EgoDex includes hand pose annotations, their alignment to the camera coordinate system can introduce minor inaccuracies, and the provided annotations cannot be directly fitted to the MANO hand model. Therefore, we apply the hand reconstruction and refinement pipeline introduced in Section 3.2 to recover accurate MANO meshes. The reconstructed meshes are then rendered into the video domain to serve as intermediate vision inputs for the condition network. We discard a small number of videos with excessive duration or missing hand visibility.

Implementation Details. The video components of our framework, including the video VAE, text tokenizer, text encoder, and video diffusion transformer, are initialized from the pretrained Wan2.2-TI2V-5B model [33], which provides strong video generation capability. The parameters of blocks for the video input in the condition network are copied from the pretrained diffusion model, and additional linear adapter layers are zero-initialized to enable smooth fine-tuning. All other components, including the action diffusion transformer and action branch in the condition network, are randomly initialized.

During training, all video frames are resized to a resolution of 832×480 and temporally clipped to a maximum of 49 frames, or the largest valid sequence satisfying $T = 1 + 4t$ as required by the VAE design. The action sequences are temporally aligned with the corresponding video clips, ensuring one-to-one correspondence across frames. Text prompts are coarse action descriptions provided by the dataset and are encoded using the pretrained text encoder. We train our framework on 2×8 NVIDIA H100 GPUs with 80 GB memory. More training details are provided in the supplementary material.

4.1. Hand-Centric Video Generation

Evaluation Metrics. We evaluate the generated videos using multiple quantitative metrics. For visual quality, we assess semantic alignment with the frame-level CLIP score [27]. PSNR, SSIM [36], and LPIPS [43] are further employed to measure structure and perceptual similarity. Temporal coherence across frames is evaluated using the Fréchet Video Distance (FVD) [32]. To quantify the quality of the generated hand-object interactions and the accuracy of hand positioning, we compute a hand-region CLIP score ($\text{CLIP}_{\text{hand}}$) to measure semantic consistency within localized hand areas. We also apply an existing hand detector [25] to identify hand regions in the generated videos and compute the IoU between the detected regions and the hand locations in the ground-truth videos. Details of the hand-related metrics are provided in the supplementary material.

Baselines. We compare our method with both general text-to-video generation models and hand-aware baselines. The text-to-video baselines include AnimateAnything [16], CogVideoX-I2V-5B [40], and Wan2.2-TI2V-5B [33]. Except for AnimateAnything, we fine-tune other pretrained models on the EgoDex dataset to ensure fair comparison. To further compare with methods that incorporate action control, we include HANDI [20] and visual action prompt technology proposed by Wang *et al.* [35]. HANDI generates the motion region in the video where detailed activities occur, and guides a diffusion model to synthesize action-consistent videos. Wang *et al.* render actions into visual skeletons, serving as domain-agnostic representations that maintain both geometric precision and cross-domain adaptability for complex actions. Since the paper does not provide released code or pretrained weights, we reproduce the method based on the descriptions in the paper. We leverage CogVideoX [40] as the pretrained base model, adopt a conditioning mechanism with visual skeleton, and fine-tune the DiT backbone using LoRA [15].

Table 1. Comparison of hand-centric egocentric video generation.

Method	Condition	EgoDex	Visual Quality					Hand Action	
			CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	$\text{CLIP}_{\text{hand}}$ ↑	IoU ↑
AnimateAny [16]	Text		0.9152	23.97	0.867	0.210	913.9	0.8844	0.4073
CogVideoX-I2V-5B [40]	Text	✓	0.8825	19.58	0.814	0.324	568.0	0.8638	0.2770
Wan2.2-TI2V-5B [33]	Text	✓	0.9306	22.86	0.842	0.223	482.3	0.9132	0.5005
HANDI [20]	Mask		0.8898	23.55	0.866	0.226	1303.9	0.8466	0.1825
Wang *et al.* [35]	Skeleton	✓	0.9031	20.73	0.804	0.297	516.1	0.8750	0.5752
HandWorld	MANO	✓	**0.9568**	**26.27**	**0.874**	**0.132**	**133.9**	**0.9461**	**0.8291**

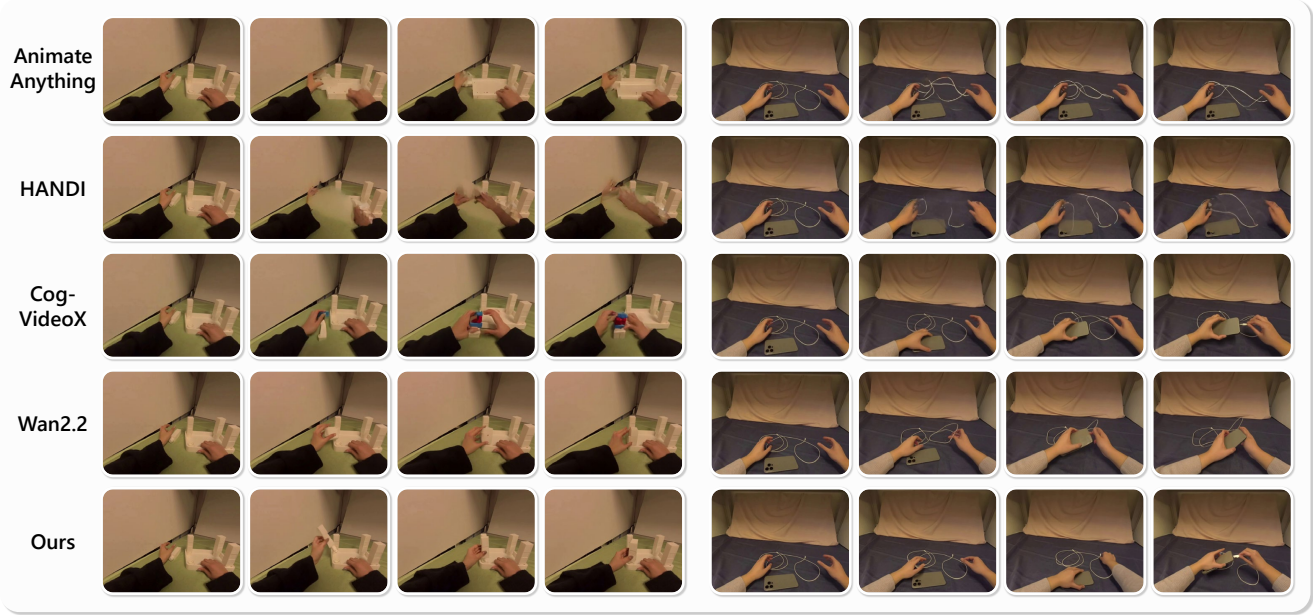

Figure 3. Qualitative comparison of hand-centric egocentric video generation between baseline models and HandWorld.

Comparison. We evaluate the ability of HandWorld to generate visually realistic and hand-centered egocentric videos controlled by action. Table 1 reports quantitative comparisons across several text-to-video and action-conditioned baselines in terms of visual quality and hand action consistency. Among all methods, HandWorld achieves the best overall performance across both visual and hand metrics. Compared to general text-to-video models such as AnimateAnything, CogVideoX, and Wan2.2, which rely only on text condition, HandWorld produces videos with precise hand details and more stable hand trajectories. The improvements in FVD demonstrate its stronger realism and temporal coherence, while higher CLIP and SSIM scores confirm its enhanced perceptual quality. When compared to hand-centered models, HANDI and Wang *et al.*, HandWorld shows clear advantages in both visual fidelity and motion controllability. HANDI achieves localized motion control through region-based conditions, but lacks precise structural constraints, leading to inconsistent hand poses and trajectories. Wang *et al.* employ skeleton-based visual prompts that improve coarse motion alignment but cannot fully capture the subtle deformation of fingers. In contrast, our shared cross-domain representation provides a physically grounded control signal, enabling natural, fine-grained hands that are consistent with surrounding environment dynamics.

Visualization. We provide qualitative comparisons with the fine-tuned Wan2.2-TI2V-5B, fine-tuned CogVideoX-I2V-5B, AnimateAnything, HANDI, and our HandWorld in Figure 3. Because most existing methods cannot support action conditions, our comparison focuses on the quality of the generated hand-object interaction rather than the accuracy of action control. For clarity, Figure 4 visualizes the consistency between the provided action condition and the generated videos. While the Wan2.2 and CogVideoX models generally deliver reasonable background fidelity and overall visual quality, they struggle to generate plausible hand motions, often resulting in distorted hand geometry or unsta-

Table 2. Comparison of hand action forecasting.

	Model (Policy)	Avg Distance		Final Distance	
		$K=1$	$K=5$	$K=1$	$K=5$
X-IL [18]	Dec-BC	0.045	0.045	0.062	0.062
	Dec-DDPM	0.053	0.044	0.071	0.050
	Dec-FM	0.052	0.042	0.071	0.049
	EncDec-BC	**0.044**	0.044	0.060	0.060
	EncDec-DDPM	0.052	0.042	0.071	0.048
	EncDec-FM	0.051	0.041	0.070	0.047
Ours	DDPM	0.047	0.043	0.053	**0.045**
	FM	**0.044**	**0.039**	**0.051**	**0.045**

ble articulations. In contrast, HandWorld produces videos with significantly improved hand fidelity, including stable geometry, precise articulation, and consistent motion across frames. Moreover, HandWorld demonstrates notably better hand-object interaction quality, maintaining more coherent contact patterns and fewer object flickering artifacts compared with prior methods. These results highlight the advantages of our cross-domain condition and MANO-based representation, which enable precise action control while improving interaction plausibility.

Computational Efficiency. Although HandWorld introduces two decoupled diffusion transformers and a shared condition network, its inference remains efficient. When performing generation in a single domain, such as video or action synthesis, only the corresponding diffusion transformer is activated. Cross-domain information remains accessible through the shared condition network, enabling effective interaction without additional computational overhead. On a single NVIDIA H100 GPU, HandWorld generates a 49-frame egocentric video in 33.8 s. For comparison, AnimateAnything completes in 34.4 s, while the two-stage HANDI requires 37.5 s. The pretrained text-to-video model Wan2.2 takes 22.3 s. HandWorld, built on the Wan2.2 backbone, delivers substantial improvements in hand fidelity without introducing a significant latency penalty, benefiting from the decoupled diffusion transformer design. All inference results are measured with 50 denoising steps.

4.2. Hand Action Prediction

Evaluation Metrics. We evaluate the action forecasting performance with a best-of-K metric following [14]. For each test sample, we generate K predictions to capture diverse possible modes. We then compute the distance between the ground truth action sequence and the closest predicted action sequence among the K samples. The distance is calculated as the Euclidean distance between predicted 3D keypoint positions and their ground truth 3D counterparts, averaged over the predicted horizon and the 12 keypoints (*i.e.*, the wrist and fingertips of both hands).

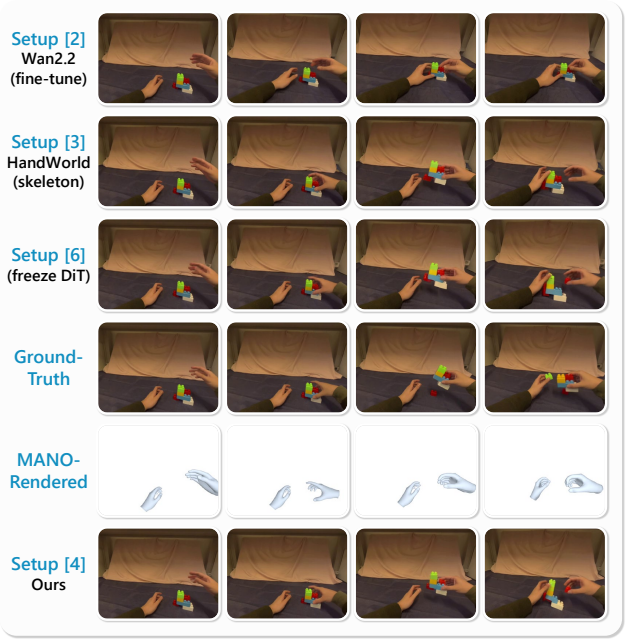

Figure 4. Visualization of different ablation setups.

Baselines. We compare our method with imitation learning policies from the X-IL framework [18]. The framework includes both decoder-only (Dec) and encoder-decoder (EncDec) transformer architectures, and evaluates three policy formulations, such as behavior cloning (BC), denoising diffusion (DDPM), and flow matching (FM). The performance of these baselines is reported in [14].

Comparison. Table 2 compares HandWorld with the X-IL baselines under the best-of-K evaluation for action forecasting. Across all metrics, our framework combined with the flow-matching strategy achieves the best performance, producing the lowest average and final prediction errors at both K=1 and K=5. This demonstrates that HandWorld is able to predict plausible future hand motions more accurately in egocentric HOI scenarios. In addition, the gap between single-prediction and multi-predictions performance is smaller for our model, indicating more stable and reliable predictions. More results on hand action forecasting are provided in the supplementary material.

4.3. Ablation Study

To understand the contribution of each component in HandWorld, we conduct a series of ablation experiments, summarized in Table 3. We also provide qualitative comparisons, as shown in Figure 4, including the ground-truth video and the corresponding MANO-rendered hands, which facilitate a clear comparison of hand geometry and action alignment across different setups. Starting from the pretrained Wan2.2 model (Setup 1) and its fine-tuned version

Table 3. Ablation study of HandWorld components and training strategies.

Setup (Architecture and Strategy)	**Condition**	**Visual Quality**					**Hand Action**	
		CLIP ↑	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓	$\text{CLIP}_{\text{hand}}$ ↑	IoU ↑
[1] Wan2.2-TI2V-5B (pre-train)	Text	0.9250	22.44	0.835	0.237	704.3	0.9047	0.4622
[2] Wan2.2-TI2V-5B (fine-tune)	Text	0.9306	22.86	0.842	0.223	482.3	0.9132	0.5005
[3] HandWorld	Skeleton	**0.9591**	25.18	0.858	0.149	245.0	0.9380	0.7904
[4] HandWorld	MANO	0.9568	**26.27**	**0.874**	0.132	**133.9**	**0.9461**	**0.8291**
[5] w/o. shared condition net	MANO	0.9358	23.32	0.830	0.241	497.0	0.9162	0.6966
[6] w/o. video DiT fine-tuning	MANO	0.9533	25.95	0.871	0.137	174.8	0.9398	0.8077
[7] w/o. multi-task training	MANO	0.9441	24.98	0.863	**0.131**	139.2	0.9384	0.7996

on EgoDex (Setup 2), we observe that the text-only condition provides reasonable visual quality but limited controllability of hand motion, as reflected in the low $\text{CLIP}_{\text{hand}}$ and visible artifacts in hand geometry. Since our video diffusion transformer, video VAE, and text encoder are all initialized from Wan2.2, this configuration serves as the baseline.

When introducing our framework HandWorld with joint action video learning and the shared cross-domain condition network (Setup 3), both visual and motion alignment improve notably, reducing the FVD from 482.3 to 245.0. The decrease in LPIPS further indicates higher perceptual fidelity. However, the skeleton-based condition remains coarse and lacks physical plausibility for hand details, as also seen in the visualization. Our full model (Setup 4), which incorporates MANO-rendered hands as an auxiliary intermediate representation, achieves the best performance across most metrics. Compared to the skeleton-based setup, the MANO-based condition significantly enhances both visual fidelity and hand consistency, as shown in Figure 4.

When the shared condition network is removed (Setup 5), we replace it with a ControlNet condition module to directly inject MANO-rendered signals. Both visual quality and hand accuracy degrade, confirming the importance of the shared condition network in maintaining coherence between the two domains. Disabling fine-tuning of the pre-trained video diffusion transformer (Setup 6) leads to noticeably worse perceptual and temporal quality (*e.g.*, FVD increases from 133.9 to 174.8), suggesting that domain adaptation to egocentric hand-object videos is crucial for capturing fine-grained dynamics, albeit at the cost of additional training overhead. Finally, removing multi-task training (Setup 7) weakens both visual and action consistency, indicating that cross-task optimization reinforces the coupling between visual and action domains.

5. Conclusion

In this work, we introduced HandWorld, a unified generative framework that jointly models egocentric videos and hand actions through shared cross-domain representations. By integrating a dual-branch condition network, a MANO-rendered intermediate representation, and two decoupled diffusion transformers operating under shared conditions, HandWorld effectively learns and predicts across hand actions and egocentric videos. A flexible multi-task training strategy further enables the framework to adapt to diverse task configurations, including action forecasting and action-conditioned video generation. Extensive experiments on large-scale egocentric hand-object interaction datasets demonstrate that HandWorld achieves high-fidelity visual synthesis and accurate hand action prediction across complex interactive scenarios.

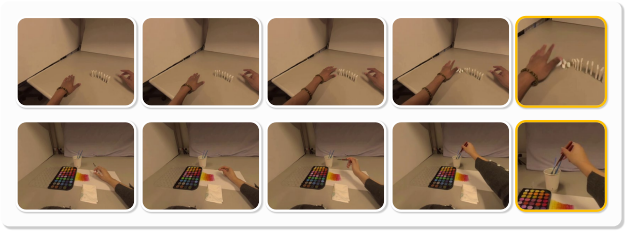

Figure 5. Failure cases of HandWorld. The last column highlights the object-level temporal consistency in the generated video.

Limitation. While HandWorld improves hand fidelity and action coherence in egocentric video generation, maintaining object-level temporal consistency remains a key challenge. As illustrated in Figure 5, our model may fail to produce plausible interactions for small objects, leading to flickering or disappearance issues, which are commonly observed in video diffusion models. A crucial limitation is the absence of object-level supervision. Large-scale datasets such as EgoDex do not provide annotations for objects, making it difficult to learn object dynamics. Future work may explore supervision, such as trajectories and contact maps, or design objectives that encourage object consistency and hand-object interaction modeling.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (No. 62521004).

References

[1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Botsort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 4

[2] Rick Akkerman, Haiwen Feng, Michael J Black, Dimitrios Tzionas, and Victoria Fernández Abrevaya. Interdyn: Controllable interactive dynamics with video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 3

[3] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 2024. 2, 3

[4] Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. *arXiv preprint arXiv:2506.21552*, 2025. 2, 3

[5] 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*, 2023. 2

[6] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[7] Homanga Bharadhwaj, Debidatta Dwibedi, Abhinav Gupta, Shubham Tulsiani, Carl Doersch, Ted Xiao, Dhruv Shah, Fei Xia, Dorsa Sadigh, and Sean Kirmani. Gen2act: Human video generation in novel scenarios enables generalizable robot manipulation. *arXiv preprint arXiv:2409.16283*, 2024. 2, 3

[8] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *International Conference on Machine Learning*, 2024. 2, 3

[9] Jiayi Chen, Wenxuan Song, Pengxiang Ding, Ziyang Zhou, Han Zhao, Feilong Tang, Donglin Wang, and Haoang Li. Unified diffusion vla: Vision-language-action model via joint discrete denoising diffusion process. *arXiv preprint arXiv:2511.01718*, 2025. 2, 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 2024. 4

[11] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation*, 2017. 3

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[13] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2

[14] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 2, 5, 7

[15] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *International Conference on Learning Representations*, 2022. 5

[16] 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*, 2024. 2, 3, 5, 6

[17] Yucheng Hu, Yanjiang Guo, Pengchao Wang, Xiaoyu Chen, Yen-Jen Wang, Jianke Zhang, Koushil Sreenath, Chaochao Lu, and Jianyu Chen. Video prediction policy: A generalist robot policy with predictive visual representations. *arXiv preprint arXiv:2412.14803*, 2024. 3

[18] Xiaogang Jia, Atalay Donat, Xi Huang, Xuan Zhao, Denis Blessing, Hongyi Zhou, Han A Wang, Hanyi Zhang, Qian Wang, Rudolf Lioutikov, et al. X-il: Exploring the design space of imitation learning policies. *arXiv preprint arXiv:2502.12330*, 2025. 7

[19] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model. *arXiv preprint arXiv:2503.00200*, 2025. 2, 3

[20] Yayuan Li, Zhi Cao, and Jason J Corso. Handi: Hand-centric text-and-image conditioned video generation. *arXiv preprint arXiv:2412.04189*, 2024. 5, 6

[21] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 4

[22] Yifang Men, Yuan Yao, Miaomiao Cui, and Liefeng Bo. Mimo: Controllable character video synthesis with spatial decomposed modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[23] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2, 4

[24] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 4

[25] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 4, 5, 1

[26] Jing Qi, Li Ma, Zhenchao Cui, and Yushu Yu. Computer vision-based hand gesture recognition for human-robot in-

teraction: a review. *Complex & Intelligent Systems*, 2024. 2

[27] 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*, 2021. 5

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3, 4

[29] Xiaoyu Shi, Zhaoyang Huang, Fu-Yun Wang, Weikang Bian, Dasong Li, Yi Zhang, Manyuan Zhang, Ka Chun Cheung, Simon See, Hongwei Qin, et al. Motion-i2v: Consistent and controllable image-to-video generation with explicit motion modeling. In *ACM SIGGRAPH Conference*, 2024. 2, 3

[30] Tomáš Souček, Dima Damen, Michael Wray, Ivan Laptev, and Josef Sivic. Genhowto: Learning to generate actions and state transformations from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[31] Sruthi Sudhakar, Ruoshi Liu, Basile Van Hoorick, Carl Vondrick, and Richard Zemel. Controlling the world by sleight of hand. In *European Conference on Computer Vision*, 2024. 3

[32] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. In *International Conference on Learning Representations Workshop*, 2019. 5

[33] 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. 5, 6

[34] Jiawei Wang, Yuchen Zhang, Jiaxin Zou, Yan Zeng, Guoqiang Wei, Liping Yuan, and Hang Li. Boximator: Generating rich and controllable motions for video synthesis. *arXiv preprint arXiv:2402.01566*, 2024. 2, 3

[35] Yuang Wang, Chao Wen, Haoyu Guo, Sida Peng, Minghan Qin, Hujun Bao, Xiaowei Zhou, and Ruizhen Hu. Precise action-to-video generation through visual action prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 5, 6

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004. 5

[37] Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH Conference*, 2024. 2, 3

[38] Mengda Xu, Zhenjia Xu, Yinghao Xu, Cheng Chi, Gordon Wetzstein, Manuela Veloso, and Shuran Song. Flow as the cross-domain manipulation interface. *arXiv preprint arXiv:2407.15208*, 2024. 3

[39] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 2, 3

[40] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 5, 6

[41] Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

[42] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 5

[44] Wenyao Zhang, Hongsi Liu, Zekun Qi, Yunnan Wang, Xinqiang Yu, Jiazhao Zhang, Runpei Dong, Jiawei He, Fan Lu, He Wang, et al. Dreamvla: a vision-language-action model dreamed with comprehensive world knowledge. *arXiv preprint arXiv:2507.04447*, 2025. 2, 3

[45] Zhenghao Zhang, Junchao Liao, Menghao Li, Zuozhuo Dai, Bingxue Qiu, Siyu Zhu, Long Qin, and Weizhi Wang. Tora: Trajectory-oriented diffusion transformer for video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. 2, 3

[46] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, 2024. 3