Egocentric World Model for Photorealistic Hand Object Interaction Synthesis

Dayou Li[1*], Lulin Liu[1,2*], Bangya Liu[3], Shijie Zhou[4], Jiu Feng[5], Ziqi Lu[6‡], Minghui Zheng[1], Chenyu You[7], Zhiwen Fan[1]

[1]Texas A&M University [2]University of Minnesota
[3]University of Wisconsin–Madison [4]University of California, Los Angeles
[5]University of Texas at Austin [6]Amazon
[7]Stony Brook University

Project Website: `https://egohoi.github.io/`

Abstract. To serve as a scalable data source for embodied AI, world models should act as true simulators that infer interaction dynamics strictly from user actions, rather than mere conditional video generators relying on privileged future object states. In this context, egocentric Hand Object Interaction (HOI) world models are critical for predicting physically grounded first-person rollouts. However, building such models is profoundly challenging due to rapid head motions, severe occlusions, and high-DoF hand articulations that abruptly alter contact topologies. Consequently, existing approaches often circumvent these physics challenges by resorting to conditional video generation with access to known future object trajectories. We introduce EgoHOI, an egocentric HOI world model that breaks away from this shortcut to simulate photorealistic, contact-consistent interactions from action signals alone. To ensure physical accuracy without future-state inputs, EgoHOI distills geometric and kinematic priors from 3D estimates into physics-informed embeddings. These embeddings regularize the egocentric rollouts toward physically valid dynamics. Experiments on the HOT3D dataset demonstrate consistent gains over strong baselines, and ablations validate the effectiveness of our physics-informed design.

Keywords: Egocentric Vision · World Model · Hand-Object Interaction

1 Introduction

Egocentric Hand Object Interaction (HOI) world models aim to simulate physically grounded interaction dynamics from a first-person perspective driven by user controls. The resulting predictive rollouts are increasingly important beyond visual synthesis; they can serve as scalable training and evaluation data for a wide

[*] Equal contribution.
[‡] Work performed independently; unrelated to the author's role at Amazon.

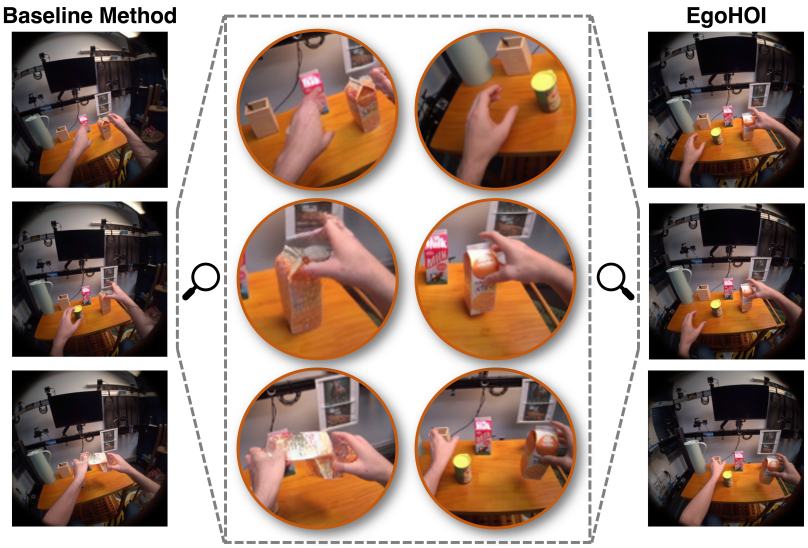

Fig. 1: Qualitative comparison between a baseline and EgoHOI. Starting from the same first frame, EgoHOI integrates physics-informed embeddings to model hand–object interaction dynamics, improving ego-motion consistency, kinematic fidelity, and object integrity over time. Zoom-ins highlight clearer contact details.

range of downstream tasks, including dexterous manipulation [10,27,49], Vision-Language-Actions Models (VLAs) [7,42,43], spatiotemporal reasoning [51], physical intelligence [24], and data augmentation [21]. Despite this potential, the inherent complexity of egocentric HOI poses major challenges. High-DoF hand articulation rapidly changes contact topologies and object states, while severe occlusions and small head motions make the interaction highly ambiguous. The difficulty of this domain is also evident in recent spatiotemporal representation research [51], even state-of-the-art VLMs still struggle to reliably understand and reason about the complex dynamics in egocentric HOI videos. If interpreting these interactions is already challenging, building an egocentric HOI world model is even harder. Such a model must go beyond passive observation and actively generate future frames that satisfy user-specified actions, physical laws, and dynamical consistency, producing reliable physical rollouts rather than only visual plausibility.

Despite recent progress in HOI video generation, a reliable egocentric HOI world model remains largely absent. This gap stems from two fundamental limitations in existing pipelines. First, current HOI video generation [28, 32, 46] is predominantly studied in static, exocentric views. These third-person perspectives do not directly support first-person rollouts with dynamic interactions, which are essential in embodied AI and robotics. Second, many existing approaches [9, 17, 18, 25] condition heavily on privileged future object information, such as ground-truth trajectories and waypoints. By explicitly defining how the

object should move in the future, these methods largely bypass the core challenge of contact-driven dynamics reasoning. This turns the task into conditional video generation, which conflicts with the formulation of a world model that must infer interaction dynamics and future states strictly from current observations and user-specified actions. Ideally, the model should predict future hand-object interaction purely from action signals, rather than replaying privileged future object states. Because this removes future-state shortcuts, simply scaling RGB data is insufficient for physical accuracy. Instead, the model should be augmented with physics-informed embeddings that inject explicit metric and kinematic structure. Fortunately, recent advances in 3D foundation models for geometry [19, 35, 36] and hand pose [29] now make it increasingly practical to extract these crucial metric and kinematic priors directly from unstructured videos, opening opportunities to model high-DoF interaction through explicit 3D hand kinematics, camera motion, and contact-aware geometry.

In this paper, we present EgoHOI, an egocentric HOI world model for photorealistic hand-object interaction simulation that integrates physics-informed embeddings into the generative rollout process. EgoHOI uses interpretable estimates from scene, hand, and geometry models, and converts them into embeddings that provide complementary physical structure for learning egocentric dynamics, including hand kinematics, metric-scale ego motion, and object integrity. These embeddings are incorporated through lightweight adapters to preserve the pretrained generative prior while imposing the metric grounding, temporal consistency, and contact-aware dynamics. Unlike earlier trials utilizing access to privileged future states, EgoHOI enables causal simulation of physically plausible hand-object interactions while retaining the generative diversity and reconstruction quality trained from massive data. We demonstrate an example in Fig. 1 to show the resulting simulated rollouts, and summarize our contributions as follows:

- We introduce EgoHOI, an egocentric HOI world model that simulates first-person hand–object interaction dynamics under user-specified actions and controls, without privileged future object states.
- We introduce physics-informed embeddings derived from 3D geometry and kinematic priors to regularize egocentric HOI rollouts toward physically consistent interaction dynamics.
- We show EgoHOI better aligns rollouts with user actions and improves hand–object interaction realism, with comprehensive metrics and ablations validating our physics-informed embeddings.

2 Related Works

Egocentric World Model & Video Generation. Egocentric World model aims to simulate the physical world by predicting future visual observations in a first-person viewpoint. Recently, there has been progress in the general world models [5, 31], which usually take in detailed language descriptions as control signals to generate future rollouts. For instance, Nvidia's Cosmos [1], which was

trained on 100M language-video pairs, can capture diverse real-world physics and be fine-tuned for specific tasks. However, most of these models operate primarily in image space, which makes physically plausible rollouts under high degrees of freedom still challenging in practice. Unlike general world models, egocentric world models often focus on more challenging tasks. In egocentric settings, viewpoints constantly change, while hand-object interactions frequently influence the dynamic evolution of the scene, making it significantly more difficult for world models to simulate [2, 12, 15]. PlayerOne [33] drives a first-person simulator with whole-body motion to maintain consistent visual rendering across scenes, while GEM [13] jointly generates RGB and depth streams conditioned on ego-trajectory and object identity features. Beyond PlayerOne and GEM, many existing efforts focus on designing view-transformation or hand-trajectory controls for first-person generation, rather than grounding rollouts in reconstruction-based 3D priors or causal hand–object dynamics. EgoExo-Gen [39] is one example that generates a first-person video by watching a third-person view of the same scene. Zhang et al. [44] first predict a coarse hand trajectory and then condition a latent diffusion model to generate future frames. Overall, prior egocentric world models focus on plausible first-person rendering, but hand–object interaction is less emphasized, and maintaining physically consistent interaction dynamics remains difficult.

Hand-Object Interactions Synthesis. Hand–object interaction (HOI) synthesis targets interactions that are visually realistic and physically plausible by jointly modeling appearance, relative geometry, and contact semantics [6, 8, 11, 20, 22, 23, 26]. In the egocentric setting, MEgoHand [50] synthesizes hand–object motion from first-person RGB together with text and an initial hand pose. Beyond egocentric inputs, a dominant trend in general HOI generation is reference-conditioned interaction alignment, where models leverage reference signals such as target poses, relative transforms, or motion trajectories to improve controllability and contact quality. These include image-level refinement of hand–object alignment [18], geometry and relation aware motion generation with diffusion [41], multimodal video generation guided by an HOI adapter [17], open world synthesis with affordance guidance and physics refinement [47], and tokenized sequence models that translate between language and long 3D HOI sequences [16].

3 Method

3.1 Preliminary

World Model Formulation A world model $\mathbf{f}_\theta$ captures action-driven internal state dynamics and enables rollouts by iteratively predicting how the state evolves under actions. Let $x_t \in \mathcal{X}$ denote the internal state and let a_t denote the action applied at time t. The action-driven transition dynamics are:

$$x_{t+1} = \mathbf{f}_\theta(x_t, a_t), \tag{1}$$

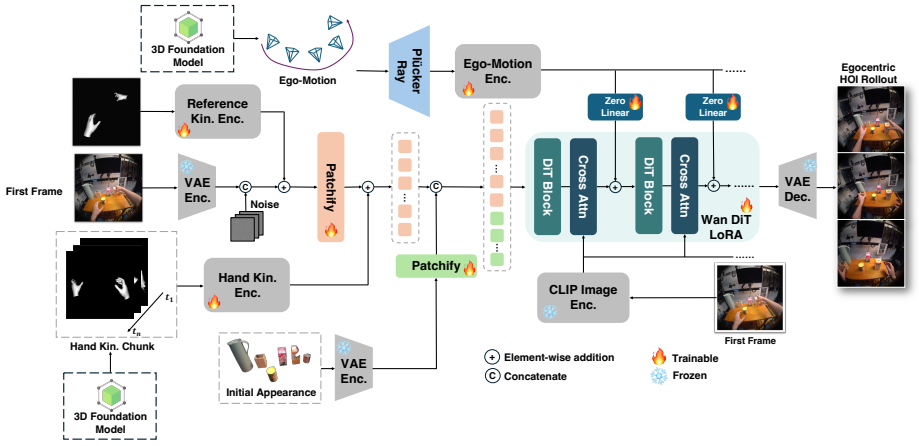

Fig. 2: Overview of EgoHOI. We formulate EgoHOI as an egocentric world model that predicts action-driven latent transitions with a DiT backbone. Lightweight adapters inject physics-informed embeddings and object appearance, improving hand–object realism, ego-motion consistency, and object identity across viewpoints.

and the observation function g_ϕ maps the internal state to the observation space:

$$o_t \sim g_\phi(x_t), \tag{2}$$

where o_t denotes the observable signal. Given an initial observation o_1, an encoder produces the initial internal state $x_1 = e_\phi(o_1)$. A rollout over horizon L is obtained by recursively applying $\mathbf{f}_\theta$ under an action sequence $\{a_t\}_{t=1}^{L}$ and mapping the resulting latent states to predicted observations via g_ϕ.

In this view, a world model serves as an action-driven simulator: actions are treated as interventions that induce state transitions, and rollouts should predict the causal effects of actions on the world state, instead of only reconstructing the agent's kinematic trajectory. In EgoHOI's setting, we instantiate the transition model $\mathbf{f}_\theta$ with a DiT backbone to realize action-driven evolution of the model's internal state in a VAE latent space; e_ϕ and g_ϕ are the VAE encoder and decoder that map video frames to and from this latent representation.

Diffusion Process Video generative models [3, 14] generates samples by iteratively denoising random noise through a learned reverse diffusion process. Formally, we consider a latent representation z_0 of an initial clean data sample obtained from a variational autoencoder (VAE) that encodes the video into its latent space. A forward process q gradually adds Gaussian noise to z_0 over T time steps, producing a sequence $z_1, z_2, \ldots, z_T$. At each step, noise is added according to a variance schedule $\beta_{t=1}^{T}$. For example, a common formulation is:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}\, z_{t-1},\, \beta_t \mathbf{I}),\ \ t = 1, ..., T \tag{3}$$

with $\beta_t \in (0,1)$ controlling the noise level. As t increases, the latent z_t becomes increasingly noisy; in the limit z_T approaches a standard Gaussian noise. Equivalently, one can express z_t in closed form as a noisy mixture of the original latent and noise:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \varepsilon \tag{4}$$

where $\alpha_t := 1-\beta_t$, and $\bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$ is the accumulated noise attenuation. Here $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ represents standard Gaussian noise. The reverse process learns to invert this degradation by removing the noise from z_t to reconstruct the original latent representation. The denoiser is trained to predict the noise added to the latent, with respect to the loss function:

$$\mathcal{L} = \mathbb{E}_{z_0,\, \varepsilon,\, t}\Big[\big\| \varepsilon - \varepsilon_\theta(z_t,\, t) \big\|_2^2 \Big] \tag{5}$$

By minimizing this objective, the model learns to remove noise and recover the clean latent z_0 from a noisy z_t.

3.2 Latent-Space World Modeling for Egocentric HOI

We model egocentric hand–object interaction (HOI) synthesis as action-driven world modeling in a compact latent state space. Given the first-frame observation $\mathbf{I_1} \in \mathbb{R}^{H\times W\times 3}$, we obtain an initial latent state z_1 via the VAE encoder. The action signal at each step t is specified by hand kinematics $\mathbf{H}^t$ and metric head motion $\mathbf{C}^t$, forming an L-step action sequence $\{(\mathbf{H}^t, \mathbf{C}^t)\}_{t=1}^{L}$. Our goal is to predict the temporal evolution of latent states under actions and decode them into future first-person observations. Concretely, the world model defines action-driven latent dynamics that evolve the state over time,

$$z_{t+1} = \mathbf{f}_\theta(z_t,\, \mathbf{H}^t,\, \mathbf{C}^t), \qquad t = 1, \ldots, L, \tag{6}$$

and the predicted observation rollout is obtained by decoding $\{z_t\}$ into $\hat{\mathbf{I}}_{2:L+1}$.

Modeling egocentric HOI dynamics in latent space is challenging due to the coupled effects of high-DoF hand articulation, metric ego motion, and contact-driven object dynamics. We regularize the action-driven latent evolution without relying on privileged future object states. The next section introduces physics-informed embeddings distilled from 3D geometric and kinematic estimates, which promote physically plausible latent dynamics.

3.3 Physics-Informed Embeddings

To constrain HOI rollouts without using privileged future object states, we distill 3D geometric and kinematic estimates into compact physics-informed embeddings. Concretely, we derive three embeddings from complementary sources of structure: (i) hand kinematics that describe high-DoF articulation, (ii) metric ego motion that encodes the viewpoint trajectory, and (iii) an object-entity anchor tied to the first frame to stabilize object-centric appearance under interaction. These embeddings are injected into the latent dynamics to provide per-step geometric and kinematic context, thereby guiding state evolution toward physically plausible dynamics.

Hand Kinematic Embeddings (HKE). The goal of HKE is to capture dense hand motion structure from reconstructed 3D hand estimates, which is essential for contact-consistent HOI rollouts under high-DoF articulation. Rather than representing actions as sparse coordinate vectors, we use explicit per-frame hand renders obtained from reconstructed meshes as a dense 3D representation. We stack a sequence of hand render frames $\{H_t\}_{t=1}^{L}$ and encode them with a temporal 3D convolutional stack with SiLU activations, which gradually increases the channel dimension and maps the hand stream to the shared the 5120-dimension latent space. To stabilize hand identity and reduce temporal drift, we add a reference-hand branch that encodes the first-frame hand pose $\mathbf{P}^1$ into a compact feature $\mathbf{R}_{\text{ref}} \in \mathbb{R}^{H'' \times W'' \times 20}$ via a six-layer 2D convolutional pyramid (3×3 Conv + SiLU). Together, the temporal 3D Conv stack and the reference encoder produce hand kinematic embeddings that preserve fine-grained articulation structure and support physically plausible hand evolution during rollouts.

Ego-Motion Embeddings (EME). Ego head motion is a dominant factor in first-person rollouts, where viewpoint changes couple tightly with hand motion and occlusions. EME encodes metric head trajectory estimates into embeddings aligned with the latent space, enabling geometry-consistent viewpoint evolution over time. Given a head pose sequence $\{\mathbf{C}^t\}_{t=1}^{L}$ with intrinsics $\mathbf{K} \in \mathbb{R}^{3\times3}$ and extrinsics $\mathbf{E} = [\mathbf{R} : \mathbf{t}] \in \mathbb{R}^{3\times4}$ (relative to the first frame), we adopt Plücker embeddings to represent the per-pixel ray field induced by the head pose. For each pixel (u, v), its embedding is $\mathbf{p}_{u,v} = (\mathbf{o} \times \mathbf{d}_{u,v}, \mathbf{d}_{u,v}) \in \mathbb{R}^6$, where $\mathbf{o}$ is the head-frame optical center and $\mathbf{d}_{u,v}$ is the world-direction ray computed as:

$$\mathbf{d}_{u,v} = \mathbf{R}\,\mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} + \mathbf{t}. \tag{7}$$

This ray-based representation is invariant to the choice of scene origin and preserves metric structure under head motion. By representing ego motion as an image-aligned ray field, it avoids collapsing geometry into a low-dimensional pose vector and instead imposes spatially resolved constraints on viewpoint evolution. We stack per-frame Plücker tensors into a trajectory representation, process it with causal 3D convolutions and hybrid spatiotemporal blocks that combine 2D residual convolutions with temporal self-attention, and finally use a projection head and patchifier to map features into the latent grid, yielding ego-motion embeddings that regularize long-horizon viewpoint dynamics.

Object Entity Embeddings (OEE). Egocentric rollouts can suffer from object entity drift, which often appears as instability in color, texture, and fine details under rapid ego motion and partial occlusions. To stabilize object entities over time, OEE anchors each object to the first frame, providing a persistent reference that supports coherent HOI evolution across the rollout. Concretely, we use the first-frame object segmentation image, obtained from an off-the-shelf

segmentation model, to isolate object-centric evidence and reduce background interference under viewpoint changes. The segmentation is encoded by the frozen video VAE encoder into a latent grid, and a 3D convolutional patchifier converts it into tokens aligned with the main latent stream. Operating in the same latent parameterization as the backbone, OEE yields features that mitigate drift and preserve object-centric details throughout the rollout.

3.4 Latent Space Integration

We incorporate physics-informed embeddings into the frozen Wan-DiT in a single token-space forward pass by first aligning each embedding stream to the main token layout and shared token dimension d (as shown in Fig. 2). Let the Wan-VAE latent be $z \in \mathbb{R}^{B\times F\times H\times W\times C}$ and the corresponding main token sequence be $\mathbf{T}_{\text{main}} \in \mathbb{R}^{B\times N\times d}$. Each embedding tensor is projected and patchified into an aligned token sequence compatible with $\mathbf{T}_{\text{main}}$, enabling Wan-DiT to leverage physics-informed structure for latent evolution without modifying the original backbone architecture.

Action-Driven Token. The hand encoder produces hand tokens $\mathbf{T}^{\text{HKE}}$ in $\mathbb{R}^{B\times N\times d}$, which are fused into the main token stream via a gated residual update:

$$\mathbf{X}_0 = \mathbf{T}_{\text{main}} + \gamma_h \mathbf{T}^{\text{HKE}} \tag{8}$$

where γ_h is a learnable gate. In parallel, the reference-hand branch produces a first-frame hand feature that is added to the first-frame latent and concatenated with the noise latent before patchification, tying the initial hand configuration to the latent state space. The ego-motion adapter outputs geometry tokens $\mathbf{T}^{\text{EME}} \in \mathbb{R}^{B\times N\times d}$. For compatibility with the extended sequence length introduced below, we zero-pad $\mathbf{T}^{\text{EME}}$ to obtain with the same length as the active token sequence. In the first D DiT blocks, we apply per-block zero-initialized linear adapters $\mathbf{U}_l : \mathbb{R}^d \to \mathbb{R}^d$ to generate residuals:

$$\Delta\mathbf{X}_l = \mathbf{U}_l\big(\mathbf{T}^{\text{EME}}\big) \tag{9}$$

which are added to the block input. This construction regularizes viewpoint evolution under metric head motion while preserving backbone stability at initialization.

Object-Entity Token. A object patchifier converts first-frame object entity features into tokens $T^{\text{OEE}} \in \mathbb{R}^{B\times N\times d}$. We assign shifted RoPE to spatially anchor these tokens as an independent reference stream, then append them to form the extended sequence

$$T_{\text{all}} = [X_0;\, T^{\text{OEE}}] \tag{10}$$

Within each DiT block, the object-entity tokens participate as KV-only context during self-attention; only the main tokens are decoded, which provides a persistent entity reference that reduces drift in long rollouts.

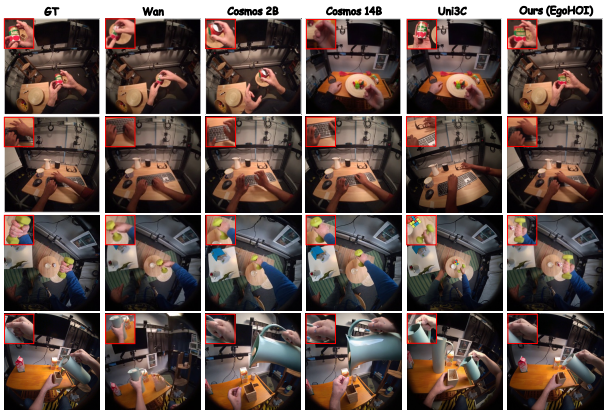

Fig. 3: Qualitative comparison with baselines. Columns from left to right show the ground truth (GT), Wan, Cosmos-2B, Cosmos-14B, Uni3C, and our EgoHOI model. All methods receive the same first frame as input; compared with the baselines, our model better preserves hand and object geometry, maintains object identity, and produces more stable interaction dynamics over time. See the zooming boxes for comparison of the fine-grained details.

First-Frame Anchor The first frame is encoded by the Wan VAE encoder as $\mathbf{Y}$, and the reference-hand feature is added to $\mathbf{Y}$ to couple initial kinematics and appearance. Before patchification, we concatenate $\mathbf{Y}$ with the noise latent along the channel dimension. In parallel, the first frame is processed by the CLIP image encoder [30] to obtain global image embeddings; these embeddings are linearly projected and provided as fixed keys and values to the cross-attention layers of Wan-DiT, encouraging preservation of overall scene structure and appearance consistency during denoising.

4 Experiment

4.1 Experiment Setup

Implementation Details We choose Wan 2.1 14B [34] as our model backbone. We start from the official checkpoint with LoRA rank set to 128 and LoRA alpha set to 128. The inference step and the learning rate are set to 50 and 1×10^{-5}. We use the Adam optimizer with BF16 precision. The cfg is set to 1.0. We train our model for 8000 steps on 16 NVIDIA H100 GPUs with a batch size of 1 and a sample resolution of 480×480. The training process lasts for approximately 1 day. During training and inference, our frame length is set to 81. After training, our model can achieve 16 FPS when simulating hand-object interaction.

Benchmark Since there is no publicly available benchmark for world modeling in egocentric hand–object interaction, we construct an evaluation benchmark

Table 1: Comparison with baselines. We compare EgoHOI against a Wan backbone, Cosmos world models (2B, 14B) and Uni3C on frame prediction (PSNR/SSIM/LPIPS/Object-CLIP), ego-motion consistency (ATE/RRE/RPE), and kinematic fidelity (MR/MPJPE/RMSE). ↑ higher is better, ↓ lower is better.

Method	Frame Prediction				Ego-Motion Consistency			Kinematic Fidelity		
	PSNR ↑	SSIM ↑	LPIPS ↓	Object-CLIP ↑	ATE ↓	RRE ↓	RPE ↓	MR ↓	MPJPE ↓	RMSE ↓
Wan [34]	14.78	0.46	0.52	0.86	0.124	10.687	0.036	19.67%	0.049	0.083
Cosmos 2B [1]	15.49	0.56	0.41	0.87	0.116	9.515	0.026	16.23%	0.044	0.078
Cosmos 14B [1]	15.89	0.59	0.38	0.88	0.112	9.046	0.022	14.61%	0.041	0.074
Uni3C [4]	14.21	0.50	0.55	0.82	0.168	14.209	0.026	20.50%	0.073	0.089
Ours (EgoHOI)	**21.05**	**0.65**	**0.27**	**0.92**	**0.084**	**5.192**	**0.021**	**5.84%**	**0.014**	**0.044**

from HOT3D. We select 100 clips from scenes that are excluded from the training set. Each clip contains 150 frames and provides hand mesh renders and camera trajectories. Although our model operates on 81-frame rollouts, we make full use of each 150-frame clip by extracting multiple 81-frame sub-clips with a sliding-window scheme, where the window length is fixed to 81 frames and is shifted along the clip to form evaluation samples.

Metrics We evaluate the model from three complementary aspects: visual prediction, ego-trajectory consistency, and kinematic fidelity. Visual prediction is measured by PSNR, SSIM [37], LPIPS [45], and Object-CLIP [40]. We additionally report VBench [48] scores to assess perceptual and temporal quality, including Subject Consistency, Background Consistency, Motion Smoothness, Dynamic Degree, Aesthetic Quality, and Imaging Quality. Ego-motion consistency is quantified by ATE, RPE, and RRE between the estimated and ground-truth camera trajectories. Kinematic accuracy is computed with HaMeR [29] on both ground-truth and generated results, reporting Missing Ratio (MR), MPJPE, and RMSE on hand segmentation images. Details are in the supplementary material.

4.2 Comparison with Baselines

We compare EgoHOI with three tiers of baselines for egocentric HOI rollouts. Wan [34] serves as a backbone-level video generation baseline. Cosmos 2B and 14B [1] are general-purpose world models at two parameter scales. We also include Uni3C [4], which offers stronger camera and human motion control. Wan and Cosmos are fine-tuned from released checkpoints using the same data and protocol as EgoHOI, and all methods are evaluated on the same datasets, metrics, and pipeline. Several recent egocentric world models [33, 38] are omitted because their implementations are not publicly available at submission time, which prevents a reproducible evaluation under our protocol.

Table 1 reports frame prediction, ego-motion consistency, and kinematic fidelity. Scaling from Wan to Cosmos yields modest gains in frame quality, but

motion-related performance remains limited. For example, Cosmos 14B reaches ATE 0.112 and MR 14.61%. Uni3C improves motion control, yet it degrades under rollout evaluation, with the largest ego-motion errors and weakest kinematic fidelity (ATE 0.168, RRE 14.209, MR 20.50%), suggesting that motion control alone does not recover contact-driven interaction dynamics. In contrast, EgoHOI performs best across all metric groups. It increases PSNR to 21.05 while substantially reducing ego-motion and hand kinematic errors, indicating that our physics-informed embedding extraction and injection improve interaction dynamics and temporal physical consistency, rather than only per-frame quality. Additional object-level comparisons are provided in the supplementary material, including OPE and OOE from our ablation study.

For VBench in Tab. 2, EgoHOI achieves the highest subject consistency (95.51%), background consistency (94.91%), aesthetic quality (52.03%), and imaging quality (64.48%). Its lower Dynamic Degree (53.29%) reflects a design choice that prioritizes stable hand–object interactions over highly dynamic camera motion. Qualitatively (Fig. 3), Wan exhibits drift; Cosmos produces sharper frames but shows contact-region errors and small-object inconsistencies; Uni3C constrains camera motion but remains unreliable around contact. Overall, changing backbones, scaling model size, or enforcing motion control does not address the core challenge of egocentric HOI world models, whereas EgoHOI regularizes action-driven dynamics with physics-informed embeddings distilled from 3D priors to yield more plausible rollouts. More visualization is included in the supplementary material.

Table 2: VBench Evaluation of the Baselines. Reported scores cover subject consistency, background consistency, motion smoothness, dynamic degree, aesthetic quality, and imaging quality. For the consistency and perceptual quality metrics, higher values correspond to better results, whereas dynamic degree characterizes how dynamic the generated content is rather than its realism.

Method	Subject Consistency	Background Consistency	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
Wan [34]	91.40%	93.14%	99.19%	64.75%	46.57%	58.58%
Cosmos 2B [1]	89.01%	93.00%	99.27%	87.50%	46.88%	62.98%
Cosmos 14B [1]	93.21%	93.56%	99.32%	**89.50%**	49.83%	63.78%
Uni3C [4]	91.91%	94.29%	99.33%	49.00%	47.03%	61.47%
Ours (EgoHOI)	**95.51%**	**94.91%**	**99.40%**	53.29%	**52.03%**	**64.48%**

4.3 Ablation Study

We provide a detailed ablation analysis of EgoHOI by isolating how each component supports egocentric hand–object rollouts along three aspects: kinematic fidelity, object entity stability, and ego-motion consistency. Across kinematic fidelity (Tab. 3), ego-motion consistency (Tab. 4), and object integrity (Tab. 5), we compare several configurations of EgoHOI. Base uses the Wan 2.1 DiT video backbone without HKE, OEE, or EME. The HKE Injection, OEE Injection,

Fig. 4: Qualitative results for ablation on kinematic fidelity. Rows: GT, Base, HKE Injection, Ours (EgoHOI). HKE improves hand articulation stability and reduces local shape distortion. Zoom-in boxes highlight clearer hand–object contact details. The full model is the most stable, matching MR/MPJPE/RMSE.

Table 3: Ablation on kinematic fidelity. We ablate HKE Injection and report hand motion fidelity in generated videos using MR, MPJPE, and RMSE; ↓ indicates lower is better.

Method	MR ↓	MPJPE ↓	RMSE ↓
Ours (base)	28.77%	0.576	0.089
HKE Injection	6.47%	0.015	0.044
Ours (EgoHOI)	**5.84%**	**0.014**	**0.044**

and EME Injection variants each enable a single component on top of this backbone, which isolates its contribution under otherwise identical settings. EgoHOI activates three components jointly and corresponds to the model used in the main experiments. All variants share the same data, training and optimization settings. We measure kinematic fidelity with MR, MPJPE, and RMSE, object entity with Object-CLIP together with VBench Subject Consistency and Imaging Quality, and ego-motion consistency with ATE, RPE, and RRE.

Study on Kinematic Fidelity Tab. 3 examines the effect of HKE on hand state evolution over egocentric rollouts. Compared with the base variant, adding HKE reduces the MR and MPJPE to 6.47% and 0.015 from 28.77% and 0.576, respectively, indicating more reliable hand localization and substantially improved joint accuracy under interaction and partial occlusions. The reduction in RMSE further suggests that the hand region matches the ground truth more closely in both shape and RGB appearance over the sequence. The full model, which combines HKE with OEE and EME, further improves these metrics to 5.84%, 0.014, and 0.044. Qualitatively (Fig. 4), HKE mitigates hand-shape collapse and temporal fluctuations, while the full model yields the most stable hand evolution across time with fewer distortions during frequent hand–object interaction.

Table 4: Ablation on ego-motion consistency. We evaluate the EME injection variant on long-horizon egocentric rollouts using ATE, RRE, and RPE, and compare it with the base backbone and the full EgoHOI model; ↓ indicates more consistent ego-motion trajectories and reduced accumulated drift.

Method	ATE ↓	RRE ↓	RPE ↓
Ours (base)	0.133	15.249	0.039
EME Injection	0.096	6.525	0.023
Ours (EgoHOI)	**0.084**	**5.192**	**0.021**

Fig. 5: Qualitative results for ablation of ego-motion consistency Rows: GT, Base, EME Injection, Ours (EgoHOI). EME mitigates long-horizon ego-motion drift and improves viewpoint stability. Zoom-ins highlight cleaner details under viewpoint changes. The full model best matches ATE/RRE/RPE gains.

Study on Ego-Motion Consistency Tab. 4 and Fig. 5 examine ego-motion consistency over egocentric rollouts using ATE, RRE, and RPE. Compared with the base variant, enabling EME reduces these ego-motion errors and alleviates accumulated viewpoint drift over time. The full model achieves the best ego-motion metrics, reaching ATE/RRE/RPE of 0.084/5.192/0.021, which indicates closer agreement with the ground-truth viewpoint changes. Qualitatively (Fig. 5), EME injection yields more stable viewpoints with less drift across time, and the full model further improves global temporal coherence. Overall, improving ego-motion consistency is important for egocentric rollouts, and our ablation shows that EME is a key factor in reducing ATE/RRE/RPE.

Study on Object Integrity. Tab. 5 examines the effect of OEE on object integrity over egocentric rollouts. Compared with the base variant, OEE injection improves Object-CLIP from 0.81 to 0.83 and increases VBench Subject Consistency and Imaging Quality from 85.26% and 52.55% to 87.90% and 54.75%, respectively, indicating more stable object appearance over time. We additionally report object pose errors in this ablation: object position error (OPE) and 2D orientation error (OOE) computed from object masks (details are included

Table 5: Ablation on object integrity. We evaluate the OEE injection variant on egocentric rollouts, where object integrity refers to preserving object identity and maintaining stable object pose under interaction. We report Object CLIP, Subject Consistency, and Imaging Quality (↑ is better), together with object position error (OPE) and 2D orientation error (OOE) computed from object masks (↓ is better).

Method	Object CLIP ↑	Subject Consistency ↑	Imaging Quality ↑	OPE ↓	OOE ↓
Ours (base)	0.81	85.26%	52.55%	0.141	27.739
OEE Injection	0.83	87.90%	54.75%	0.131	24.124
Ours (EgoHOI)	**0.92**	**95.51%**	**64.48%**	**0.015**	**9.412**

Fig. 6: Qualitative results for ablation of object integrity. Rows: GT, Base, OEE Injection, Ours (EgoHOI). OEE reduces texture flicker and appearance drift, preserving object-specific details through interaction and occlusion. Zoom-ins highlight more stable patterns and edges. The full model yields the most consistent results.

in supplementary materials). OEE injection reduces OPE from 0.141 to 0.131 and OOE from 27.739 to 24.124. The full model further improves these metrics to 0.92, 95.51%, and 64.48%, while substantially reducing OPE/OOE to 0.015/9.412. Qualitatively (Fig. 6), OEE injection mitigates drift in object appearance and reduces pose deviations, and the full model yields the most coherent object evolution during periods of frequent hand–object interaction.

5 Conclusion

We introduce EgoHOI, an egocentric world model for photorealistic, action-driven hand–object interaction synthesis. EgoHOI generates physically plausible HOI rollouts from user actions while preserving object identity and realistic interactions. On HOT3D, EgoHOI improves visual quality, ego-motion consistency, and kinematic fidelity, with ablations validating the roles of physics-informed embeddings and first-frame appearance. Its egocentric HOI rollouts further suggest potential applications in VLA pre-training, latent action modeling, policy evaluation, and egocentric video understanding.

Acknowledgments. This work used computational resources provided by the Texas A&M VISION GPU Cluster and was partially supported by the U.S. National Science Foundation under Grant No. 2527316. Zhiwen Fan was supported in part by gifts from Meta and the US AI Glasses Impact Grants program.

References

1. Agarwal, N., Ali, A., Bala, M., Balaji, Y., Barker, E., Cai, T., Chattopadhyay, P., Chen, Y., Cui, Y., Ding, Y., et al.: Cosmos world foundation model platform for physical ai. arXiv preprint arXiv:2501.03575 (2025)
2. Bai, Y., Tran, D., Bar, A., LeCun, Y., Darrell, T., Malik, J.: Whole-body conditioned egocentric video prediction. arXiv preprint arXiv:2506.21552 (2025)
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. CoRR (2023)
4. Cao, C., Zhou, J., Li, S., Liang, J., Yu, C., Wang, F., Xue, X., Fu, Y.: Uni3c: Unifying precisely 3d-enhanced camera and human motion controls for video generation. arXiv preprint arXiv:2504.14899 (2025)
5. Chen, D., Moutakanni, T., Chung, W., Bang, Y., Ji, Z., Bolourchi, A., Fung, P.: Planning with reasoning using vision language world model. arXiv preprint arXiv:2509.02722 (2025)
6. Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauser, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
7. Dai, W., Lan, K., Zhou, J., Zhao, B., Su, X., Tong, J., Guan, W., Yang, S.: Conla: Contrastive latent action learning from human videos for robotic manipulation. arXiv preprint arXiv:2602.00557 (2026)
8. Diller, C., Dai, A.: Cg-hoi: Contact-guided 3d human-object interaction generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19888–19901 (2024)
9. Fan, Y., Yang, Q., Wang, K., Zhou, H., Li, Y., Feng, H., Ding, E., Wu, Y., Wang, J.: Re-hold: Video hand object interaction reenactment via adaptive layout-instructed diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17550–17560 (2025)
10. Gavryushin, A., Wang, X., Malate, R.J., Yang, C., Liconti, D., Zurbrügg, R., Katzschmann, R.K., Pollefeys, M.: Maple: Encoding dexterous robotic manipulation priors learned from egocentric videos. arXiv preprint arXiv:2504.06084 (2025)
11. Ghosh, A., Dabral, R., Golyanik, V., Theobalt, C., Slusallek, P.: Imos: Intent-driven full-body motion synthesis for human-object interactions. In: Computer Graphics Forum. vol. 42, pp. 1–12. Wiley Online Library (2023)
12. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., 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. pp. 18995–19012 (2022)
13. Hassan, M., Stapf, S., Rahimi, A., Rezende, P., Haghighi, Y., Brüggemann, D., Katircioglu, I., Zhang, L., Chen, X., Saha, S., et al.: Gem: A generalizable ego-vision multimodal world model for fine-grained ego-motion, object dynamics, and scene composition control. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22404–22415 (2025)

14. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
15. Huang, J., Hao, S., Hu, B., Wang, G.: Understanding dynamic scenes in ego centric 4d point clouds. arXiv preprint arXiv:2508.07251 (2025)
16. Huang, M., Chu, F., Tekin, B., Liang, K.J., Ma, H., Wang, W., Chen, X., Gleize, P., Xue, H., Lyu, S., Kitani, K., Feiszli, M., Tang, H.: Hoigpt: Learning long-sequence hand-object interaction with language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7136–7146 (June 2025)
17. Huang, Z., Zhou, Z., Cao, J., Ma, Y., Chen, Y., Rao, Z., Xu, Z., Wang, H., Lin, Q., Zhou, Y., et al.: Hunyuanvideo-homa: Generic human-object interaction in multimodal driven human animation. arXiv preprint arXiv:2506.08797 (2025)
18. Jiang-Lin, J.Y., Huang, K.Y., Lo, L., Huang, Y.N., Lin, T., Wu, J.C., Shuai, H.H., Cheng, W.H.: Record: Reasoning and correcting diffusion for hoi generation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 9465–9474 (2024)
19. Keetha, N.V., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. In: ICCV 2025 Workshop on Wild 3D: 3D Modeling, Reconstruction, and Generation in the Wild
20. Lee, J., Saito, S., Nam, G., Sung, M., Kim, T.K.: Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 527–537 (2024)
21. Lepert, M., Fang, J., Bohg, J.: Masquerade: Learning from in-the-wild human videos using data-editing. arXiv preprint arXiv:2508.09976 (2025)
22. Li, J., Clegg, A., Mottaghi, R., Wu, J., Puig, X., Liu, C.K.: Controllable human-object interaction synthesis. In: European Conference on Computer Vision. pp. 54–72. Springer (2024)
23. Li, M., Christen, S., Wan, C., Cai, Y., Liao, R., Sigal, L., Ma, S.: Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17416–17425 (2025)
24. Lin, X., Lian, S., Yu, B., Yang, R., Wu, C., Miao, Y., Jin, Y., Shi, Y., Huang, C., Cheng, B., et al.: Physbrain: Human egocentric data as a bridge from vision language models to physical intelligence. arXiv preprint arXiv:2512.16793 (2025)
25. Liu, B., Gong, X., Zhao, Z., Song, Z., Lu, Y., Wu, S., Zhang, J., Banerjee, S., Zhang, H.: Byteloom: Weaving geometry-consistent human-object interactions through progressive curriculum learning. arXiv preprint arXiv:2512.22854 (2025)
26. Liu, X., Yi, L.: Geneoh diffusion: Towards generalizable hand-object interaction denoising via denoising diffusion. arXiv preprint arXiv:2402.14810 (2024)
27. Luo, H., Wang, Y., Zhang, W., Zheng, S., Xi, Z., Xu, C., Xu, H., Yuan, H., Zhang, C., Wang, Y., et al.: Being-h0. 5: Scaling human-centric robot learning for cross-embodiment generalization. arXiv preprint arXiv:2601.12993 (2026)
28. Pang, Y., Shao, R., Zhang, J., Tu, H., Liu, Y., Zhou, B., Zhang, H., Liu, Y.: Manivideo: Generating hand-object manipulation video with dexterous and generalizable grasping. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12209–12219 (June 2025)
29. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9826–9836 (2024)

30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). pp. 8748–8763. PMLR (2021)
31. Routray, S., Pan, H., Jain, U., Bahl, S., Pathak, D.: Vipra: Video prediction for robot actions. arXiv preprint arXiv:2511.07732 (2025)
32. Shen, Z., Wu, C., Zhou, J., Zhao, C., Wang, K., Zhou, H., Li, Y., Feng, H., He, W., Wang, J.: idit-hoi: Inpainting-based hand object interaction reenactment via video diffusion transformer. arXiv preprint arXiv:2506.12847 (2025)
33. Tu, Y., Luo, H., Chen, X., Bai, X., Wang, F., Zhao, H.: Playerone: Egocentric world simulator. arXiv preprint arXiv:2506.09995 (2025)
34. Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., et al.: Wan: Open and advanced large-scale video generative models. CoRR (2025)
35. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
36. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
37. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
38. Xie, L., Sun, L.C., Neall, A., Wu, T., Cai, S., Wetzstein, G.: Generated reality: Human-centric world simulation using interactive video generation with hand and camera control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3998–4008 (2026)
39. Xu, J., Huang, Y., Pei, B., Hou, J., Li, Q., Chen, G., Zhang, Y., Feng, R., Xie, W.: Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. In: The Thirteenth International Conference on Learning Representations (2025)
40. Xu, Z., Huang, Z., Cao, J., Zhang, Y., Cun, X., Shuai, Q., Wang, Y., Bao, L., Li, J., Tang, F.: Anchorcrafter: Animate cyberanchors saling your products via human-object interacting video generation. arXiv preprint arXiv:2411.17383 (2024)
41. Xue, M., Liu, Y., Guo, L., Huang, S., Ding, C.: Guiding human-object interactions with rich geometry and relations. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22714–22723 (2025)
42. Yang, R., Yu, Q., Wu, Y., Yan, R., Li, B., Cheng, A.C., Zou, X., Fang, Y., Cheng, X., Qiu, R.Z., et al.: Egovla: Learning vision-language-action models from egocentric human videos. arXiv preprint arXiv:2507.12440 (2025)
43. Yoshida, T., Kurita, S., Nishimura, T., Mori, S.: Developing vision-language-action model from egocentric videos. arXiv preprint arXiv:2509.21986 (2025)
44. Zhang, B., Shou, M.Z.: Ego-centric predictive model conditioned on hand trajectories. arXiv preprint arXiv:2508.19852 (2025)
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595 (2018)
46. Zhang, W., Foo, L.G., Beeler, T., Dabral, R., Theobalt, C.: Vhoi: Controllable video generation of human-object interactions from sparse trajectories via motion densification. arXiv preprint arXiv:2512.09646 (2025)

47. Zhang, Z., Shi, Y., Yang, L., Ni, S., Ye, Q., Wang, J.: Openhoi: Open-world hand-object interaction synthesis with multimodal large language model. arXiv preprint arXiv:2505.18947 (2025)
48. Zheng, D., Huang, Z., Liu, H., Zou, K., He, Y., Zhang, F., Gu, L., Zhang, Y., He, J., Zheng, W.S., et al.: Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. arXiv preprint arXiv:2503.21755 (2025)
49. Zheng, R., Niu, D., Xie, Y., Wang, J., Xu, M., Jiang, Y., Castañeda, F., Hu, F., Tan, Y.L., Fu, L., et al.: Egoscale: Scaling dexterous manipulation with diverse egocentric human data. arXiv preprint arXiv:2602.16710 (2026)
50. Zhou, B., Zhan, Y., Zhang, Z., Lu, Z.: Megohand: Multimodal egocentric hand-object interaction motion generation. arXiv preprint arXiv:2505.16602 (2025)
51. Zhou, S., Vilesov, A., He, X., Wan, Z., Zhang, S., Nagachandra, A., Chang, D., Chen, D., Wang, X.E., Kadambi, A.: Vlm4d: Towards spatiotemporal awareness in vision language models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8600–8612 (2025)