HUMANWM: A WORLD ACTION MODEL FOR EGOCENTRIC HUMAN MOTION RECONSTRUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing human motion from egocentric video requires inferring a largely unseen body from viewpoint changes and interactions with the environment. We present HUMANWM, a World Action Model for Egocentric Motion that adapts a pretrained world model to jointly model egocentric video and structured human actions. The model introduces Coupled Kinematic Supervision that jointly constrains spatial structure and acceleration for egocentric motion reconstruction. At inference, full-sequence optimization balances fidelity to predicted motion with rotational and root-dynamics regularization. On 3,000 sequences from 57 Nymeria test recordings, HUMANWM outperforms all evaluated baselines across pose, root-trajectory, motion-dynamics, and foot-skating metrics using RGB alone. Compared with RGB-only ReViV, the RGB-only configuration reduces globally aligned and Procrustes-aligned mean per-joint position errors by 38.2% and 44.9%, respectively. Task text further improves pose and root-trajectory accuracy, yielding 78.10 mm GA-MPJPE, 46.49 mm PA-MPJPE, and 0.216 m final root trajectory error.

1 INTRODUCTION

Understanding human motion from egocentric video is important for embodied learning from human experience. However, egocentric RGB video typically provides only partial observations of the wearer's body, making it challenging to reconstruct wearer motion from visible body cues alone. Viewpoint changes and interactions with the surrounding environment provide indirect evidence of wearer motion; the challenge is to connect these visual changes to the evolution of the wearer's body state. We introduce HUMANWM, a World Action Model for Egocentric Motion that adapts a pretrained world model to jointly model structured wearer motion and egocentric visual dynamics. Figure 1 illustrates qualitative motion reconstructions from in-the-wild egocentric videos, showing reconstructed body poses and trajectories alongside the corresponding input frames.

Existing egocentric motion methods use device measurements and learned motion priors to constrain the ambiguity of an unseen body. HMD^2 and EgoAllo combine egocentric imagery with SLAM-derived head motion to infer the wearer's movement in the surrounding scene (Guzov et al., 2025; Yi et al., 2025). These measurements provide useful spatial constraints, but their availability is an additional assumption beyond monocular RGB video. More recent methods broaden the scope of egocentric modeling: UniEgoMotion unifies motion estimation, forecasting, and generation (Patel et al., 2025), while EgoPriMo connects language-conditioned egocentric motion priors to humanoid control (Ge et al., 2026). ReViV further demonstrates joint estimation of wearer motion, camera motion, and scene geometry directly from monocular video (Lyu et al., 2026).

In parallel, recent embodied research increasingly connects action learning to world prediction. DreamZero adapts a pretrained video generation backbone to jointly model future video and robot actions (Ye et al., 2026), while EgoWAM investigates how world-prediction targets enable transfer from egocentric human data to robot policies (Li et al., 2026). Pretrained world models capture rich priors about real-world visual dynamics. World action models leverage these priors to predict actions from language instructions and scene observations. We aim to adapt these capabilities to egocentric wearer-motion reconstruction, where visible body cues are limited.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

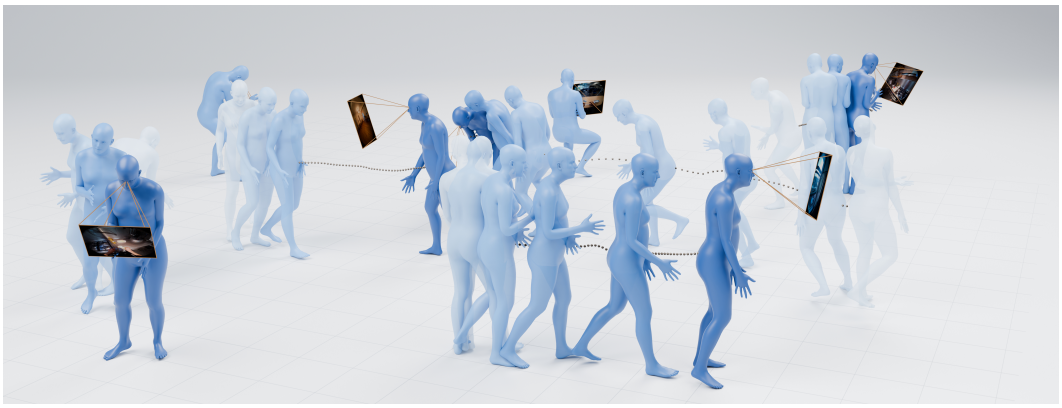

Figure 1: **In-the-wild egocentric motion reconstruction.** HUMANWM reconstructs wearer motion from egocentric videos in the wild. Reconstruction results from multiple sequences are presented together in a single composite visualization.

HUMANWM adapts the pretrained Cosmos 3 world model (NVIDIA, 2026), mapping human motion onto its native action interface. The model reconstructs wearer motion from egocentric RGB video, optionally conditioned on task text. The same model also supports joint video generation from an initial frame. Accurate poses at individual frames do not necessarily ensure coherent root trajectories or temporal dynamics. The base action objective does not explicitly supervise root trajectory structure or acceleration. We therefore introduce Coupled Kinematic Supervision (CKS), which supervises overall path structure and local acceleration from the same predicted horizontal root trajectory in a canonical ground frame. Auxiliary geometric supervision further constrains human joint positions and displacements, together with foot displacements. To improve temporal consistency at inference, we introduce full-sequence inference optimization. Starting from aligned and fused window predictions, we refine root and body motion over the complete sequence by balancing prediction fidelity with regularization of angular-motion changes, root acceleration, and jerk. The optimization operates only on inference, without additional training.

On 57 test recordings, HUMANWM outperforms all evaluated baselines across pose, root trajectory, motion-dynamics, and foot-skating metrics using RGB alone. Compared with RGB-only ReViV, it reduces GA-MPJPE and PA-MPJPE by 38.2% and 44.9%, respectively, and mean root trajectory error by 35.0%. Adding task text further improves pose and root trajectory accuracy, reaching 78.10 mm GA-MPJPE and 46.49 mm PA-MPJPE, whereas the RGB-only configuration yields lower acceleration and jerk errors. Our main contributions can be summarized as follows:

- We introduce HUMANWM, a World Action Model for Egocentric Motion that adapts a pretrained world model's native action interface to structured wearer motion. A single model supports motion reconstruction and video generation.
- We introduce Coupled Kinematic Supervision (CKS), which jointly constrains root-trajectory structure and temporal dynamics through complementary path and acceleration supervision.
- We introduce full-sequence inference optimization, which refines root and body motion by balancing prediction fidelity with rotational and root-dynamics regularization.

2 RELATED WORK

2.1 EGOCENTRIC HUMAN MOTION MODELING

Egocentric motion reconstruction combines indirect visual evidence with motion priors to infer a partially observed body. EgoEgo uses SLAM and learned head-motion estimation to obtain a head trajectory, which conditions a diffusion model for body motion (Li et al., 2023). HMD^2 incorporates head motion, image embeddings, and a SLAM point cloud (Guzov et al., 2025), while EgoAllo combines a head-conditioned motion prior with image-based guidance to estimate body and hand

motion (Yi et al., 2025). EgoExoMoCap additionally exploits views from other wearable cameras and tracked head or wrist signals (Jiang et al., 2026). These methods relate wearer motion to the surrounding scene under different sensing assumptions.

Recent work expands both the reconstructed signals and the task interface. ReViV jointly estimates wearer motion, camera motion, depth, and gaze from monocular RGB video (Lyu et al., 2026). UniEgoMotion uses visual scene context and a head-centric representation for motion estimation, forecasting, and generation (Patel et al., 2025); EgoPriMo further integrates language into egocentric motion priors and connects them to humanoid control (Ge et al., 2026). Our focus is adapting a pretrained world model's native action interface to structured wearer motion, with explicit supervision of root-trajectory structure and dynamics. HUMANWM predicts camera transitions together with body motion and does not require externally supplied head trajectories. The auxiliary inputs used in each evaluated configuration are specified in Section 4.

2.2 PRETRAINED WORLD MODELS FOR ACTION LEARNING

World action models connect actions to their visual consequences. DreamZero adapts a pretrained video diffusion backbone to jointly model robot actions and future video (Ye et al., 2026). EgoWAM studies world-prediction targets for transferring egocentric human experience to robot policies (Li et al., 2026), while MotionWAM conditions humanoid control on intermediate features of a video world model (Zheng et al., 2026). For human motion, EgoTwin jointly generates egocentric video and wearer motion by modeling the relationship between body movement and viewpoint changes (Xiu et al., 2025). These advances motivate learning actions together with visual dynamics, rather than treating motion solely as an output of visual perception.

Building on Cosmos 3 (NVIDIA, 2026), HUMANWM adapts an existing action interface to encode camera transitions, root states, and body states within a shared model of video and motion. The adaptation combines this structured action domain with kinematic and geometric supervision, rather than using the world model only as a visual feature extractor. Different video-observation masks provide reconstruction, generation, and forecasting conditions during training; our experiments focus on motion reconstruction and video generation conditioned on an initial image and task text.

2.3 MOTION REPRESENTATIONS AND TEMPORAL REGULARIZATION

Coordinate choices determine how motion states remain comparable across viewpoints and time. GVHMR introduces gravity–view coordinates for motion recovery from videos that observe a person, using gravity and the viewing direction to relate body orientation to the camera (Shen et al., 2024). For egocentric motion, EgoAllo designs head-motion conditioning around spatial and temporal invariance (Yi et al., 2025), while UniEgoMotion adopts a head-centric representation (Patel et al., 2025). Our hybrid representation places root positions in a fixed canonical ground frame and root rotations in a time-varying gravity–view frame. This keeps the trajectory comparable across time while relating body rotation to the current viewpoint, providing the spatial basis for CKS.

Kimodo separates a smoothed root path from body motion and uses a two-stage denoiser for controllable motion synthesis (Rempe et al., 2026). Inspired by structured root-motion learning, CKS supervises the overall locomotion path and local acceleration of the same predicted horizontal root trajectory. Low-pass filtering is used only to compute the path loss; the model still predicts the original root coordinates without a separate root generator. We further distinguish this training supervision from full-sequence inference optimization: CKS matches reference trajectories and dynamics, whereas inference optimization regularizes changes in rotation increments, root acceleration, and jerk while anchoring the result to the model predictions. The latter operates on aligned and fused window predictions without ground-truth motion or network updates.

3 METHOD

HUMANWM reconstructs the wearer's body pose, root trajectory, and camera motion from monocular egocentric RGB video. It also jointly generates future video and motion from partial observations (Fig. 2). We represent structured wearer motion through the native action interface of a pretrained world model. During training, Coupled Kinematic Supervision (CKS) jointly supervises trajectory

structure and acceleration. At inference, full-sequence inference optimization balances fidelity to the predicted motion with temporal regularity.

3.1 HUMANWM: HUMAN WORLD ACTION MODEL

World action models can jointly generate robot actions and future video conditioned on visual observations and language instructions. We adopt the pretrained Cosmos 3 Edge model (NVIDIA, 2026) as our base model. Cosmos 3 supports multiple action domains for different embodiments. We extend this framework with a dedicated domain for structured human motion. This domain represents each 144-dimensional action using three components: a camera transition, a root state, and body states. We define the pelvis center as the root joint. Crucially, root state prediction is incorporated into the human motion action domain to provide the structured motion representation used by Coupled Kinematic Supervision to jointly constrain the spatial structure and temporal dynamics of root motion.

Given an egocentric video clip of $T+1$ frames, we reconstruct the wearer's motion as a sequence of T actions, $\mathbf{A} = \{\mathbf{a}_t\}_{t=1}^{T}$. Each action comprises the camera transition from frame $t-1$ to t and the root state and body states at frame t:

$$\mathbf{a}_t = \left[\Delta\mathbf{C}_t,\, \mathbf{s}_t^{\mathrm{root}},\, \boldsymbol{\theta}_t^{\mathrm{body}}\right] \in \mathbb{R}^{144}. \tag{1}$$

Where, $\Delta\mathbf{C}_t \in \mathbb{R}^9$ contains a 3D relative translation and a 6D relative rotation. The root state $\mathbf{s}_t^{\mathrm{root}} = [\mathbf{p}_t, \mathbf{r}_t^{\mathrm{root}}] \in \mathbb{R}^9$ combines the 3D root position $\mathbf{p}_t \in \mathbb{R}^3$ with its continuous 6D rotation $\mathbf{r}_t^{\mathrm{root}} \in \mathbb{R}^6$. The body $\boldsymbol{\theta}_t^{\mathrm{body}} \in \mathbb{R}^{126}$ concatenates the 6D parent-relative rotations of 21 SMPL-X body joints (Pavlakos et al., 2019). HUMANWM reconstructs human motion from complete egocentric video.

3.2 COUPLED KINEMATIC SUPERVISION

In egocentric video, camera rotation continuously changes the visual reference frame, making root position and rotation difficult to model consistently over time. Even with low per-frame root-position error, consistently biased errors can accumulate into gradual path drift or incorrect curvature, whereas alternating errors across adjacent frames can produce root jitter and erratic acceleration. We therefore introduce *Coupled Kinematic Supervision* (CKS), which first stabilizes the motion space through gravity–view grounding and then applies multi-scale root trajectory supervision to jointly constrain the spatial structure and temporal dynamics of root motion.

To obtain a root state $\mathbf{s}_t^{\mathrm{root}}$ that remains stable under egocentric viewpoint changes, we adopt a hybrid grounded gravity–view representation. The root position is expressed in a canonical ground frame $\mathbf{G}_0$, whose vertical axis is aligned with gravity and whose horizontal heading is defined by the first-frame viewing direction projected onto the ground plane. Its vertical coordinate is encoded as $z_t - g_{\mathrm{sup}}$, where g_{sup} is an episode-level support reference estimated from contacted toe and heel segments. In contrast, the continuous 6D root rotation is expressed in a time-varying gravity–view frame $\mathbf{G}_t$, constructed at each frame from gravity and the current viewing direction projected onto the ground plane. By expressing all root positions in the same $\mathbf{G}_0$ frame, this representation gives their temporal differences a consistent physical reference.

From the predicted action sequence $\mathbf{A}$, we collect the horizontal components $\mathbf{p}_{t,xy}$ of the root positions defined in Section 3.1 to form the predicted trajectory $\mathbf{P}_{xy} = (\mathbf{p}_{t,xy})_{t=1}^{T}$. Let $\mathbf{A}^\star$ denote the ground-truth action sequence; its root states analogously define the ground-truth trajectory $\mathbf{P}^\star_{xy}$. To supervise the overall locomotion path, we apply the same symmetric low-pass filter $\mathcal{H}$ to both trajectories:

$$\mathcal{L}_{\mathrm{path}} = \left\| \mathcal{H}(\mathbf{P}_{xy}) - \mathcal{H}(\mathbf{P}^\star_{xy}) \right\|_F^2 . \tag{2}$$

This objective emphasizes low-frequency errors in the route, including gradual drift and incorrect path curvature. Filtering is used only to compute this loss; the model still predicts the original root coordinates. Path agreement alone does not ensure accurate local dynamics. We therefore match second-order temporal differences of the *unfiltered* trajectories:

$$\mathcal{L}_{\mathrm{dyn}} = \left\| \Delta^2\mathbf{P}_{xy} - \Delta^2\mathbf{P}^\star_{xy} \right\|_F^2 . \tag{3}$$

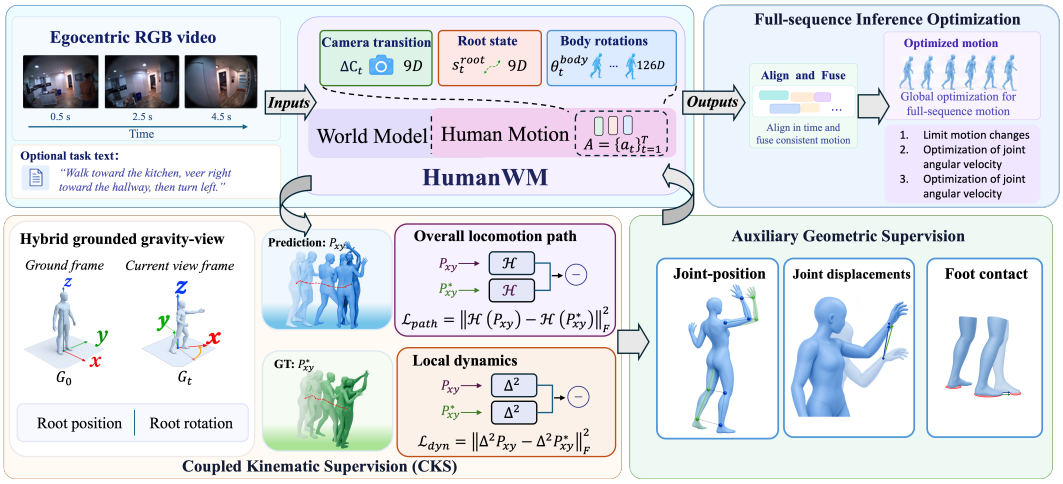

Figure 2: HUMANWM adapts one pretrained backbone and action interface for motion reconstruction. Coupled kinematic supervision learns spatial states and trajectory dynamics within the same action sequence. Camera motion is predicted jointly and used to decode root orientation; auxiliary geometry supervision is used only during training.

Here, $\Delta^2 \mathbf{p}_{t,xy} = \mathbf{p}_{t+1,xy} - 2\mathbf{p}_{t,xy} + \mathbf{p}_{t-1,xy}$ is proportional to horizontal root acceleration at the fixed sampling rate. The objective penalizes deviations from the reference dynamics, including spurious jitter, rather than driving acceleration toward zero. Genuine accelerations during starts, stops, and turns therefore remain part of the supervision target.

The resulting objective combines full-action supervision with the two trajectory terms:

$$\mathcal{L}_{\text{CKS}} = \mathcal{L}_{\text{flow}} + \lambda_{\text{path}}\mathcal{L}_{\text{path}} + \lambda_{\text{dyn}}\mathcal{L}_{\text{dyn}}, \tag{4}$$

where $\mathcal{L}_{\text{flow}}$ is the standard flow loss inherited from Cosmos 3 and applied to the complete 144-dimensional action. The path and dynamics objectives act on the same predicted horizontal root trajectory and jointly constrain its spatial structure and temporal evolution. The vertical root position remains supervised by $\mathcal{L}_{\text{flow}}$, without an additional smoothing objective.

3.3 AUXILIARY GEOMETRIC SUPERVISION

To enhance the geometric constraints, we introduce auxiliary geometric losses on the decoded body motion. During training, we recover 22 joints and foot-surface geometry from the root and body action components. The joint-position loss $\mathcal{L}_{\text{joint}}$ matches predicted and reference root-local joint positions to supervise body configuration, while $\mathcal{L}_{\text{jvel}}$ matches joint displacements in the canonical frame to constrain their temporal evolution. $\mathcal{L}_{\text{contact}}$ additionally matches the reference foot displacements rather than enforcing zero velocity or a fixed foot height. These complementary terms form the auxiliary geometric objective:

$$\mathcal{L}_{\text{geo}} = \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} + \lambda_{\text{jvel}}\mathcal{L}_{\text{jvel}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}}. \tag{5}$$

When computing these losses, we decode both the predicted and reference motion using ground-truth camera rotations. This isolates the geometric supervision from camera-rotation prediction errors. The geometric losses backpropagate through the root and body action components, but not through the camera components, which remain supervised by the base action objective.

The overall training combines CKS, auxiliary geometric supervision, and video supervision:

$$\mathcal{L} = \lambda_A\big(\mathcal{L}_{\text{CKS}} + \mathcal{L}_{\text{geo}}\big) + \mathcal{L}_{\text{video}}, \tag{6}$$

where λ_A balances action and video supervision. The video loss $\mathcal{L}_{\text{video}}$ applies only to unobserved latents, and each loss is normalized over its active targets.

3.4 Full-Sequence Inference Optimization

We introduce full-sequence inference optimization to reduce residual motion jitter while limiting deviations from the model predictions. Window predictions are aligned and fused to initialize the complete sequence. We then refine the decoded root positions, root rotations, and parent-relative body rotations by balancing prediction fidelity with temporal regularity:

$$\mathcal{E}_{\text{inf}} = \mathcal{E}_{\text{anchor}} + \mathcal{E}_{\text{angular}} + \mathcal{E}_{\text{root}}. \tag{7}$$

The anchoring term $\mathcal{E}_{\text{anchor}}$ uses Smooth L1 penalties on root-position deviations and rotation corrections relative to the fused predictions. The angular term $\mathcal{E}_{\text{angular}}$ penalizes changes in frame-to-frame rotation increments, discouraging abrupt variations in rotational motion. The root term $\mathcal{E}_{\text{root}}$ penalizes second- and third-order differences of root positions to regularize acceleration and jerk. Root rotations are evaluated in the shared canonical frame, while body rotations remain parent-relative. Each term includes its component weights, and all penalties are averaged over their active entries. Starting from the fused predictions, we use L-BFGS to optimize root positions and rotation-vector corrections over the complete sequence. The first and last states remain fixed, preserving the predicted boundary positions and poses while refining the interior motion. The optimization uses only model predictions and leaves network parameters unchanged.

4 Experiments

4.1 Experimental Setup

Dataset and split. We construct the HumanWM-Nymeria split from 573 converted recordings of Nymeria (Ma et al., 2024), with 500 recordings for training and 57 recordings from 18 participants for testing. The remaining 16 recordings belong to these test participants and are excluded from both training and scored evaluation, ensuring a person-disjoint split.

Evaluation metrics. For motion evaluation, we resample baseline predictions to the 15 FPS timestamps of HumanWM and evaluate all methods on a common set of 21 body joints. For reconstruction, we report globally aligned mean per-joint position error (GA-MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), and pelvis-aligned MPJPE, all in millimeters. GA-MPJPE uses a single similarity transform over the entire sequence, whereas PA-MPJPE aligns each frame independently. Pelvis-aligned MPJPE removes root translation in each frame. We evaluate trajectory accuracy using mean and final root position errors in meters, after a single translation aligning the initial root positions. To assess motion dynamics, we report joint and root acceleration errors (J-Accel and R-Accel, in m/s^2) and root jerk error (R-Jerk, in m/s^3) relative to ground truth.

For video generation, PSNR, SSIM (Wang et al., 2004), and DreamSim (Fu et al., 2023) measure agreement with recorded futures. We additionally use RAFT (Teed & Deng, 2020) optical flow and the Dynamic Degree and Motion Smoothness metrics from VBench (Huang et al., 2024) to compare motion statistics between generated and recorded videos. We report absolute log-ratio motion-magnitude error, absolute smoothness error, seed-level Dynamic Degree agreement, and task-level Spearman correlation between generated and recorded motion magnitudes.

Implementation details. We train HumanWM with 480p inputs on eight A100 80GB GPUs using FSDP and bfloat16 precision with a global batch size of eight. We use Adam with a base learning rate of 10^{-5} and a $10\times$ learning-rate multiplier for the action projections and action modality embedding. At inference, we use the UniPC sampler with 30 sampling steps and a shift parameter of 10. For full-sequence reconstruction, we align and fuse predictions from overlapping 33-frame windows with a stride of 16 frames.

4.2 Human Motion Reconstruction

We evaluate full-sequence motion reconstruction on 3,000 sequences sampled from the 57 test recordings. We compare HumanWM with ReViV (Lyu et al., 2026), EgoAllo (Yi et al., 2025), EgoEgo (Li et al., 2023), UniEgoMotion (Patel et al., 2025), and EgoPriMo (Ge et al., 2026).

With RGB-only input, HumanWM outperforms all evaluated baselines across the five pose and root trajectory metrics (Table 1). Compared with RGB-only ReViV, it reduces GA-MPJPE from

Table 1: Full-sequence human motion reconstruction on 3,000 test sequences. Auxiliary inputs refer to task text or ground-truth camera trajectories. "None" indicates RGB-only input without auxiliary conditions. ↓ indicates that lower values are better.

Method	Auxiliary inputs	MPJPE ↓			Root error↓	
		GA-MPJPE	PA-MPJPE	Pelvis-MPJPE	Mean error	Final error
EgoEgo	GT Cam. Traj.	192.32	133.16	186.54	0.504	0.886
EgoAllo	GT Cam. Traj.	188.51	123.25	170.50	0.535	0.944
UniEgoMotion	GT Cam. Traj.	319.50	233.96	394.02	0.543	0.908
EgoPriMo	Text	176.00	75.38	197.01	0.426	0.810
ReViV	None	137.11	91.88	207.95	0.180	0.322
HUMANWM	None	84.78	50.64	91.28	0.117	0.232
	Text	**78.10**	**46.49**	**84.73**	**0.110**	**0.216**

137.11 to 84.78 mm and PA-MPJPE from 91.88 to 50.64 mm, corresponding to reductions of 38.2% and 44.9%, respectively. Mean and final root trajectory errors also decrease from 0.180 to 0.117 m and from 0.322 to 0.232 m.

Adding task text further improves all five pose and root trajectory metrics. GA-MPJPE and PA-MPJPE decrease to 78.10 mm and 46.49 mm, respectively, improving on the RGB-only configuration by 7.9% and 8.2%. The resulting PA-MPJPE is also 38.3% lower than EgoPriMo's 75.38 mm, the lowest baseline error on this metric. The text-conditioned configuration achieves lower GA-MPJPE than ReViV on 94.73% of test sequences, indicating that the improvement is not limited to a small subset. Figure 3 compares HUMANWM with EgoEgo and ReViV on five example sequences, showing both body configurations and root trajectories.

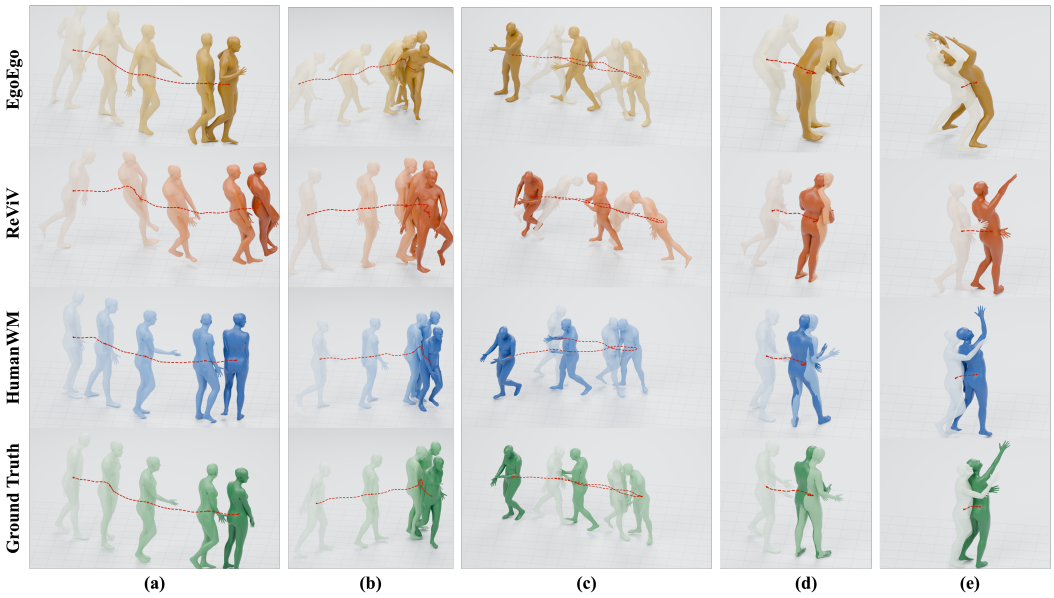

Figure 3: Qualitative comparison of human motion reconstruction on five test sequences, shown in columns (a)–(e). Rows show EgoEgo, ReViV, HUMANWM, and ground truth. Sampled poses are overlaid to visualize motion over time, and red dashed curves trace the pelvis (root) trajectories.

Both HUMANWM configurations achieve lower acceleration and jerk errors and less foot skating than all evaluated baselines (Table 2). Compared with EgoEgo, the strongest baseline on acceleration and jerk metrics, the RGB-only configuration reduces root acceleration and root jerk errors by 42.5% and 46.4%, respectively. It also reduces foot skating by 20.2% relative to EgoPriMo, the strongest baseline on this metric.

Table 2: Motion dynamics and foot skating on the same test set and input configurations as Table 1. Lower values are better, and bold indicates the best result.

Method	Auxiliary inputs	J-Accel ↓	R-Accel ↓	R-Jerk ↓	Foot Skate ↓
EgoEgo	GT Cam. Traj.	2.261	1.401	25.832	0.247
EgoAllo	GT Cam. Traj.	2.594	1.740	34.338	0.356
UniEgoMotion	GT Cam. Traj.	8.536	8.624	152.193	0.727
EgoPriMo	Text	2.502	1.903	44.816	0.183
ReViV	None	4.681	2.867	68.889	0.428
HUMANWM	None	**1.278**	**0.806**	**13.843**	**0.146**
	Text	1.451	1.001	18.436	0.150

Text conditioning improves pose and root trajectory accuracy, whereas the RGB-only configuration achieves lower acceleration and jerk errors and slightly less foot skating. Section 4.4 separately examines the effect of full-sequence inference optimization on these temporal metrics.

4.3 VIDEO GENERATION

We evaluate video generation on 100 test windows and compare HUMANWM with CogVideoX-5B-I2V (Yang et al., 2024), Wan2.2-TI2V-5B (Team Wan et al., 2025), and the unadapted Cosmos3-Edge backbone (NVIDIA, 2026). All models receive the same initial image and the literal prompt *First-person view. Task description:* followed by the original task description, without prompt expansion, future RGB frames, or ground-truth actions. Each model generates three samples per test window using different random seeds.

Table 3: Video generation on 100 test windows with three sampling seeds per window. Visual agreement and motion diagnostics are evaluated against recorded futures. Motion metrics exclude the conditioning frame. Bold indicates the best reported mean.

	Visual agreement			Motion diagnostics			
Method	PSNR ↑	SSIM ↑	DreamSim ↓	Motion-mag. error ↓	Smooth. error ↓	Dynamic agree. ↑	Motion rank ρ ↑
CogVideoX-5B-I2V	15.0855	0.4233	0.2251	1.0725	0.0205	0.6433	0.2868
Wan2.2-TI2V-5B	14.2372	0.3847	0.2436	0.8474	0.0200	0.7500	0.2448
Cosmos3-Edge	14.8231	0.3630	0.2441	0.8689	0.0140	0.6900	0.3088
HUMANWM	**15.6491**	**0.4475**	**0.2147**	**0.4831**	**0.0115**	**0.8700**	**0.6501**

HUMANWM achieves the best reported mean on all seven metrics in Table 3. Relative to the unadapted Cosmos3-Edge backbone, it improves PSNR by 0.8260 dB and SSIM by 0.0845, while reducing DreamSim by 0.0294. The paired 95% participant-cluster confidence intervals for all three differences exclude zero on this evaluation subset.

The visual improvements are accompanied by closer agreement with the recorded motion statistics. HUMANWM's mean motion magnitude, estimated using RAFT, is 14.45, close to 14.86 in the recorded futures. Its motion-magnitude error is 0.4831, compared with 0.8689 for Cosmos3-Edge. It also achieves lower Motion Smoothness error, higher Dynamic Degree agreement, and stronger task-level motion-rank correlation than the evaluated baselines.

Motion-stratified evaluation further shows that the gains persist when the recorded future contains appreciable motion. For medium- and high-motion windows, HUMANWM improves PSNR over Cosmos3-Edge by 1.46 and 1.08 dB and SSIM by 0.104 and 0.111, respectively. Motion-magnitude error decreases from 0.790 to 0.490 in the medium-motion group and from 1.381 to 0.369 in the high-motion group. Cosmos3-Edge nevertheless achieves slightly lower motion-magnitude error in the low-motion group and higher balanced Dynamic Degree accuracy overall. Together, these diagnostics support the interpretation that the visual gains are not explained solely by reduced motion. Additional results and evaluation details are provided in Appendix B.2.

4.4 Ablation Studies

Table 4 examines multi-task learning, Coupled Kinematic Supervision (CKS), and full-sequence inference optimization on the same 3,000 test sequences. CKS combines the hybrid grounded gravity–view representation with path-structure and acceleration supervision. The *w/o Coupled Trajectory Losses* variant retains the hybrid grounded gravity–view representation and the base action flow objective, but removes supervision of the overall locomotion path and local dynamics.

Table 4: Ablation studies on the same reconstruction test set. Recon. denotes reconstruction-only training; Multi-task includes reconstruction, generation, and forecasting. Inf. opt. indicates whether full-sequence inference optimization is enabled. The full-model row uses task-text conditioning. Bold indicates the best result at the reported precision.

Variant	Training tasks	Inf. opt.	GA-MPJPE ↓	PA-MPJPE ↓	Root Mean ↓	R-Accel ↓	Foot Skate ↓
w/o CKS	Recon.	Off	115.65	53.24	0.216	7.724	0.361
w/o CKS	Recon.	On	114.42	51.86	0.215	1.919	0.210
w/o CKS	Multi-task	On	103.79	51.02	0.187	1.751	0.180
w/o Coupled Trajectory Losses	Multi-task	On	83.56	48.27	0.121	1.030	0.161
w/o Inference Optimization	Multi-task	Off	78.49	46.60	0.113	3.088	0.232
HumanWM (Full)	Multi-task	On	**78.10**	**46.49**	**0.110**	**1.001**	**0.150**

The reconstruction-only and multi-task settings use the same Nymeria training data. Reconstruction-only training observes the complete video window, whereas multi-task training additionally uses initial-frame and video-prefix conditioning to jointly predict actions and unobserved video. Thus, multi-task training changes how the paired data are used rather than introducing additional training data. With inference optimization enabled, the multi-task configuration without CKS outperforms its reconstruction-only counterpart across all five metrics. GA-MPJPE decreases from 114.42 to 103.79 mm, PA-MPJPE from 51.86 to 51.02 mm, and mean root error from 0.215 to 0.187 m, alongside lower root acceleration error and foot skating. This comparison favors the multi-task configuration, although training-time text conditioning also differs between the two settings. The *w/o Coupled Trajectory Losses* variant achieves a GA-MPJPE of 83.56 mm and a mean root error of 0.121 m. The full model achieves 78.10 mm and 0.110 m, respectively, together with the lowest PA-MPJPE and root acceleration error among the evaluated variants. These results are consistent with the benefit of CKS for root-trajectory structure and temporal dynamics.

To isolate the effect of full-sequence inference optimization, each before-and-after comparison uses identical checkpoint weights and cached window predictions, with the same window alignment and fusion. For the full model, optimization reduces root acceleration error from 3.088 to 1.001 m/s^2 and foot skating from 0.232 to 0.150 m/s, corresponding to reductions of 67.6% and 35.3%, respectively. Pose and root trajectory errors change only slightly: GA-MPJPE decreases from 78.49 to 78.10 mm, and mean root error decreases from 0.113 to 0.110 m. The reconstruction-only checkpoint exhibits the same pattern: root acceleration error decreases from 7.724 to 1.919 m/s^2 and foot skating from 0.361 to 0.210 m/s, with smaller changes in pose errors. These controlled comparisons show that inference optimization primarily reduces acceleration error and foot skating while retaining pose and root trajectory accuracy.

5 Conclusion

HumanWM adapts the native action interface of a pretrained world model to structured human motion, jointly modeling wearer actions and egocentric visual dynamics. Coupled Kinematic Supervision combines gravity–view grounding with multi-scale root trajectory supervision, while full-sequence inference optimization balances prediction fidelity with temporal regularity. Experiments on Nymeria demonstrate improved pose and root trajectory reconstruction, lower acceleration errors and foot skating, and closer agreement between generated videos and recorded futures, using a single checkpoint for motion reconstruction and video generation. Together, these results highlight the value of combining pretrained visual-dynamics priors with structured motion representations and kinematic supervision for egocentric human motion modeling.

AI USE STATEMENT

An AI coding and writing assistant helped inspect experiment records, organize the manuscript, draft and edit LaTeX prose, and check compilation. The authors defined the research questions and methods, ran the experiments, supplied the source records, and remain responsible for verifying every technical statement, citation, and numerical claim before submission.

REFERENCES

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.09344.

Haoyang Ge, Peng Ren, Yukun Shi, Cong Huang, Kun Li, and Kai Chen. Egoprimo: Egocentric motion generation for interactive humanoid control. *arXiv preprint arXiv:2606.08495*, 2026. URL https://arxiv.org/abs/2606.08495.

Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://github.com/songweige/content-debiased-fvd.

Vladimir Guzov, Yifeng Jiang, Fangzhou Hong, Gerard Pons-Moll, Richard Newcombe, C. Karen Liu, Yuting Ye, and Lingni Ma. HMD^2: Environment-aware motion generation from single egocentric head-mounted device. In *International Conference on 3D Vision*, pp. 1394–1405, 2025. doi: 10.1109/3DV66043.2025.00132.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VBench_Comprehensive_Benchmark_Suite_for_Video_Generative_Models_CVPR_2024_paper.html.

Jiaxi Jiang, Bharat Lal Bhatnagar, Nan Yang, Lingni Ma, Sebastian Starke, Robin Kips, Nadine Bertsch, Christian Holz, and Federica Bogo. EgoExoMoCap: Distributed ego-exo human motion capture. *arXiv preprint arXiv:2607.15868*, 2026. URL https://arxiv.org/abs/2607.15868.

Baoyu Li, Xinchen Yin, Mengying Lin, Yixin Zhang, and Danfei Xu. Egowam: World action models beyond pixels with in-the-wild egocentric human data. *arXiv preprint arXiv:2607.08436*, 2026. URL https://arxiv.org/abs/2607.08436.

Jiaman Li, C. Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17142–17151, 2023.

Xiaozhong Lyu, Gen Li, Zhiyin Qian, Xucong Zhang, Marc Pollefeys, and Siyu Tang. ReViV: Reconstructing the viewer and the view in 4D from monocular egocentric video. In *European Conference on Computer Vision*, 2026.

Lingni Ma, Yuting Ye, Fangzhou Hong, Vladimir Guzov, Yifeng Jiang, Rowan Postyeni, Luis Pesqueira, Alexander Gamino, Vijay Baiyya, Hyo Jin Kim, Kevin Bailey, David S. Fosas, C. Karen Liu, Ziwei Liu, Jakob Engel, Renzo De Nardi, and Richard Newcombe. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. In *European Conference on Computer Vision*, pp. 445–465. Springer, 2024. doi: 10.1007/978-3-031-72691-0_25.

NVIDIA. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026. URL https://research.nvidia.com/labs/cosmos-lab/cosmos3/technical-report.pdf.

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. *arXiv preprint arXiv:2508.01126*, 2025. URL https://arxiv.org/abs/2508.01126.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10967–10977, 2019. doi: 10.1109/CVPR.2019.01123.

Davis Rempe, Mathis Petrovich, Ye Yuan, Haotian Zhang, Xue Bin Peng, Yifeng Jiang, Tingwu Wang, Umar Iqbal, David Minor, Michael de Ruyter, Jiefeng Li, Chen Tessler, Edy Lim, Eugene Jeong, Sam Wu, Ehsan Hassani, Michael Huang, Jin-Bey Yu, Chaeyeon Chung, Lina Song, Olivier Dionne, Jan Kautz, Simon Yuen, and Sanja Fidler. Kimodo: Scaling controllable human motion generation. *arXiv preprint arXiv:2603.15546*, 2026. URL https://arxiv.org/abs/2603.15546.

Zehong Shen, Huaijin Pi, Yan Xia, Zhi Cen, Sida Peng, Zechen Hu, Hujun Bao, Ruizhen Hu, and Xiaowei Zhou. World-grounded human motion recovery via gravity-view coordinates. In *SIGGRAPH Asia Conference Papers*, pp. 1–11, 2024. doi: 10.1145/3680528.3687565.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://github.com/Wan-Video/Wan2.2.

Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, pp. 402–419. Springer, 2020. doi: 10.1007/978-3-030-58536-5_24. URL https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/3526_ECCV_2020_paper.php.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. URL https://www.cns.nyu.edu/~lcv/ssim/.

Jingqiao Xiu, Fangzhou Hong, Yicong Li, Mengze Li, Wentao Wang, Sirui Han, Liang Pan, and Ziwei Liu. Egotwin: Dreaming body and view in first person. *arXiv preprint arXiv:2508.13013*, 2025. URL https://arxiv.org/abs/2508.13013.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. CogVideoX: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. URL https://arxiv.org/abs/2408.06072.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, Ayaan Malik, Kyungmin Lee, William Liang, Nadun Ranawaka, Jiasheng Gu, Yinzhen Xu, Guanzhi Wang, Fengyuan Hu, Avnish Narayan, Johan Bjorck, Jing Wang, Gwanghyun Kim, Dantong Niu, Ruijie Zheng, Yuqi Xie, Jimmy Wu, Qi Wang, Ryan Julian, Danfei Xu, Yilun Du, Yevgen Chebotar, Scott Reed, Jan Kautz, Yuke Zhu, Linxi Jim Fan, and Joel Jang. World action models are zero-shot policies, 2026. URL https://arxiv.org/abs/2602.15922.

Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pp. 7072–7084, 2025. doi: 10.1109/CVPR52734.2025.00663.

Jia Zheng, Teli Ma, Yudong Fan, Zifan Wang, Shuo Yang, and Junwei Liang. Motionwam: Towards foundation world action models for real-time humanoid loco-manipulation. *arXiv preprint arXiv:2606.09215*, 2026. URL https://arxiv.org/abs/2606.09215.

APPENDIX

A ACTION AND TRAINING DETAILS

Table 5 summarizes the 144D action representation. Continuous 6D rotations are converted to matrices by column-wise Gram–Schmidt orthonormalization.

Table 5: HumanWM action layout. Each action describes the body state at RGB frame t, with the camera block describing the transition from $t-1$ to t.

Channels	Width	Meaning
0:9	9	camera translation + relative rotation
9:12	3	canonical root XY + support-relative Z
12:18	6	root rotation in per-frame GV coordinates
18:144	126	21 parent-relative body rotations

Single-task and multi-task training. Both settings use the same Nymeria training split and paired RGB–motion data; multi-task training introduces no additional dataset or motion annotations. Reconstruction-only training conditions on the complete video window and predicts its action sequence. Multi-task training varies the observed video through task-specific conditioning masks: reconstruction observes the full window, generation observes only the initial frame, and forecasting observes the first nine frames of a 33-frame window. These masks specify clean conditioning video latents; the remaining latents are noised prediction targets rather than black RGB frames. All three tasks supervise the complete action sequence, including actions within the observed prefix, while video prediction is supervised only at unobserved latents. Windows are deterministically assigned to reconstruction, generation, and forecasting with target proportions of 0.4, 0.3, and 0.3 using a 64-bit BLAKE2b hash of `recording:episode:window:42`. The multi-task modes share the 144D action domain and use the original nonempty task descriptions; malformed or multiply labeled episodes are excluded. In the reported single-task versus multi-task ablation, the former uses a fixed training instruction whereas the latter uses episode-specific text, so this comparison evaluates the training configurations rather than isolating task mixing alone.

B ADDITIONAL EXPERIMENTAL DETAILS AND RESULTS

B.1 METRIC AND STATISTICAL DETAILS

We evaluate $J = 21$ joints: the first 22 SMPL-X joints excluding index 3. Let $\widehat{\mathbf{x}}_{t,j}, \mathbf{x}_{t,j} \in \mathbb{R}^3$ be predicted and reference positions in millimeters over T common valid frames. Padding, repeated tail frames, and extrapolated frames are excluded from alignment and scoring; sequence-level metrics are averaged with equal sequence weights.

Globally aligned MPJPE (GA-MPJPE) first fits one similarity transform to all valid frames and joints of a complete sequence:

$$(s^\star, \mathbf{R}^\star, \mathbf{b}^\star) = \underset{\substack{s>0,\ \mathbf{R}\in \mathrm{SO}(3)\\ \mathbf{b}\in\mathbb{R}^3}}{\arg\min} \sum_{t=1}^{T}\sum_{j=1}^{J} \|s\mathbf{R}\widehat{\mathbf{x}}_{t,j} + \mathbf{b} - \mathbf{x}_{t,j}\|_2^2 . \tag{8}$$

The fitted scale, rotation, and translation are shared across all frames:

$$\text{GA-MPJPE} = \frac{1}{TJ}\sum_{t=1}^{T}\sum_{j=1}^{J} \|s^\star\mathbf{R}^\star\widehat{\mathbf{x}}_{t,j} + \mathbf{b}^\star - \mathbf{x}_{t,j}\|_2 . \tag{9}$$

GA-MPJPE removes one sequence-wide similarity transform, retaining residual pose and trajectory errors but not evaluating absolute world position, orientation, or scale.

Procrustes-aligned MPJPE (PA-MPJPE) instead fits an independent similarity transform in each frame:

$$(s_t^\star, \mathbf{R}_t^\star, \mathbf{b}_t^\star) = \underset{\substack{s>0,\ \mathbf{R}\in \mathrm{SO}(3)\\ \mathbf{b}\in\mathbb{R}^3}}{\arg\min} \sum_{j=1}^{J} \|s\mathbf{R}\widehat{\mathbf{x}}_{t,j} + \mathbf{b} - \mathbf{x}_{t,j}\|_2^2, \quad t = 1, \ldots, T, \tag{10}$$

$$\text{PA-MPJPE} = \frac{1}{TJ} \sum_{t=1}^{T} \sum_{j=1}^{J} \|s_t^\star \mathbf{R}_t^\star \widehat{\mathbf{x}}_{t,j} + \mathbf{b}_t^\star - \mathbf{x}_{t,j}\|_2 . \tag{11}$$

PA-MPJPE removes framewise rotation, translation, and scale, so it does not assess global trajectories or temporal consistency of orientation and scale. Both alignments use SVD-based least-squares Sim(3) fitting with $\det \mathbf{R} = 1$; fitting minimizes squared distances, while MPJPE averages unsquared distances without temporal alignment or joint reassignment.

Root trajectory errors align only the initial pelvis positions by translation; pelvis-aligned MPJPE instead subtracts each frame's pelvis, without rotation or scale fitting. Acceleration and jerk errors compare predicted and reference finite differences in physical units using the actual frame interval.

Foot skating is horizontal foot speed on ground-truth contact frames, defined by height within 0.05 m of the fifth-percentile foot height and vertical speed at most 0.20 m/s.

B.2 Video Generation Protocol and Additional Results

Video evaluation uses 100 windows selected before generation from 44 recordings and 18 participants, with 5–6 windows per participant.

Models share the prompt in Section 4.3 but retain their native inference settings. Inputs preserve aspect ratio; scoring removes padding to retain a common field of view. We score 16 future timestamps $t_k = k/8$ seconds ($k = 1, \ldots, 16$), selecting nearest frames by presentation timestamp without interpolation or playback-rate changes.

PSNR and SSIM use RGB values in $[0, 255]$; SSIM uses Gaussian weighting ($\sigma = 1.5$) and population covariance. DreamSim uses its three-backbone ensemble and 224×224 preprocessing. Frame-level scores are averaged within videos, then over seeds within tasks, and finally over tasks. Motion magnitude m is the temporal average of the largest 5% of RAFT flow magnitudes per adjacent frame pair; its error is $|\log((m_{\text{pred}} + 0.1)/(m_{\text{gt}} + 0.1))|$. VBench supplies Motion Smoothness scores and Dynamic Degree labels; we report absolute score differences and label agreement against recorded futures, averaging seeds within tasks. Motion bands use thresholds 9.964 and 19.222 fixed from the 1,098-task candidate pool's recorded futures before inspecting model rankings.

Table 6: Video metric means and 95% participant-cluster bootstrap intervals on 100 tasks, after averaging three seeds per task. Bold denotes the best mean; gray highlights HumanWM. Intervals are not adjusted for multiple comparisons.

Method	PSNR ↑	SSIM ↑	DreamSim ↓
CogVideoX-5B-I2V	15.0855 [14.3960, 15.7811]	0.4233 [0.3945, 0.4529]	0.2251 [0.2041, 0.2454]
Wan2.2-TI2V-5B	14.2372 [13.6682, 14.7876]	0.3847 [0.3575, 0.4124]	0.2436 [0.2218, 0.2673]
Cosmos3-Edge	14.8231 [14.2683, 15.3242]	0.3630 [0.3383, 0.3875]	0.2441 [0.2207, 0.2683]
HumanWM	**15.6491** [15.0313, 16.2011]	**0.4475** [0.4140, 0.4800]	**0.2147** [0.1963, 0.2344]

Mean motion magnitudes are 14.86 for recorded futures and 14.45/8.18/10.61/8.85 for HumanWM/Cosmos3-Edge/Wan2.2/CogVideoX. Including the conditioning frame preserves the ranking of all four motion metrics. The alternative task-majority balanced Dynamic Degree accuracy is 0.708 for HumanWM and 0.741 for Cosmos3-Edge, unlike the seed-level agreement reported in the main table. Paired differences from Cosmos3-Edge are -0.3858 $[-0.4982, -0.2558]$ for motion-magnitude error, -0.0026 $[-0.0043, -0.0009]$ for smoothness error, and $+0.1800$ $[0.1023, 0.2484]$ for Dynamic Degree agreement.

The low/medium/high-motion groups contain 34/38/28 tasks. In this order, HumanWM versus Cosmos3-Edge obtains motion errors of 0.569/0.490/0.369 versus 0.535/0.790/1.381,

SSIM of 0.417/0.462/0.465 versus 0.376/0.358/0.354, and PSNR of 15.81/15.66/15.44 versus 15.90/14.20/14.36 dB. Medium/high-motion DreamSim is 0.235/0.287 versus 0.279/0.312.

Confidence intervals average seeds within each task, then resample all tasks of each of the 18 participants together over 2,000 cluster-bootstrap replicates. Paired HUMANWM-minus-Cosmos3-Edge differences are 0.8260 dB $[0.4952, 1.1421]$ for PSNR, 0.0845 $[0.0696, 0.0986]$ for SSIM, and -0.0294 $[-0.0404, -0.0190]$ for DreamSim. The DreamSim difference from CogVideoX has an interval containing zero. These experiments compare complete systems under shared image and text inputs while retaining model-specific inference settings. The results should therefore be interpreted as system-level comparisons rather than an isolated assessment of training effects.

Table 7: Supplementary video distribution distances: mean ± sample standard deviation over three seeds, each evaluated on 100 clips. Bold denotes the lowest mean; gray highlights HUMANWM. The feature spaces have different scales; finite-sample estimates are not used for a definitive ranking.

Method	CD-FVD ↓	I3D-400 FVD ↓
CogVideoX-5B-I2V	771.53 ± 61.46	539.63 ± 132.50
Wan2.2-TI2V-5B	758.12 ± 15.54	525.46 ± 43.11
Cosmos3-Edge	777.83 ± 24.43	734.39 ± 63.88
HUMANWM	$\mathbf{592.79} \pm 31.52$	$\mathbf{263.03} \pm 3.77$

CD-FVD (Ge et al., 2024) uses 1,408D VideoMAEv2 features and I3D-400 FVD uses 400D logits, both with population covariance. Although HUMANWM has lower mean distances, 100 clips per seed are fewer than either feature dimension, and seed averaging does not eliminate finite-sample bias; these scores do not establish temporal or physical correctness.

C FULL-SEQUENCE INFERENCE OPTIMIZATION DETAILS

Full-sequence inference aligns and fuses window predictions, then jointly refines root positions and rotations without updating the network. Unlike training-time CKS, it regularizes predicted motion without ground-truth targets.

For a 15-FPS video $\mathbf{V}_{0:T} = \{\mathbf{I}_t\}_{t=0}^{T}$ with at least 33 frames, each 33-frame window predicts 32 action endpoints:

$$\mathcal{W}_k = \{\mathbf{I}_{s_k}, \ldots, \mathbf{I}_{s_k+32}\}, \qquad \mathcal{T}_k = \{s_k + 1, \ldots, s_k + 32\}, \tag{12}$$

Window starts are spaced by 16 frames, with an additional start at $T - 32$ when needed to cover the endpoint without padding. Predictions cover frames $1, \ldots, T$; frame zero is observation-only. Task text is supplied only in the text-conditioned setting; sampling follows Section 4.1. The decoded window motion is

$$\mathcal{M}_k = \left\{ \mathbf{p}_{k,t}, \mathbf{R}^g_{k,t}, \mathbf{R}^b_{k,t,1}, \ldots, \mathbf{R}^b_{k,t,21} \right\}_{t \in \mathcal{T}_k}, \tag{13}$$

where $\mathbf{p}_{k,t}$ and $\mathbf{R}^g_{k,t}$ are canonical root position and rotation, and $\mathbf{R}^b_{k,t,j}$ are parent-relative body rotations.

Each new window is aligned to the accumulated prediction over overlap Ω_k, with current root estimates $\bar{\mathbf{p}}_t, \bar{\mathbf{R}}^g_t$. For N_k valid target frames and $i = 0, \ldots, N_k - 1$, alignment and fusion share the center weights

$$w_{k,s_k+1+i} = \sin^2\left(\frac{\pi(i+1)}{N_k + 1}\right). \tag{14}$$

Relative root rotations determine per-frame yaw offsets:

$$\mathbf{Q}_{k,t} = \bar{\mathbf{R}}^g_t (\mathbf{R}^g_{k,t})^\top, \qquad \theta_{k,t} = \operatorname{atan2}([\mathbf{Q}_{k,t}]_{2,1}, [\mathbf{Q}_{k,t}]_{1,1}). \tag{15}$$

Their weighted circular mean gives the window alignment:

$$\theta_k = \operatorname{atan2}\left(\sum_{t \in \Omega_k} w_{k,t} \sin\theta_{k,t}, \sum_{t \in \Omega_k} w_{k,t} \cos\theta_{k,t} \right), \qquad \mathbf{A}_k = R_z(\theta_k). \tag{16}$$

After yaw alignment, horizontal translation is the componentwise median overlap residual:

$$\mathbf{b}_k^{xy} = \mathrm{median}_{t\in\Omega_k}\left[\bar{\mathbf{p}}_t^{xy} - (\mathbf{A}_k\mathbf{p}_{k,t})^{xy}\right]. \tag{17}$$

With $\mathbf{b}_k = [(\mathbf{b}_k^{xy})^\top, 0]^\top$,

$$\mathbf{p}'_{k,t} = \mathbf{A}_k\mathbf{p}_{k,t} + \mathbf{b}_k, \qquad \mathbf{R}^{g\prime}_{k,t} = \mathbf{A}_k\mathbf{R}^g_{k,t}. \tag{18}$$

Only yaw and horizontal translation change; gravity, root height, and local joint rotations are preserved. The first window defines the common reference.

For overlapping windows $\mathcal{K}_t = \{k : t \in \mathcal{T}_k\}$, root positions use a weighted Euclidean mean:

$$\bar{\mathbf{p}}_t = \frac{\sum_{k\in\mathcal{K}_t} w_{k,t}\mathbf{p}'_{k,t}}{\sum_{k\in\mathcal{K}_t} w_{k,t}}. \tag{19}$$

Root and local joint rotations use a weighted geodesic mean:

$$\bar{\mathbf{R}}_t = \operatorname*{arg\,min}_{\mathbf{R}\in\mathrm{SO}(3)} \sum_{k\in\mathcal{K}_t} w_{k,t}\left\|\mathrm{Log}\big(\mathbf{R}^\top\mathbf{R}'_{k,t}\big)\right\|_2^2. \tag{20}$$

Here Log returns a rotation vector and $\mathbf{R}'_{k,t}$ is an aligned root or unchanged local rotation. Starting from the highest-weight rotation, we iteratively average relative rotation vectors and apply their exponential map, stopping below an update norm of 10^{-10} or after 20 iterations before projection onto $\mathrm{SO}(3)$.

We refine fused positions $\mathbf{p}_t^0$ and rotations $\mathbf{R}_{t,j}^0$ over all valid frames, with $J = 22$ rotations ($j = 0$ for the root). Optimization variables are root positions $\widetilde{\mathbf{p}}_t$ and rotation-vector corrections $\boldsymbol{\delta}_{t,j}$:

$$\widetilde{\mathbf{R}}_{t,j} = \mathbf{R}_{t,j}^0\,\mathrm{Exp}(\boldsymbol{\delta}_{t,j}), \qquad \boldsymbol{\omega}_{t,j} = \mathrm{Log}\Big(\widetilde{\mathbf{R}}_{t,j}^\top\widetilde{\mathbf{R}}_{t+1,j}\Big). \tag{21}$$

The exponential map preserves valid rotations, and $\boldsymbol{\omega}_{t,j}$ denotes inter-frame angular increments. Temporal penalties use 15-FPS frame differences without physical-unit rescaling. For residual tensor $\mathbf{x}$, the elementwise mean Smooth L1 penalty is

$$\mathrm{SL1}_\beta(\mathbf{x}) = \frac{1}{|\mathbf{x}|}\sum_u h_\beta(x_u), \qquad h_\beta(x) = \begin{cases} x^2/(2\beta), & |x| < \beta, \\ |x| - \beta/2, & \text{otherwise.} \end{cases} \tag{22}$$

The refinement objective is

$$\begin{aligned} \mathcal{E}_{\mathrm{inf}} = {} & \lambda_{\mathrm{rot}}\,\mathrm{SL1}_{0.1}\left(\boldsymbol{\delta}_{2:T-1}\right) \\ & + \frac{\lambda_{\mathrm{ang}}}{3J(T-2)}\sum_{t=1}^{T-2}\sum_{j=0}^{J-1}\|\boldsymbol{\omega}_{t+1,j} - \boldsymbol{\omega}_{t,j}\|_2^2 \\ & + \lambda_{\mathrm{trans}}\,\mathrm{SL1}_{0.01}\left(\widetilde{\mathbf{P}} - \mathbf{P}^0\right) \\ & + \frac{\lambda_{\mathrm{acc}}}{3(T-2)}\sum_{t=1}^{T-2}\left\|\Delta^2\widetilde{\mathbf{p}}_t\right\|_2^2 + \frac{\lambda_{\mathrm{jerk}}}{3(T-3)}\sum_{t=1}^{T-3}\left\|\Delta^3\widetilde{\mathbf{p}}_t\right\|_2^2, \end{aligned} \tag{23}$$

where $\widetilde{\mathbf{P}}$ and $\mathbf{P}^0$ collect refined and fused positions, and $\boldsymbol{\delta}_{2:T-1}$ contains interior-frame corrections. The forward difference Δ gives $\Delta^2\widetilde{\mathbf{p}}_t = \widetilde{\mathbf{p}}_{t+2} - 2\widetilde{\mathbf{p}}_{t+1} + \widetilde{\mathbf{p}}_t$ and $\Delta^3\widetilde{\mathbf{p}}_t = \Delta^2\widetilde{\mathbf{p}}_{t+1} - \Delta^2\widetilde{\mathbf{p}}_t$; denominators normalize over active entries. Anchoring limits changes to the prediction, while temporal terms regularize angular-increment changes, root acceleration, and jerk rather than matching ground-truth derivatives. We use fixed weights $(\lambda_{\mathrm{rot}}, \lambda_{\mathrm{ang}}, \lambda_{\mathrm{trans}}, \lambda_{\mathrm{acc}}, \lambda_{\mathrm{jerk}}) = (1, 20, 1, 20, 2)$. The first and last predicted states are held fixed:

$$\widetilde{\mathbf{p}}_1 = \mathbf{p}_1^0, \qquad \widetilde{\mathbf{p}}_T = \mathbf{p}_T^0, \qquad \boldsymbol{\delta}_{1,j} = \boldsymbol{\delta}_{T,j} = \mathbf{0}. \tag{24}$$

Starting from zero rotation corrections and fused positions, L-BFGS uses learning rate 0.5, at most 80 iterations, strong-Wolfe line search, history size 20, gradient tolerance 10^{-9}, and change tolerance 10^{-12}. Sequences shorter than three frames remain unchanged; jerk is omitted below four frames.

Refinement and scoring use native 15-FPS timestamps. For source-video visualization only, with $\alpha = (\tau - \tau_t)/(\tau_{t+1} - \tau_t)$, translations and rotations are interpolated as

$$\begin{aligned} \mathbf{p}(\tau) &= (1-\alpha)\widetilde{\mathbf{p}}_t + \alpha\widetilde{\mathbf{p}}_{t+1}, \\ \mathbf{R}_j(\tau) &= \mathrm{SLERP}\Big(\widetilde{\mathbf{R}}_{t,j}, \widetilde{\mathbf{R}}_{t+1,j}; \alpha\Big). \end{aligned} \tag{25}$$

Visualization timestamps outside the predicted interval are clamped to its endpoints, without adding scored frames. All inference stages use predictions only; the ground-truth transforms used for GA-MPJPE and PA-MPJPE are evaluation-only.