arXiv:2605.15477v1 [cs.CV] 14 May 2026

EgoExo-WM: Unlocking Exo Video for Ego World Models

Danny Tran[*1] **Roberto Martín-Martín**[†] **Kristen Grauman**[†]

The University of Texas at Austin

Abstract

Egocentric world models present a promising direction for enabling agents to predict and plan, but their performance is constrained by the limited availability of egocentric training data and its inherent partial observability of humans' physical actions. In contrast, exocentric video is abundant and reveals body poses well, but lacks direct alignment with an agent's action space—and is not egocentric. We propose a method to bridge this gap by extracting structured body pose from exocentric video as a representation of action and transforming the exocentric video to egocentric video, informed by a human kinematics prior. This process unlocks the integration of in-the-wild exocentric data for egocentric world model training. We show that training whole-body action-conditioned egocentric world models with our converted data significantly improves both prediction quality and downstream planning performance, where we infer the sequence of body poses needed to achieve a visual goal state. Our approach paves the way to enlist arbitrary in-the-wild videos for building powerful egocentric world models, furthering applications in robot planning and augmented-reality guidance. [1]

1 Introduction

Humans can learn about the world not only through direct experience, but also by observing others [7, 70]. We may learn to knead dough from a cooking video, study a friend's brushstrokes while painting, or watch a parent guide thread through the loops of a knitting pattern. From these exocentric observations, we can mentally imagine how the same interactions would unfold through our own eyes: where our hands would move, how objects would shift, resist, or deform under contact, and how each action would reshape the visual scene before us. This ability to transform observed behavior into anticipated first-person experience allows us to learn predictive structure about the world without directly performing every action ourselves [69].

Egocentric world models aim to provide agents with this first-person predictive understanding by modeling how future states evolve under embodied action. Such models could serve as a foundation for future systems that must anticipate the consequences of human motion, including robots [25, 4, 55, 27], wearable augmented reality agents that reason from a user's viewpoint [13], and assistive technologies that predict and support everyday interactions [19]. In human-centered egocentric settings, the relevant action space is naturally 3D human motion: egocentric observations change with the actor's movement and through interactions with the physical space, shaping egocentric observations.

† Indicates Equal advising.

[1] Project page: `https://vision.cs.utexas.edu/projects/EgoExo-WM/`.

Preprint.

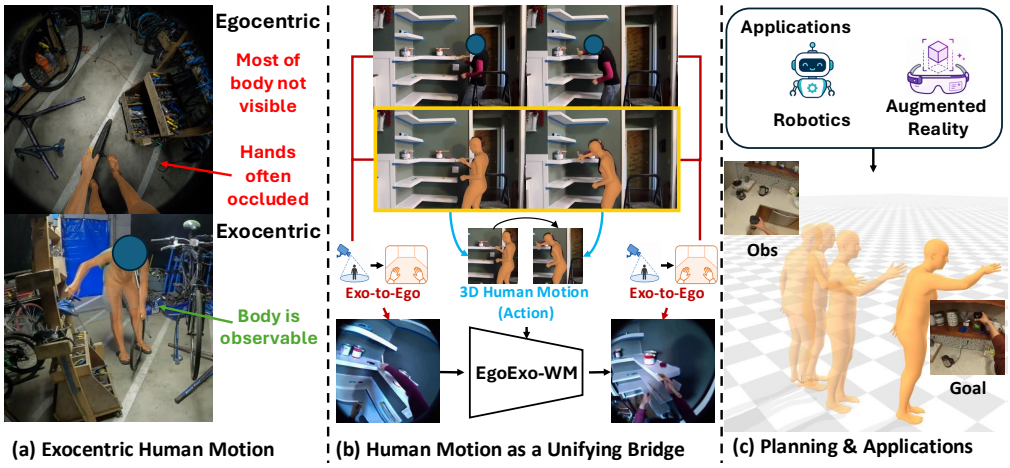

Figure 1: **Overview.** Egocentric video provides an embodied view but often hides the body and occludes the hands, while exocentric video often reveals full-body motion (a). EgoExo-WM uses recovered 3D human motion as a bridge for learning an egocentric world-model with exocentric video: it defines the action sequence and guides exo-to-ego synthesis into action-aligned egocentric observations (b). The learned world model enables goal-conditioned planning for applications such as robotics and augmented reality by selecting actions that best reach a target state (c).

Current egocentric world model training processes remain limited by existing egocentric datasets [6, 56, 71, 25]. However, learning such models from existing egocentric datasets alone creates a dual bottleneck. First, egocentric video is difficult to collect at scale because it requires people to wear cameras while performing diverse activities across many environments—today's largest egocentric captures [28, 18, 50] are still dwarfed by (largely exocentric) internet-scale video resources [52, 43, 11, 64]. Second, the 3D human motion needed to define the action space and condition the world model is difficult to obtain at scale: egocentric video alone provides only a very partial view of the body, and hands are frequently occluded during interaction. More fundamentally, an embodied agent should not be limited to learning only from experiences recorded through its own viewpoint. To build scalable world models, we need mechanisms that can convert arbitrary observations of human behavior into useful first-person predictive supervision.

Taking inspiration from human observational learning, we introduce EgoExo-World Model (WM), a framework for leveraging *exocentric* videos to train *egocentric* world models. See Fig. 1. Exocentric video offers a complementary source of supervision: it exists at internet scale, captures diverse activities and environments, and often reveals the 3D human motion hidden in egocentric views. Our key insight is that 3D human motion can unify exocentric observation, world model learning, and egocentric video synthesis. We extract 3D human motion from exocentric video and use it as the action space for world model training and to guide exo-to-ego synthesis. Building on EgoX [41], our viewpoint conversion module incorporates body motion priors so that synthesized egocentric videos are not only visually plausible, but also aligned with the underlying human action.

Using this approach, we convert diverse exocentric videos from internet sources such as HowTo100M [52], CrossTask [81], and 100 Days of Hands [64] into paired egocentric visual observations with 3D human motion sequences. We then inject this converted data into training a human action-conditioned egocentric world model, together with existing egocentric data. Given a history of egocentric observations and a 3D human motion, our world model predicts the resulting future egocentric visual state. Together, these components form EgoExo-WM, a framework that demonstrates how egocentric WMs can move beyond purely egocentric training by incorporating heterogeneous exocentric observations as diverse visual experience and human-action supervision.

Evaluating on three diverse egocentric datasets (Home Action Genome [62], LEMMA [39], and Ego-Exo4D [29]), we demonstrate that EgoExo-WM consistently outperforms SOTA egocentric world-model baselines in future rollout predictions. Beyond prediction, EgoExo-WM also enables stronger downstream planning: given contextual egocentric observations and a target visual goal state, the learned world model can select the sequence of 3D human motions that best drives the agent toward the desired visual outcome—e.g., how the person would need to move within the scene

and/or manipulate objects in order to achieve a target state (Fig. 1(c)). Together, these results show how EgoExo-WM can unlock exocentric video as a scalable source of visual experience and action supervision for egocentric world modeling.

2 Related Work

World Models. World Models are predictive models that enable agents to predict future observations and reason about the consequences of actions, making them a promising paradigm for embodied intelligence. They have been used for generating synthetic training data [38, 2], improving policies [61, 21, 32], and learning action-conditioned world representations for planning and control [25, 4, 8, 33, 31, 30]. A growing line of work explores egocentric video due to its alignment with the first-person perceptual inputs and onboard sensors used for control [28, 80, 42]. Building on this motivation, recent *egocentric world models* learn predictive dynamics from wearable camera videos, often conditioning on explicit body, hand, or head motion to capture interaction dynamics relevant to humanoid control and dexterous manipulation [71, 6, 56, 27, 24]. However, these approaches remain limited by the scarcity of egocentric data with accurate, low-level action annotations—and body pose visibility is inherently restricted in egocentric views. Some works avoid explicit actions by learning latent actions [25] or conditioning on text descriptions [76], but such action representations are difficult to interpret or too coarse to capture the fine-grained kinematics needed for precise embodied control. Thus, a central challenge remains: how to scale egocentric world modeling?

To address this challenge, we scale egocentric world modeling with abundant exocentric video while preserving physically grounded supervision. Our model extracts action signals from 3D human pose, providing a richer and more interpretable representation of agent dynamics than latent or low-dimensional actions, and it introduces an auxiliary wrist-position consistency loss to keep the learned dynamics grounded in the agent's pose rather than only future visual appearance. Most importantly, we show that transforming exo data to ego unlocks an existing resource for WM training, translating third-person observations into first-person experience.

Exo-to-Ego Generation. Exocentric-to-egocentric view translation synthesizes egocentric observations from exocentric videos, but remains challenging due to large viewpoint gaps, occlusions, and ambiguity in human motion. Prior work addresses this with stronger supervision, such as the first egocentric frame [73] or synchronized multi-view exocentric videos [47], and with geometric representations that improve cross-view consistency [17, 41, 9, 10, 53, 14, 60, 15]. However, these methods largely emphasize scene-level alignment; methods that transfer human information typically rely on egocentric hand priors [49, 73] rather than full-body dynamics. EgoX [41] is a strong framework combining 4D reconstruction [36], video generation priors [72], and VLM-based conditioning [1], but its lack of an explicit human prior can lead to action misalignment, including incorrect hand-object interactions, swapped hands, or hallucinated arm motion, as shown in Figure 3. EgoExo-WM builds on EgoX by adding full-body human priors, which provide kinematic cues for projecting hands into the egocentric view consistently with the estimated body pose. We use this conversion step to unlock exocentric videos as training data for egocentric world models. While we observe it to have good precision advantages in practice, exo-to-ego conversion itself is not our technical focus, and other conversion methods could also be incorporated into our framework [17, 41, 49].

Human Pose Estimation. Human pose estimation aims to recover 2D or 3D human structure from monocular images or videos, including body and hand motion [20, 26, 66, 78, 74, 59]. These methods either estimate skeletons directly [46] or fit parametric models such as SMPL [48, 58], MANO [63], and MHR [22]. Recent work has made substantial progress in improving the accuracy, robustness, and temporal consistency of these predictions across diverse scenes and motions [23, 75, 12, 79]. By converting raw visual observations into compact, semantically meaningful structure, pose provides a useful human prior for downstream tasks such as motion prediction [3, 57, 16] and robot learning [45, 5, 68]. We leverage the human pose present in exocentric videos as a prior in two ways: to guide ego conversion and to serve as the action space for our world model.

3 EgoExo-WM

In this section we introduce our EgoExo-WM. We start by presenting our world model (Section 3.1), followed by our method for converting exocentric videos to egocentric videos (Section 3.2). Finally, we describe how we can plan with EgoExo-WM by ranking action sequences (Section 3.3).

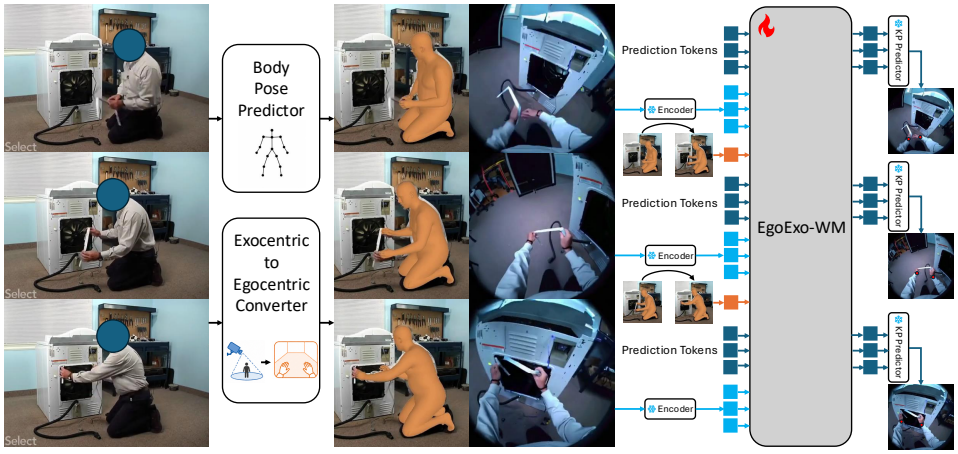

Figure 2: **World Model Training.** EgoExo-WM unlocks exocentric video for egocentric world model training. Given an exocentric video, we recover 3D human motion which we use alongside the original video to ground our exo-to-ego conversion. The 3D human motion becomes our actions and the converted exocentric video becomes the egocentric observation. We then train EgoExo-WM autoregressively with teacher forcing. We apply a lightweight wrist keypoint predictor on predicted latents and use a wrist keypoint consistency loss to preserve relevant pose cues.

3.1 World Model Training

We formulate EgoExo-WM as a human motion conditioned egocentric world model. Given a history of egocentric states and a 3D human motion, the model learns to predict how the egocentric state evolves after executing that action.

Let $\mathbf{x}_t$ denote the egocentric observation at time t, and let $\mathbf{a}_t$ denote the 3D human motion action that moves the agent from time t to $t+1$. We represent $\mathbf{a}_t$ as a structured vector that captures both global body motion and local joint articulation. Specifically, $\mathbf{a}_t \in \mathrm{R}^N$ consists of the change in root translation and the relative rotations of 22 body joints, with each joint rotation parameterized by Euler angles, and thus $N = 3 + 22 \times 3 = 69$.

We train the world model in a visual latent space. Pixel-level prediction is expensive and can overemphasize low-level appearance details, whereas latent prediction provides a more compact target that preserves semantic and geometric information relevant for future egocentric observations. Specifically, we use a pretrained DINOv3-L [67] encoder E to map each observation $\mathbf{x}_t$ to a visual latent $\mathbf{z}_t = E(\mathbf{x}_t)$. Given a fixed-length context window of H past visual latents $\mathbf{z}_{t-H+1:t}$ and the next 3D human motion action $\mathbf{a}_t$, the world model f_θ predicts the next visual latent,

$$\hat{\mathbf{z}}_{t+1} = f_\theta(\mathbf{z}_{t-H+1:t}, \mathbf{a}_t).$$

We supervise this prediction by encoding the future egocentric observation $\mathbf{x}_{t+1}$ with the same visual encoder and regressing to its latent representation $\mathbf{z}_{t+1} = E(\mathbf{x}_{t+1})$ using $\mathcal{L}_{\text{latent}} = \|\hat{\mathbf{z}}_{t+1} - \mathbf{z}_{t+1}\|_2^2$.

Because human pose is central to grounding egocentric dynamics, we encourage the predicted future state to preserve relevant pose cues by adding an auxiliary wrist-position consistency objective, inspired by prior work [27, 49, 73]. In our implementation, a lightweight head h_ϕ predicts a wrist heatmap $\hat{\mathbf{V}}_{t+1} = h_\phi(\hat{\mathbf{s}}_{t+1})$ from the predicted latent state, which is supervised using wrist pseudo-labels $\mathbf{V}_{t+1}$ extracted from the corresponding training frame. The full training objective is thus

$$\mathcal{L} = \underbrace{\|\hat{\mathbf{z}}_{t+1} - \mathbf{z}_{t+1}\|_2^2}_{\mathcal{L}_{\text{latent}}} + \lambda \underbrace{\|\hat{\mathbf{V}}_{t+1} - \mathbf{V}_{t+1}\|_2^2}_{\mathcal{L}_{\text{wrist}}}.$$

where λ is used to scale the wrist position consistency loss; we find $\lambda = 1$ works well.

This formulation enables joint training on real egocentric and converted exocentric videos in a shared observation-action format. Real egocentric datasets require external sensing for full-body motion, such as the motion-capture suit used in Nymeria [50]. For converted exocentric videos, we instead estimate 3D human motion directly from the exocentric view and pair it with synthesized

egocentric observations. Critically, the synthesized egocentric videos complement the real egocentric videos with a broader range of activities, scenes, and interactions. Concretely, this means we can take a specialized dataset like Nymeria and join it with more in-the-wild, broadly available internet video like HowTo100M [52] or 100 Days of Hands [64]. Figure 2 overviews how we incorporate synthesized egocentric videos.

3.2 Exo-to-Ego Data Generation

Synthesizing egocentric video from an exocentric viewpoint is inherently underdetermined, since the exocentric camera cannot always observe the objects or fine-grained interactions that appear in the actor's egocentric view. To address this, we develop EgoX-Body, a method designed to bridge this gap by grounding the generation process in explicit human kinematic structure.

Building on Ego-X [41], EgoX-Body incorporates the kinematics at two complementary levels. On the exocentric side, while the input video contains rich motion data, it is often entangled with viewpoint and background. We introduce kinematic conditioning by overlaying human skeletons (SAM 3D Body representation [23, 75]) onto the conditioning frames. This provides a structured, canonical representation of behavior that grounds the synthesis in the actor's physical movement.

On the egocentric side, we introduce an egocentric hand kinematics conditioning to directly reflect the interactions that define egocentric video. We condition the model with a drawn hand-skeleton overlay that exposes hand kinematics, helping generate consistent hand motion and interaction-relevant cues.

During training, we instantiate hand kinematics using ground-truth poses extracted with HaMeR [59]. At inference time, we reconstruct the hand kinematics directly from the exocentric input by projecting the 3D hand positions from the estimated body pose into this rendered view, creating a conditioning signal that aligns the generated egocentric video with the observed human motion. This dual-prior approach ensures that EgoX-Body generates videos that are not only visually plausible but faithfully represent the underlying human action. The overall inference pipeline can be seen in Figure 4. A qualitative sample comparing EgoX and EgoX-Body can be seen in Figure 3. More qualitative samples can be seen in the Supplementary Video.

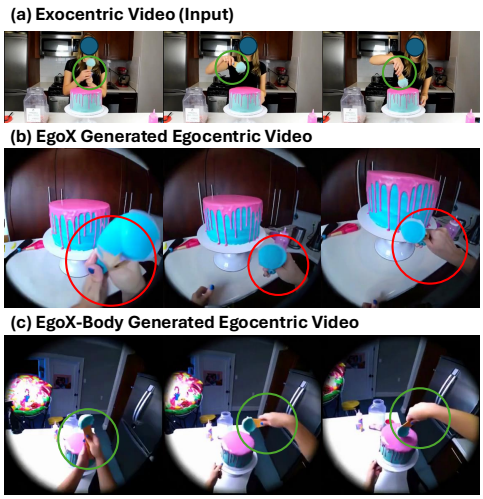

Figure 3: **EgoX-Body Qualitative Comparison.** EgoX-Body better grounds generated egocentric video in human motion and interaction structure.

3.3 Planning with EgoExo-WM

Having introduced our world model and the viewpoint conversion model, we now explain how it can be used for planning. A key advantage of an action-conditioned egocentric world model is that it can be used not only to predict future observations, but also to choose actions that make progress toward a desired future state. For example, a robot could given the textual goal 'load my laundry' along with a *visual goal image* of the laundry in the washing machine and subsequently plan a sequence of actions that reach that goal [21, 4, 8, 34, 77]. Similarly, in an augmented reality coaching setting, a visual goal could depict a soccer ball reaching the top corner of the soccer goal net. Given the user's current observation, the planner could determine a human motion sequence that is predicted to achieve this outcome, and then provide guidance on how to position the support foot, align the body, strike the ball, and follow through from the user's current position.

We instantiate this idea with a motion predictive control (MPC)-style planning procedure over candidate human motion sequences. Given the current egocentric observation and a visual goal, we sample candidate action sequences, rollout each sequence forward with our learned world model, and select the sequence whose predicted future state is closest to the goal. To obtain realistic and physically plausible action candidates, we use UniEgoMotion [57] as an action proposal model. While other methods could be used [65, 40], UniEgoMotion provides temporally consistent sequences of

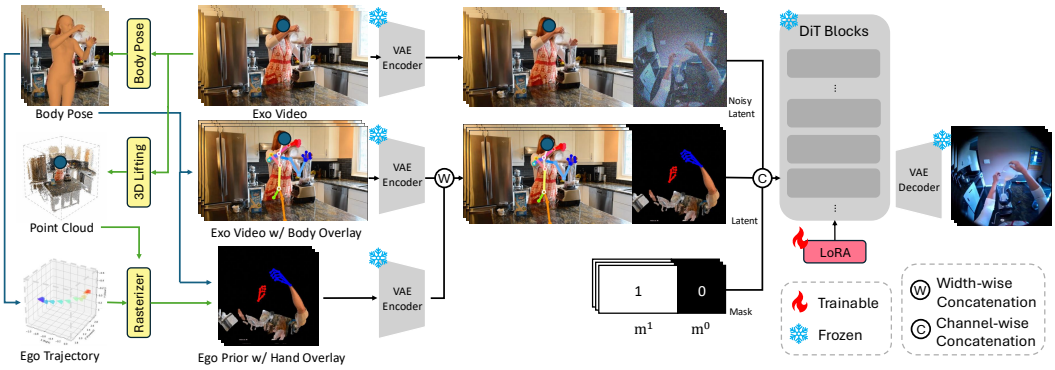

Figure 4: **EgoX-Body Inference Overview.** From exocentric videos, we extract body pose and lift the scene into a 3D point cloud. The body skeleton is overlaid onto the exocentric video, while the same pose and geometry are used to render an egocentric prior with predicted hand locations. We form two latent inputs: (1) the clean exocentric latent concatenated with noise, and (2) the body-overlaid exocentric latent concatenated with the egocentric hand prior. These are combined channel-wise with a mask and fed into a video diffusion model to generate the egocentric video. Incorporating these human kinematic priors produces outputs that are better grounded in human actions (Fig. 3).

whole-body actions grounded on a current egocentric observation, bounding the sampling space to human-like motions for our EgoExo-WM-based MPC method to evaluate and select the most likely to achieve the desired egocentric outcome.

At each planning step, we sample N candidate action sequences $\{a^{(i)}_{t:t+H}\}_{i=1}^{N}$ over a horizon H. For each candidate, we simulate its future outcome by autoregressively rolling out our learned world model in latent space:

$$z^{(i)}_{t+1:t+H} = f_\theta(z_t, a^{(i)}_{t:t+H}),$$

where z_t is the current latent observation. We assume access to a desired goal image, which in practice would be provided by a user-captured image, selected from a demonstration video, or generated from a high-level language instruction describing the desired outcome. We encode the goal image as a target latent z_g and score each candidate sequence by its closest predicted rollout state to the goal, $\mathcal{C}^{(i)} = \|z^{(i)}_{t+H} - z_g\|_2^2$. We then select the candidate action sequence with the lowest predicted cost:

$$\hat{i} = \arg\min_{i\in\{1,\dots,N\}} \mathcal{C}^{(i)}, \qquad a^{*}_{t:t+H} = a^{(\hat{i})}_{t:t+H}.$$

This procedure performs MPC-style planning by evaluating multiple horizon-length action proposals through the world model and choosing the sequence whose predicted latent trajectory best matches the desired goal. Note that ultimately the success or failure of the plan will rest on how accurately EgoExo-WM predicts its rollouts.

4 Experiments and Results

4.1 Experimental Setup

Training Data. Our training data consists of (1) real egocentric video and (2) synthetic egocentric data generated from exocentric sources. As **egocentric training data,** we use Nymeria [50], an existing egocentric dataset with paired 3D human motion trajectories. Nymeria provides egocentric video synchronized with full-body motion capture, recorded in real-world environments such as homes and outdoor parking lots using an Xsens system [54]. We follow the preprocessing steps of PEVA [6], resulting in approximately ~200 hours of training data.

Our approach further injects converted **exocentric training data** from multiple large-scale video datasets. We first preprocess them into clips where the head and body are visible and an action is being performed; details are provided in Supp. A.3. We then convert these clips into egocentric data using the pipeline in Section 3.2 and Figure 4. Our exocentric sources are: (1) HowTo100M [52], a large-scale instructional video dataset, from which we sample ~5 hours from the Food and Entertainment

HOMAGE				
Model	L2 ↓		PCK@20 ↑	
	2s	Avg	2s	Avg
PEVA-L	0.115	0.108	0.326	0.402
PEVA-XL	0.112	0.107	0.308	0.357
PEVA-XXL	0.109	0.103	0.321	0.365
EgoControl*	0.099	0.087	0.352	0.458
Ego-WM	0.069	0.058	0.313	0.396
Naive EgoExo-WM	0.065	0.053	0.347	0.447
EgoExo-WM	**0.057**	**0.047**	**0.404**	**0.531**

LEMMA				
Model	L2 ↓		PCK@20 ↑	
	2s	Avg	2s	Avg
PEVA-L	0.115	0.107	0.439	0.448
PEVA-XL	0.110	0.106	0.324	0.265
PEVA-XXL	0.109	0.101	0.363	0.420
EgoControl*	0.091	0.077	0.343	0.470
Ego-WM	0.068	0.055	0.433	0.527
Naive EgoExo-WM	0.064	0.049	0.439	0.561
EgoExo-WM	**0.058**	**0.045**	**0.515**	**0.618**

Ego-Exo4D-Bike				
Model	L2 ↓		PCK@20 ↑	
	2s	Avg	2s	Avg
PEVA-L	0.108	0.099	0.340	0.427
PEVA-XL	0.106	0.096	0.319	0.397
PEVA-XXL	0.103	0.093	0.255	0.420
EgoControl*	0.085	0.073	0.414	0.537
Ego-WM	0.050	0.042	0.468	0.561
Naive EgoExo-WM	**0.048**	**0.040**	0.382	0.525
EgoExo-WM	0.049	**0.040**	**0.489**	**0.603**

Ego-Exo4D-Cooking				
Model	L2 ↓		PCK@20 ↑	
	2s	Avg	2s	Avg
PEVA-L	0.105	0.097	0.210	0.298
PEVA-XL	0.105	0.097	0.210	0.301
PEVA-XXL	0.102	0.095	0.197	0.303
EgoControl*	0.090	0.078	0.223	0.378
Ego-WM	0.063	0.053	0.460	0.459
Naive EgoExo-WM	**0.062**	**0.052**	0.368	0.443
EgoExo-WM	**0.062**	**0.052**	**0.460**	**0.515**

Table 1: Open-loop world model evaluation across four datasets. We report final 2s error and average error over the rollout using DINOv3-L latent L_2 distance, along with wrist PCK@20. **EgoExo-WM** achieves the best overall performance, showing that converted exocentric data improves egocentric future prediction beyond egocentric-only training and naive use of raw exocentric video. We find that standard deviation < 0.005 and PCK < 0.03 for PEVA and EgoControl.

category; (2) CrossTask [81], a task-oriented instructional video dataset, from which we sample ~1 hour; and (3) 100 Days of Hands [64], a large collection of everyday hand-object interaction videos, from which we sample ~4 hours. In total, we use a diverse ~10-hour subset across these sources, which is the volume of data our compute could accommodate given the video-generation inference time; the framework can easily incorporate additional converted exocentric clips as compute allows.

Evaluation Data. We evaluate on diverse held-out datasets to assess both egocentric latent state prediction and downstream planning. Unlike prior methods evaluated primarily on training-domain datasets [6, 56], our evaluation spans broader variation in activities, environments, and action sequences. We use: (1) Home Action Genome (HOMAGE) [62], paired egocentric/exocentric videos of household activities such as cooking, cleaning, laundry, and grooming; (2) LEMMA [39], paired egocentric/exocentric videos of daily activities such as watering plants, making cereal, and sweeping. For HOMAGE and LEMMA, we extract full-body poses with SAM-Body4D and sample 150 clips, using single-agent videos for LEMMA. (3) Ego-Exo4D [29], a large-scale paired egocentric/exocentric dataset for which we use provided UniEgoMotion poses [57] and sample 125 clips each from the Bike and Cooking subsets.

Baselines and Ablations. We choose baselines to cover the key alternatives for evaluating EgoExo-WM. We compare against prior egocentric world models that predict first-person futures from human motion, include UniEgoMotion [57] as both a motion-generation baseline and action proposal model for planning, and ablate our own framework by holding architecture and objective fixed while varying the use exocentric data. All model variants use the same total training budget and resolution.

- **PEVA** [6] is a SOTA diffusion-based, human-motion-conditioned egocentric video prediction model. We compare against PEVA-L, PEVA-XL, and PEVA-XXL, trained on all 200 hours of Nymeria at 224×224 resolution, using 7, 15, and 15 context frames, respectively.
- **EgoControl*** [56] is also a SOTA diffusion-based, human-motion-conditioned egocentric video prediction model. The * indicates our adapted reimplementation, which we train on all 200 hours of Nymeria at 224×224 resolution because the original code and models are not currently released.

Model	HOMAGE		LEMMA		Ego-Exo4D-Bike		Ego-Exo4D-Cooking	
	MPJPE ↓	Wrist MPJPE ↓	MPJPE ↓	Wrist MPJPE ↓	MPJPE ↓	Wrist MPJPE ↓	MPJPE ↓	Wrist MPJPE ↓
UniEgoMotion	0.404 ± 0.035	0.471 ± 0.038	0.444 ± 0.016	0.493 ± 0.016	0.292 ± 0.044	0.367 ± 0.039	0.533 ± 0.011	0.58 ± 0.011
UniEgoMotion + Ego-WM	0.383 ± 0.010	0.447 ± 0.014	0.414 ± 0.012	0.455 ± 0.010	0.267 ± 0.007	0.341 ± 0.005	0.519 ± 0.015	0.568 ± 0.023
UniEgoMotion + EgoExo-WM	$\mathbf{0.362 \pm 0.012}$	$\mathbf{0.421 \pm 0.012}$	$\mathbf{0.396 \pm 0.008}$	$\mathbf{0.438 \pm 0.006}$	$\mathbf{0.245 \pm 0.016}$	$\mathbf{0.320 \pm 0.013}$	$\mathbf{0.498 \pm 0.018}$	$\mathbf{0.549 \pm 0.016}$

Table 2: **Model-Predictive Control Planning Results.** We evaluate world-model guided planning by ranking candidate trajectories proposed by UniEgoMotion [57]. EgoExo-WM achieves the lowest error across all four evaluation datasets, spanning diverse environments and actions, suggesting that converted exocentric-to-egocentric data improves coverage of human motions and interactions for selecting goal-directed trajectories.

- **UniEgoMotion** [57] is a scene-aware egocentric motion model. We use its Egocentric Motion Generation module as both a planning baseline and an action sampler whose trajectories are ranked by our world model in Table 2.
- **Ego-WM** is our egocentric-only variant, trained on 200 hours of Nymeria with the same architecture and objective as our full model, but without converted exocentric data.
- **Naive EgoExo-WM** uses the same architecture and objective, but trains on 190 hours of Nymeria plus 10 hours of raw exocentric video without EgoX-Body conversion. This tests whether gains come from exocentric data alone or from alignment to the egocentric observation-action format.

EgoExo-WM denotes our full method trained with 190 hours of Nymeria and 10 hours of converted exo-to-ego trajectories. Note that for fairness, all methods see the *same volume of data* at training.

Evaluation Protocol. All open-loop rollout experiments use a 2-second horizon, corresponding to 8 frames at 4 Hz. EgoExo-WM and PEVA are rolled out autoregressively by feeding each predicted observation back with the next action. EgoControl predicts the full 2-second sequence in one pass at 16 Hz, which we downsample to 4 Hz. We report both final 2-second performance and the average over all 8 rollout frames. For visual prediction quality, we measure L_2 distance between predicted and ground-truth future observations in DINOv3-L feature space; pixel-space outputs are first encoded with DINOv3-L, while latent-space predictions are evaluated directly. For human-action consistency, we report wrist PCK@20, using ViTPose [74] on generated frames for pixel-space methods and our wrist keypoint predictor for latent-space methods.

For model-predictive control planning, we use the world model to score sampled action sequences. Given the last observation, UniEgoMotion [57] samples 4 candidate 3D human motion sequences of horizon 8. EgoExo-WM rolls out future latents under each candidate and selects the sequence whose final latent is closest to the goal latent. We evaluate with whole-body and Wrist MPJPE, standard metrics for 3D human motion [35, 37, 51], and report mean and std dev over 5 runs.

4.2 Direct Evaluation of World Model Accuracy

Open-loop rollout evaluation is a standard mechanism for assessing whether a dynamics model can accurately predict future observations under a prescribed action sequence [6, 27, 56]. We evaluate EgoExo-WM on open-loop rollouts against PEVA and EgoControl, and conduct ablations to measure the contribution of our exocentric-to-egocentric training framework.

Table 1 shows that EgoExo-WM outperforms all PEVA variants and EgoControl under a matched 200-hour training data budget. Across all datasets, EgoExo-WM achieves lower L_2 embedding error, indicating more accurate 2-second open-loop rollouts. The gains are largest on HOMAGE and LEMMA, where the average L_2 error is reduced by more than half relative to the strongest PEVA baseline. This improvement is consistent with our method design: by converting diverse exocentric videos into egocentric training data, EgoExo-WM exposes the world model to a broader distribution of human motions, object interactions, and environments. This added coverage is especially valuable for HOMAGE and LEMMA whose actions and environments are underrepresented in Nymeria. EgoExo-WM also improves wrist PCK@20, reflecting the effect of our wrist-consistency loss in preserving fine-grained hand motion and interaction cues.

We ablate the role of our EgoExo-WM framework by comparing against variants that remove the exo-to-ego conversion step (Naive EgoExo-WM) or the exocentric data altogether (Ego-WM). EgoExo-WM improves performance on all datasets, with the clearest gains on HOMAGE and LEMMA, which contain home and object-centric activities that are well represented by the exocentric

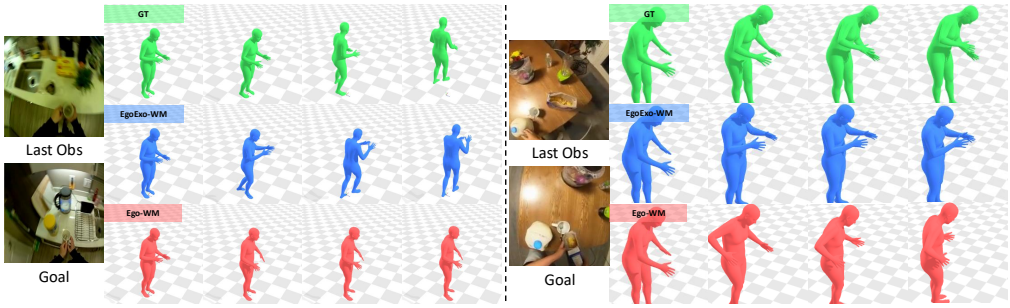

Figure 5: **Qualitative planning results.** From an observation and a visual goal, a trajectory sampler proposes candidate motion sequences, and the world model ranks them to select the one whose predicted outcome best matches the goal. In the first example, the goal is to move left toward the sink, whereas in the second, the goal is to pour cereal. EgoExo-WM chooses trajectories that better match the ground-truth behavior than Ego-WM in both cases.

interactions unlocked through our approach. Improvements in L_2 error are smaller on EgoExo4D-Bike and EgoExo4D-Cooking, likely because cooking is already well covered by Nymeria while biking is not well represented in data we converted. Even so, EgoExo-WM improves wrist PCK@20, indicating better fine-grained action consistency across the board. Compared to the naive exocentric variant, these results show that the gains do not come from adding raw exocentric videos alone, but from converting exocentric interactions into egocentric, action-aligned trajectories suitable for egocentric world-model training.

4.3 Impact for Model-Predictive Control Planning

These results demonstrate that EgoExo-WM can be used for goal-directed planning. The model can evaluate candidate human motion action sequences and select the one most likely to reach a desired visual goal. This requires reasoning about both navigation and coordinated body motion: for example, reaching for a mug may require walking to the shelf, tiptoeing, and reaching with the body. Although our SMPL-based action space does not model articulated hands, the planning task still depends on coordinated body motion. Our experiment isolates whether EgoExo-WM improves whole-body action selection compared to alternative world-model evaluators and a no-sampling baseline.

Table 2 shows that world-model-based ranking improves both body and wrist MPJPE over the no-sampling baseline, indicating that predictive models can provide guidance for whole-body action selection. However, EgoExo-WM achieves the strongest performance across the evaluation datasets, suggesting that its predictions are more informative for choosing goal-directed motions. These gains reflect the benefit of incorporating diverse converted exocentric videos, which cover a wider range of environments, body motions, and object-centric actions than egocentric data alone. This broader coverage allows EgoExo-WM to better evaluate candidate action sequences, leading to improved whole-body planning and finer-grained wrist motion across the diverse evaluation. See Fig. 5 and Supp. Video for more qualitative planning samples.

5 Limitations and Conclusion

EgoExo-WM is limited in temporal scale: EgoX-Body supports 49-frame clips, so converted exocentric data captures short interactions, and our planning experiments use a 2-second horizon. Longer-horizon planning under compounding prediction errors remains future work. Exo-to-ego conversion also struggles with complex human-object interactions, especially occlusion, precise contact, and small-object manipulation, and some predictions degenerate into mostly black or white frames. See Supp. A.3.2 and Supp. Video for examples.

We presented EgoExo-WM, a framework for using exocentric video to train egocentric world models. By recovering 3D human motion, we use pose both to guide the exo-to-ego conversion and as the action representation for world-model training. Our results show that converted exocentric data improves egocentric prediction and goal-directed planning across diverse evaluation datasets, suggesting a scalable path toward learning first-person predictive models from abundant exocentric human video.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Hassan Abu Alhaija, Jose Alvarez, Maciej Bala, Tiffany Cai, Tianshi Cao, Liz Cha, Joshua Chen, Mike Chen, Francesco Ferroni, Sanja Fidler, et al. Cosmos-transfer1: Conditional world generation with adaptive multimodal control. *arXiv preprint arXiv:2503.14492*, 2025.

[3] Kumar Ashutosh, Georgios Pavlakos, and Kristen Grauman. Fiction: 4d future interaction prediction from video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17613–17625, 2025.

[4] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[5] Arpit Bahety, Arnav Balaji, Ben Abbatematteo, and Roberto Martín-Martín. Safemimic: Towards safe and autonomous human-to-robot imitation for mobile manipulation. *arXiv preprint arXiv:2506.15847*, 2025.

[6] 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.

[7] Albert Bandura and Richard H Walters. *Social learning theory*, volume 1. Prentice-hall Englewood Cliffs, NJ, 1977.

[8] Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.

[9] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021.

[10] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5470–5479, 2022.

[11] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015.

[12] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36:11454–11468, 2023.

[13] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. Augmented reality technologies, systems and applications. *Multimedia tools and applications*, 51(1): 341–377, 2011.

[14] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4217–4229, 2023.

[15] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2416–2425, 2023.

[16] Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18000–18010, 2023.

[17] Feng Cheng, Mi Luo, Huiyu Wang, Alex Dimakis, Lorenzo Torresani, Gedas Bertasius, and Kristen Grauman. 4diff: 3d-aware diffusion model for third-to-first viewpoint translation. In *European Conference on Computer Vision*, pages 409–427. Springer, 2024.

[18] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.

[19] Jessica Dashner, Kerri Morgan, Sue Tucker, Carla Walker, and Sandra Martina Espín Tello. The human activity assistive technology model. In *Routledge Companion to Occupational Therapy*, pages 508–517. Routledge, 2025.

[20] Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, Yao Feng, and Michael J Black. Tokenhmr: Advancing human mesh recovery with a tokenized pose representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1323–1333, 2024.

[21] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.

[22] Aaron Ferguson, Ahmed AA Osman, Berta Bescos, Carsten Stoll, Chris Twigg, Christoph Lassner, David Otte, Eric Vignola, Fabian Prada, Federica Bogo, et al. Mhr: Momentum human rig. *arXiv preprint arXiv:2511.15586*, 2025.

[23] Mingqi Gao, Yunqi Miao, and Jungong Han. Sam-body4d: Training-free 4d human body mesh recovery from videos. *arXiv preprint arXiv:2512.08406*, 2025.

[24] Quankai Gao, Jiawei Yang, Qiangeng Xu, Le Chen, and Yue Wang. Lome: Learning human-object manipulation with action-conditioned egocentric world model. *arXiv preprint arXiv:2603.27449*, 2026.

[25] Shenyuan Gao, William Liang, Kaiyuan Zheng, Ayaan Malik, Seonghyeon Ye, Sihyun Yu, Wei-Cheng Tseng, Yuzhu Dong, Kaichun Mo, Chen-Hsuan Lin, et al. Dreamdojo: A generalist robot world model from large-scale human videos. *arXiv preprint arXiv:2602.06949*, 2026.

[26] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14783–14794, 2023.

[27] Raktim Gautam Goswami, Amir Bar, David Fan, Tsung-Yen Yang, Gaoyue Zhou, Prashanth Krishnamurthy, Michael Rabbat, Farshad Khorrami, and Yann LeCun. World models can leverage human videos for dexterous manipulation. *arXiv preprint arXiv:2512.13644*, 2025.

[28] 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*, pages 18995–19012, 2022.

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

[30] Agrim Gupta, Stephen Tian, Yunzhi Zhang, Jiajun Wu, Roberto Martín-Martín, and Li Fei-Fei. Maskvit: Masked visual pre-training for video prediction. In *The Eleventh International Conference on Learning Representations*, 2022.

[31] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3):440, 2018.

[32] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[33] Noriaki Hirose, Amir Sadeghian, Fei Xia, Roberto Martín-Martín, and Silvio Savarese. Vunet: Dynamic scene view synthesis for traversability estimation using an rgb camera. *IEEE Robotics and Automation Letters*, 4(2):2062–2069, 2019.

[34] Noriaki Hirose, Fei Xia, Roberto Martín-Martín, Amir Sadeghian, and Silvio Savarese. Deep visual mpc-policy learning for navigation. *IEEE Robotics and Automation Letters*, 4(4):3184–3191, 2019.

[35] Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5344–5354, 2025.

[36] Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025.

[37] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013.

[38] Joel Jang, Seonghyeon Ye, Zongyu Lin, Jiannan Xiang, Johan Bjorck, Yu Fang, Fengyuan Hu, Spencer Huang, Kaushil Kundalia, Yen-Chen Lin, et al. Dreamgen: Unlocking generalization in robot learning through video world models. *arXiv preprint arXiv:2505.12705*, 2025.

[39] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for le arning m ulti-agent m ulti-task a ctivities. In *European Conference on Computer Vision*, pages 767–786. Springer, 2020.

[40] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023.

[41] Taewoong Kang, Kinam Kim, Dohyeon Kim, Minho Park, Junha Hyung, and Jaegul Choo. Egox: Egocentric video generation from a single exocentric video. *arXiv preprint arXiv:2512.08269*, 2025.

[42] Simar Kareer, Dhruv Patel, Ryan Punamiya, Pranay Mathur, Shuo Cheng, Chen Wang, Judy Hoffman, and Danfei Xu. Egomimic: Scaling imitation learning via egocentric video. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13226–13233. IEEE, 2025.

[43] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[44] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148, 2021.

[45] Jinhan Li, Yifeng Zhu, Yuqi Xie, Zhenyu Jiang, Mingyo Seo, Georgios Pavlakos, and Yuke Zhu. Okami: Teaching humanoid robots manipulation skills through single video imitation. *arXiv preprint arXiv:2410.11792*, 2024.

[46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[48] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. 2023.

[49] Mi Luo, Zihui Xue, Alex Dimakis, and Kristen Grauman. Put myself in your shoes: Lifting the egocentric perspective from exocentric videos. In *European Conference on Computer Vision*, pages 407–425. Springer, 2024.

[50] Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pages 445–465. Springer, 2024.

[51] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2891–2900, 2017.

[52] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019.

[53] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65 (1):99–106, 2021.

[54] Movella. *MVN User Manual*. Movella, 2021. `https://www.movella.com/hubfs/MVN_User_Manual.pdf`.

[55] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

[56] Enrico Pallotta, Sina Mokhtarzadeh Azar, Lars Doorenbos, Serdar Ozsoy, Umar Iqbal, and Juergen Gall. Egocontrol: Controllable egocentric video generation via 3d full-body poses. *arXiv preprint arXiv:2511.18173*, 2025.

[57] Chaitanya Patel, Hiroki Nakamura, Yuta Kyuragi, Kazuki Kozuka, Juan Carlos Niebles, and Ehsan Adeli. Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10318–10329, 2025.

[58] 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*, pages 10975–10985, 2019.

[59] 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*, pages 9826–9836, 2024.

[60] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10318–10327, 2021.

[61] Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. Worldgym: World model as an environment for policy evaluation. *arXiv preprint arXiv:2506.00613*, 2025.

[62] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021.

[63] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.

[64] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020.

[65] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Shanshe Wang, Siwei Ma, and Wen Gao. P-stmo: Pre-trained spatial temporal many-to-one model for 3d human pose estimation. In *European Conference on Computer Vision*, pages 461–478. Springer, 2022.

[66] Soyong Shin, Juyong Kim, Eni Halilaj, and Michael J Black. Wham: Reconstructing world-grounded humans with accurate 3d motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2070–2080, 2024.

[67] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

[68] Himanshu Gaurav Singh, Pieter Abbeel, Jitendra Malik, and Antonio Loquercio. Deep sensorimotor control by imitating predictive models of human motion. *arXiv preprint arXiv:2508.18691*, 2025.

[69] Bérangere Thirioux, Manuel R Mercier, Gérard Jorland, Alain Berthoz, and Olaf Blanke. Mental imagery of self-location during spontaneous and active self–other interactions: An electrical neuroimaging study. *Journal of Neuroscience*, 30(21):7202–7214, 2010.

[70] Michael Tomasello, Sue Savage-Rumbaugh, and Ann Cale Kruger. Imitative learning of actions on objects by children, chimpanzees, and enculturated chimpanzees. *Child development*, 64(6):1688–1705, 1993.

[71] Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. Playerone: Egocentric world simulator. *arXiv preprint arXiv:2506.09995*, 2025.

[72] 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.

[73] Jilan Xu, Yifei Huang, Baoqi Pei, Junlin Hou, Qingqiu Li, Guo Chen, Yuejie Zhang, Rui Feng, and Weidi Xie. Egoexo-gen: Ego-centric video prediction by watching exo-centric videos. *arXiv preprint arXiv:2504.11732*, 2025.

[74] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35:38571–38584, 2022.

[75] Xitong Yang, Devansh Kukreja, Don Pinkus, Anushka Sagar, Taosha Fan, Jinhyung Park, Soyong Shin, Jinkun Cao, Jiawei Liu, Nicolas Ugrinovic, et al. Sam 3d body: Robust full-body human mesh recovery. *arXiv preprint arXiv:2602.15989*, 2026.

[76] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023.

[77] Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

[78] Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21232, 2023.

[79] Wanqi Yin, Zhongang Cai, Ruisi Wang, Ailing Zeng, Chen Wei, Qingping Sun, Haiyi Mei, Yanjun Wang, Hui En Pang, Mingyuan Zhang, et al. Smplest-x: Ultimate scaling for expressive human pose and shape estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[80] Ruijie Zheng, Dantong Niu, Yuqi Xie, Jing Wang, Mengda Xu, Yunfan Jiang, Fernando Castañeda, Fengyuan Hu, You Liang Tan, Letian Fu, et al. Egoscale: Scaling dexterous manipulation with diverse egocentric human data. *arXiv preprint arXiv:2602.16710*, 2026.

[81] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019.

A Technical appendices and supplementary material

In this supplementary material, we provide the following:

1. **World Model Implementation Details (Sec. A.1)** — We describe our architecture and training details, action representation, and wrist-position consistency loss.
2. **EgoX-Body implementation details (Sec. A.2)** — We describe the human motion extraction, egocentric prior construction, hand-conditioning procedure, diffusion-model inputs, and implementation details used in EgoX-Body.
3. **Exocentric video selection and filtering (Sec. A.3)** — We describe how exocentric videos are selected, segmented, filtered, and post-processed before being used for world-model training.

A.1 World Model Implementation Details

A.1.1 Architecture and Training Details

We train our world model with a CDiT-L/2 [8, 6] backbone consisting of 24 layers, hidden size 1024, 16 attention heads, and MLP ratio 4. The model takes as input DINOv3 ViT-L/16 patch tokens extracted from 224×224 images, producing a 14×14 token grid with 196 tokens of dimension 1024. We condition on 3.

We optimize with AdamW using a constant learning rate of 8×10^{-5}, β_1=0.9, β_2=0.95, zero weight decay, gradient clipping at 10.0, and `bfloat16` mixed precision. We train with a per-GPU batch size of 64 on 8 A40 GPUs, resulting in an effective global batch size of 512 and train for a total of $100,000$ iterations. For evaluation, we use an exponential moving average of the model weights, and the model is compiled with `torch.compile` during training and evaluation.

A.1.2 Action Representation

We use a 3D human-motion action space based on the 22-joint SMPL body skeleton, as described in the main paper in Section 3.1. PEVA and EgoControl use Xsens motion trajectories with 23 body joints, while our model uses the 22-joint SMPL skeleton. We therefore convert actions to the skeleton required by each model. The primary mismatch is the torso chain: Xsens uses four spine joints, `L5`, `L3`, `T12`, and `T8`, while SMPL uses three. For Xsens-to-SMPL conversion, we drop the intermediate lumbar joint `L3` and map `L5`, `T12`, and `T8` to the three SMPL spine joints. For SMPL-to-Xsens conversion, we duplicate the corresponding lumbar spine joint to recover the `L3` entry. This conversion only changes the skeleton layout; the action parameterization remains unchanged.

A.1.3 Wrist-Position Consistency Objective

The wrist consistency loss encourages the world model's predicted next-frame representations to remain spatially faithful to hand placement. We attach a frozen wrist-heatmap decoder on top of the world model's predicted DINOv3 patch tokens and penalize disagreement with a 2D ground-truth heatmap. The decoder is a 6-layer Transformer head (256-dim, 16 heads, MLP ratio 4) that takes the 14×14 grid of DINOv3 ViT-L/16 patch tokens (dim 1024) from a 224×224 image and predicts a 16×16 local patch per token, tiled back into a single-channel 224×224 heatmap. The decoder is pretrained on EgoExo4D and Nymeria with per-pixel MSE against ground-truth Gaussian heatmaps before being frozen for world-model training.

Ground-truth heatmaps are generated from ViTPose wrist keypoints, keeping only detections with mean confidence above 0.3 and de-duplicating wrists within 5 px of each other. Both wrists are rendered onto a single channel by splatting an isotropic 2D Gaussian with $\sigma = 3.0$ px and combining the two via element-wise max, which preserves a unit peak when the wrists are co-located. During world-model training, the frozen decoder is applied to the predicted next-frame patch tokens, and we add an MSE loss between the resulting heatmap and the ground-truth, masked to frames in which at least one wrist is visible.

A.2 EgoX-Body Details

A.2.1 Release

When releasing EgoX-Body, which entails image generation, we will state guidelines for use.

A.2.2 Compute Details

We train EgoX-Body on 4 GH200s for $20,000$ iterations and perform inference on GH200 gpus with each sample taking of 49 frames taking 3.5 minutes.

A.2.3 Body Pose and Scene Reconstruction

Given an in-the-wild exocentric video, we estimate full-body pose using SAM-Body4D and obtain a 4D scene reconstruction using ViPE [36] following EgoX. The extracted body pose is converted from the MHR representation to SMPL-X for compatibility with the rest of the pipeline. The body pose provides the actor's motion and 3D hand positions, while the 4D reconstruction provides scene geometry for rendering an approximate egocentric view.

A.2.4 Egocentric Prior and Hand Overlay

We render the SMPL-X skeleton recovered from the exocentric video into an egocentric view using a pinhole projection from a head-anchored virtual camera. The virtual camera center is defined as the midpoint of the SMPL-X eye joints, `L_EYE`=23 and `R_EYE`=24. We additionally push the virtual camera forward 0.1m and remove points too close to avoid rendering artifacts. We construct the camera basis using the eye-to-eye direction as the horizontal axis $\mathbf{x}$, the camera-center-to-`NECK` direction as the vertical axis $\mathbf{y}$, and $\mathbf{z} = \mathbf{x} \times \mathbf{y}$ as the forward-looking axis.

For the hand overlay, we use the SMPL-X wrist and finger joints for each hand, with 16 joints and 15 bones per hand. Bones connect the wrist to each finger and proceed distally along the kinematic chain. The left hand is drawn in red and the right hand in blue, with circles for joints and lines for bones.

A.2.5 Practical Modifications

We make several practical modifications to improve robustness and efficiency. We add 250 training samples from the H2o [44] dataset, focusing on sequences where the person faces the camera, which better matches many in-the-wild exocentric videos. We remove the Geometry Guided Attention (GGA) module from EgoX, reduce the generation resolution from 448×448 to 384×384, and save generated videos at 16 Hz. On an NVIDIA GH200, these changes reduce inference time from approximately 17.5 minutes per video to approximately 3.25 minutes per video. Following EgoX, we perform inference on 49-frame sequences.

A.3 Exocentric Video Selection and Filtering

To construct valid training clips from raw videos, we apply an automatic scene filtering pipeline. Each video is first partitioned into candidate scenes using adaptive scene detection, and scenes below a minimum duration are removed. For each remaining scene, we sample three representative frames and require a single visible person to be detected in all of them. This is checked using ViTPose [74], where a scene is retained only when reliable head and upper-body keypoints are present. We also estimate camera motion over the scene using ORB correspondences and affine motion estimation, rejecting clips with noticeable zoom, translation, or rotation.

The remaining candidates are passed to GPT-4o-mini as a conservative visual validator. Given the sampled frames, GPT-4o-mini [1] predicts whether the scene contains a clear human action, estimates the maximum overlay coverage, and identifies photographic overlays such as picture-in-picture videos. We accept a scene only when a visible human action is present, the estimated overlay coverage is below 20%, and no photographic overlay is detected. We include the prompt we use below.

VLM Filtering Prompt.

```
Check these images and answer STRICTLY in JSON.
```

```
You are judging the scene across all provided images together.

Definitions:
- A "human action" means a visible person actively doing something.
- human_action should be true ONLY if a person is clearly performing
  a visible action in at least one image, and the scene overall appears
  to depict a human action.
  Examples: cutting, cooking, typing, walking, opening something,
  holding or manipulating objects, cleaning, assembling, playing.

- human_action should be false for:
  - static poses (standing, sitting without action)
  - portraits/selfies
  - empty scenes
  - landscapes
  - objects moving without human involvement
  - animals acting without humans
  - a person just talking to the camera (e.g. vlog-style, speaking, presenting)
  - a person facing the camera without clear interaction with objects

Overlay rules:
- An "overlay" is anything drawn on top of the main scene.
- overlay_is_photographic should be true ONLY if any image contains a
  natural photo or video image as an overlay (e.g. picture-in-picture, inset video).
- overlay_is_photographic should be false for logos, branding, text,
  subtitles, UI elements, icons, flat illustrations, watermarks.

Rules:
- human_action: true only if a clear human action is visible across the set
- overlay_pct: estimated percent (0-100) of the image covered by overlays;
  if images differ, use the highest / worst-case estimate
- passes = human_action AND overlay_pct < 20 AND (overlay_is_photographic == false)

Be conservative if unsure.
```

A.3.1 Post-Processing

EgoX-Body converts each exocentric clip into a 49-frame egocentric sequence at 16 Hz. For world model training, we temporally downsample these converted clips to 8 Hz, yielding 25 frames per sequence. We apply the same temporal sampling to the corresponding 3D human-motion trajectories, so each converted clip becomes an egocentric observation-action sequence aligned with the format of real egocentric trajectories. This allows converted exocentric interactions to be used jointly with real egocentric data during training.

After conversion, we standardize generated clips for world model training. Each converted video is saved at 384×384 resolution and 16 Hz. We center crop at 85%, resize to 224×224, and then we downsample to 8 Hz for world model training and pair each egocentric observation sequence with the corresponding whole-body action trajectory derived from the estimated body pose.

A.3.2 Failure Cases

We observe that Internet ego-view collections contain several recurring failure cases that make them unhelpful for training. These include black or near-black cuts, title cards or overexposed frames, motion-blurred clips, and sped-up tutorials with abrupt frame-to-frame changes. Such clips either lack useful visual content or contain temporal artifacts that can degrade world model training. We provide some qualitative examples of failure cases in Figure 6 and in the Supplementary Video.

A.3.3 Filtering EgoX-Body Generated Videos

To remove these failure cases, we implement a simple automatic quality filter on the 326×326 ego crop of each source clip. Specifically, we compute four scalar statistics: `black_fraction_mean`, the average fraction of near-black pixels over time; `white_fraction_mean`, the average fraction of near-white pixels over time; `blur_median`, the median Laplacian variance across frames; and `motion_median`, the median mean-absolute-difference between consecutive grayscale frames.

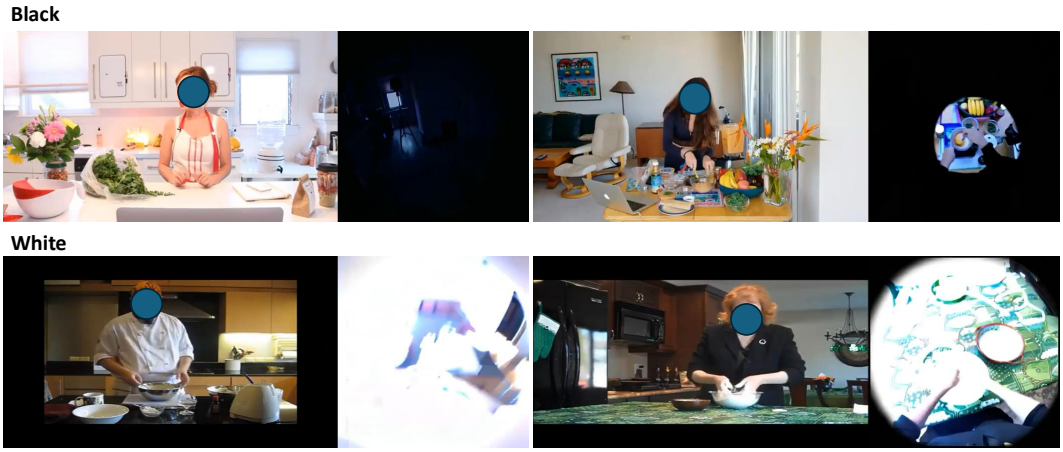

Figure 6: **Examples of failure cases in Internet ego-view videos.** We show representative clips with large white or black regions, which commonly arise from videos where the person is directly facing the camera where the egocentric prior is essentially black. These failure cases provide little useful training signal, motivating the automatic filtering criteria described in Section A.3.3.

We retain clips satisfying `black_fraction_mean` < 0.30, `white_fraction_mean` < 0.20, `blur_median` > 50, and `motion_median` < 32.5. These thresholds are calibrated by inspecting random clips near each cutoff to verify that they separate clear failure cases from usable footage. This filtering retains roughly 80% of candidate clips for training.