HUMANWORLD: FROM EXPLORABLE TO HUMAN-CENTRIC INTERACTIVE WORLD MODEL

Anonymous authors
Paper under double-blind review

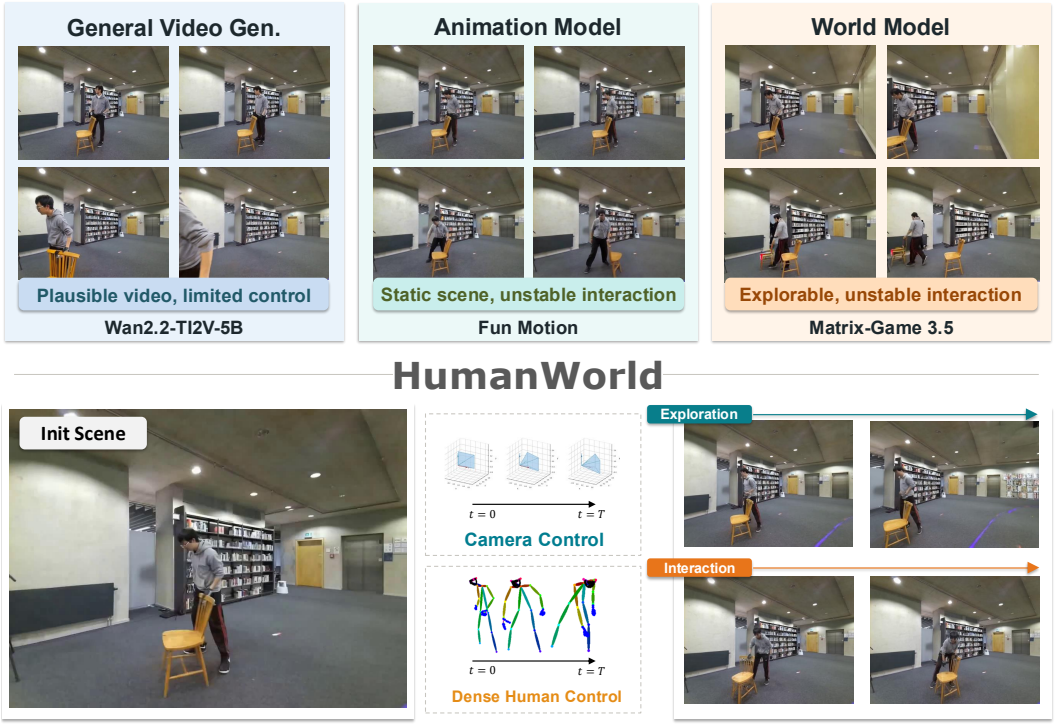


Figure 1: **From world exploration to human-centric interaction.** General video generators offer limited action control, while human animation models and camera-controlled world models produce unstable human–object interactions. HumanWorld combines independent camera trajectories and full-body human motion, letting users explore the scene and change it, e.g., by moving a chair.

ABSTRACT

Recent advances in camera-controlled world models allow users to explore and navigate scenes and observe them from different viewpoints. However, viewpoint control alone does not enable users to act within a scene and change its state through interaction. We introduce HumanWorld, a human-centric interactive world model that enables users to explore a scene through camera control and actively change it by controlling the actions of a human character, including manipulating objects through actions such as picking them up and moving them. Our framework disentangles latent interaction dynamics from visual observation generation, with human motion and camera trajectories independently specified in a shared 3D world coordinate system. We learn a latent world interaction module to predict how controlled human motion changes object and scene states without directly synthesizing RGB observations. The interacted features then condition a pretrained video diffusion model through cross-attention to generate the corresponding visual observations. HumanWorld thus unifies viewpoint exploration, human motion control, and action-driven scene evolution within a framework for human-centric world modeling. Experiments across multiple datasets demonstrate that HumanWorld improves visual fidelity, human motion alignment, and human–scene interaction consistency over existing world models with the disentangled modeling.

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.

1 INTRODUCTION

Recent advances in video generation have improved the visual fidelity, motion coherence, and controllability of videos synthesized from text or images (Yang et al., 2024; Kong et al., 2024; Wan et al., 2025). By learning from large-scale visual data, these models acquire appearance and motion priors that provide a foundation for simulating dynamic environments. Building on these priors, video world models predict future observations conditioned on past observations and user actions, allowing generated environments to evolve in response to external control (Zhu et al., 2026; Mao et al., 2025b; Sun et al., 2025). World models support applications ranging from policy learning and robotic planning (Hafner et al., 2025; Assran et al., 2025) to interactive virtual environments and immersive human experiences (Parker-Holder & Fruchter, 2025; Xie et al., 2026).

Most existing video world models use camera poses as the control signal, enabling users to navigate generated environments, explore unseen regions, and revisit familiar locations (Mao et al., 2025b; Sun et al., 2025; Xu et al., 2026). The resulting worlds are primarily *explorable*: users control how they observe the world, rather than directly specifying actions that change its state. A natural next step is to make these worlds *interactive*, allowing users to act within a scene and influence what happens through their actions (Figure 1). Recent work explores this direction through keyboard-controlled avatars, language-specified events (Parker-Holder et al., 2024; Parker-Holder & Fruchter, 2025; He et al., 2025), or pose-driven human–object interactions (Xie et al., 2026; Ji et al., 2026; Bai et al., 2025b; Tu et al., 2025b). However, coarse commands leave detailed body motion to the generator, while egocentric pose-driven models tie observation to the ego view of the acting person. Disentangling how to observe and how to interact with the world remains an important challenge. We seek to model the scene changes caused by controlled human actions and preserve those changes across viewpoints. For example, when a person picks up an object, the object should move with the hand regardless of the viewpoint. The key is to model how controlled human actions change the scene while preserving their consequences across independent viewpoints.

We introduce **HumanWorld**, a human-centric interactive world model that combines full-body motion control, independent viewpoint control, and human–scene interaction modeling. Our central design disentangles the prediction of latent world-state transitions from the generation of visual observations. Given an initial image, we use the reference camera geometry to lift RGB features into a latent scene representation aligned with human poses in a shared 3D world coordinate system. This representation connects visual scene evidence with the global position, orientation, and articulated motion of the controlled human. A latent world interaction module then predicts how the prescribed motion changes object states and human–scene relations over time, capturing the consequences of actions such as picking up and moving objects without directly synthesizing RGB frames. To observe the evolving world, a learned view-projection module maps the predicted scene states and human pose conditions into features aligned with the target camera viewpoint. These features condition a pretrained video diffusion model through cross-attention, which synthesizes the corresponding visual observations using its learned appearance and motion priors. This disentanglement assigns human actions the role of driving scene evolution and camera trajectories the role of specifying its observation, unifying full-body control and interaction-driven world modeling. Experiments on camera control, human animation, and human–scene interaction validate these capabilities and the roles of latent interaction modeling and camera-aware scene representation.

2 RELATED WORK

Explorable World Models. Video generation models synthesize temporally coherent visual sequences from text prompts or reference images, learning appearance and motion priors from large-scale video data (Ho et al., 2022b;a; Blattmann et al., 2023). Recent diffusion-transformer models combine spatiotemporal latent compression with scalable generative modeling to improve visual fidelity, motion coherence, and semantic alignment (Yang et al., 2024; Kong et al., 2024; Wan et al., 2025). Recent video world models turn pretrained video generators into *explorable* environments, allowing users to navigate generated scenes through camera controls or keyboard and mouse inputs (Zhu et al., 2026; Mao et al., 2025b; He et al., 2025). Progress in streaming generation further supports responsive exploration over longer horizons while preserving contextual consistency (He et al., 2025; Robbyant Team et al., 2026). Its central challenge is to generate coherent observations as users move through a scene, reveal unseen regions, and return to familiar locations (Sun et al., 2025;

Xu et al., 2026). These exploration capabilities address how users observe and navigate a world, whereas fine-grained control of articulated human motion and its effects on the surrounding scene poses an additional modeling challenge. HumanWorld extends this formulation with a human action interface, modeling how 3D human motion changes object states and human–scene relations.

Interactive World Models. Recent methods begin to explore interactive world models, which differ in the granularity of human control, the choice of observation viewpoint, and the representation of interaction dynamics. Keyboard-driven avatars and language-specified events allow users to influence generated worlds, but leave detailed full-body articulation to the generator (Parker-Holder et al., 2024; Parker-Holder & Fruchter, 2025; He et al., 2025). Tracked head–hand control enables fine-grained object manipulation from an egocentric viewpoint (Xie et al., 2026), while upper-body interaction models combine local human motion with explicit contact-state commands (Ji et al., 2026). Whole-body pose conditioning extends the action interface to coordinated human movement, although the generated observations remain tied to the perspective of the acting person (Bai et al., 2025b; Tu et al., 2025b). Third-person generation combines articulated human-motion control with independently specified camera trajectories, emphasizing motion adherence and long-horizon consistency (Wang et al., 2026). HumanWorld unifies *full-body motion control*, *independent viewpoint control*, and *human–scene interaction modeling* in a general interactive world model.

Human Animation. Our work is also related to human animation, which synthesizes videos of a reference person from driving signals such as pose sequences, motion videos, or audio (Hu et al., 2024; Zhang et al., 2025; Lin et al., 2025). Its central objective is to produce faithful and natural human performances, emphasizing identity and appearance preservation, motion alignment, and temporal coherence (Xu et al., 2023; Tu et al., 2025a; Zhang et al., 2025). Others enrich motion guidance with 3D human geometry (Zhu et al., 2024) or jointly condition animation on human and camera motion (Wang et al., 2024). Recent extensions also incorporate environmental context and interacting objects to improve the compatibility of the animated person with the scene (Hu et al., 2025). HumanWorld focuses on the environmental consequences of controlled human actions. Given an initial scene and a full-body motion sequence, we predict how the actions change object states and human–scene relations, and how these changes persist over time across viewpoints.

3 Method

3.1 From Explorable to Interactive World Models

Let I_0 be an initial observation and $V = (I_1, \dots, I_T)$ the future video. A video world model generates V conditioned on a user control signal $\mathcal{U}$:

$$p_\theta(V \mid I_0, \mathcal{U}). \tag{1}$$

Most existing video world models emphasize camera control: users specify a viewpoint trajectory to navigate a scene, reveal unseen regions, or revisit observed places (Mao et al., 2025b; Sun et al., 2025; Xu et al., 2026; He et al., 2025). Their primary control is thus *exploration*: choosing where to observe the world. Subsequent models become interactive by introducing keyboard actions, language-specified events, or body-pose controls that also influence what happens in the scene (Parker-Holder et al., 2024; Parker-Holder & Fruchter, 2025; Xie et al., 2026; Bai et al., 2025b).

We think these control signals can be categorized into two disentangled factors that determine the generated video: how the world state evolves and how that state is observed. An interaction control can change the human behaviors and thus the surrounding objects, while a viewpoint control changes the image formed from the resulting state without itself changing that state. We therefore disentangle the control signal as $\mathcal{U} = (u^{\text{int}}_{1:T}, u^{\text{view}}_{0:T})$ and formulate a world model with distinct state-transition and observation processes:

$$\mathbf{s}_0 = \text{Init}(I_0, u^{\text{view}}_0), \qquad \mathbf{s}_t = \text{Dynamics}(\mathbf{s}_{t-1}, u^{\text{int}}_t), \qquad I_t \sim p_\theta(\cdot \mid I_0, \mathbf{s}_t, u^{\text{view}}_t), \tag{2}$$

where $\mathbf{s}_t$ represents the evolving scene, including object states and their relations to an acting agent. The absence of u^{view}_t from the state update expresses the distinction: changing the viewpoint changes the observation, while the same interaction continues to evolve in the underlying world.

HumanWorld instantiates this framework from a human-centric perspective, as shown in Figure 2. Its interaction control is a sequence of 3D human poses $a_{0:T}$, each specifying the global position,

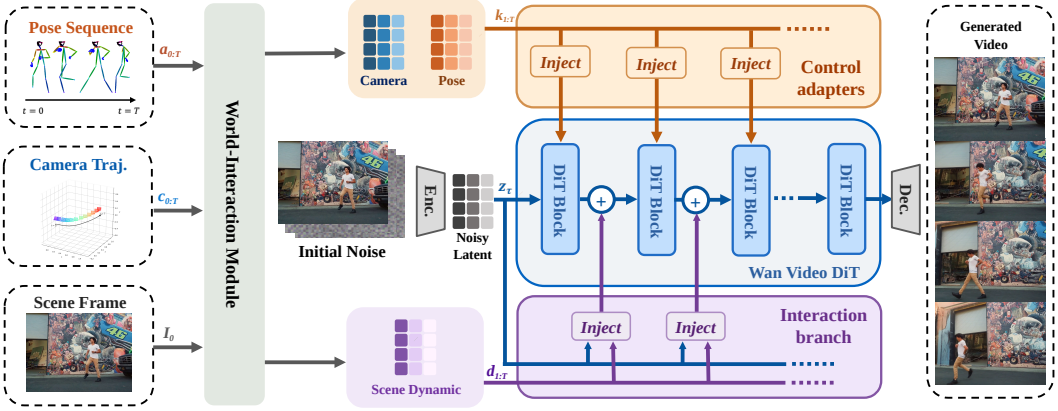


Figure 2: **Overview of HumanWorld.** Given an initial scene frame, a human pose sequence, and a camera trajectory, the World Interaction module produces encoded controls and dynamic scene conditions. Control adapters supply the prescribed motion and viewpoint to the pretrained Wan video DiT, while a separate interaction branch incorporates the scene dynamics through residual updates. The decoder maps the generated video latents to visual observations.

orientation, and articulated body configuration, and the motion between consecutive poses drives scene changes. Its observation control is an independently specified camera trajectory $c_{0:T}$. Both controls are defined in a shared 3D world coordinate system, with $u_t^{\text{int}} = a_{t-1:t}$ and $u_t^{\text{view}} = c_t$. For example, after picking up an object, it should stay with the person when the camera moves.

3.2 Latent Human–Scene Interaction Dynamics

Figure 3 illustrates the latent interaction pathway. The world interaction module takes the initial observation I_0, the world-space human poses $a_{0:T}$, and the camera poses $c_{0:T}$, and produces two conditions for video generation: an encoded control sequence and an evolving interaction state. It first encodes the camera and pose controls, constructs a camera-aware scene state from the initial scene, and then models the human–scene interaction before forming the final scene dynamic. The interaction state is driven by the human motion and conditioned on the camera-aware scene representation.

Human–Scene Representation in World Space. The initial frame provides appearance and scene layout, but its image features are tied to the reference camera c_0. A pretrained video encoder $\mathcal{E}$ extracts the scene latent, while the relative camera trajectory $\bar{c}_t = c_0^{-1} c_t$ is encoded by a camera encoder $\mathcal{C}_\theta^{\text{cam}}$. The Camera Modulation module uses the encoded camera trajectory to modulate the scene latent, producing a camera-aware scene state:

$$\mathbf{e}_t^c = \mathcal{C}_\theta^{\text{cam}}(\bar{c}_t), \qquad \mathbf{s}_t^{\text{cam}} = \mathcal{M}_\theta(\mathcal{E}(I_0), \mathbf{e}_t^c)\,. \tag{3}$$

It is aligned with the shared world coordinate system in which a_t is specified, so the global position and articulated pose relate to scene features. This is a learned latent representation rather than an explicit mesh, and it retains the scene information needed to predict the effects of human motion under different camera trajectories. The pose control is encoded as $\mathbf{e}_t^p = \mathcal{C}_\theta^{\text{pose}}(\Pi_{c_t}(a_t))$.

Action-Conditioned Interaction Dynamics. At each step, the Human-Aware Block uses a set of learnable queries $\mathbf{Q}^h$ to attend to the pose control and the current generated latent, producing an action-aware human state. The Action–Scene Block follows the same query-based design: learnable queries $\mathbf{Q}^s$ attend to the action-aware state and the camera-aware scene state to produce a scene-aware state. The Interaction Block then produces an interaction dynamic:

$$\begin{aligned} \mathbf{h}_t &= \mathcal{H}_\theta\big(\mathbf{Q}^h; \mathbf{e}_t^p, \mathbf{x}_t\big)\,, \\ \mathbf{s}_t^{\text{scene}} &= \mathcal{A}_\theta(\mathbf{Q}^s; \mathbf{h}_t, \mathbf{s}_t^{\text{cam}})\,, \\ \mathbf{r}_t &= \mathcal{I}_\theta(\mathbf{s}_t^{\text{scene}})\,. \end{aligned} \tag{4}$$

Here $\mathbf{x}_t$ denotes the current generated video latent, and $\mathbf{r}_t$ summarizes the interaction dynamic inferred from the action-aware and camera-aware states without assigning explicit object identities. The final scene dynamic is obtained by Residual Scene–Relation Fusion:

$$\mathbf{s}_t = \mathbf{s}_t^{\text{cam}} + \mathcal{F}_\theta([\text{LN}(\mathbf{s}_t^{\text{cam}})\,; \mathbf{r}_t])\,. \tag{5}$$

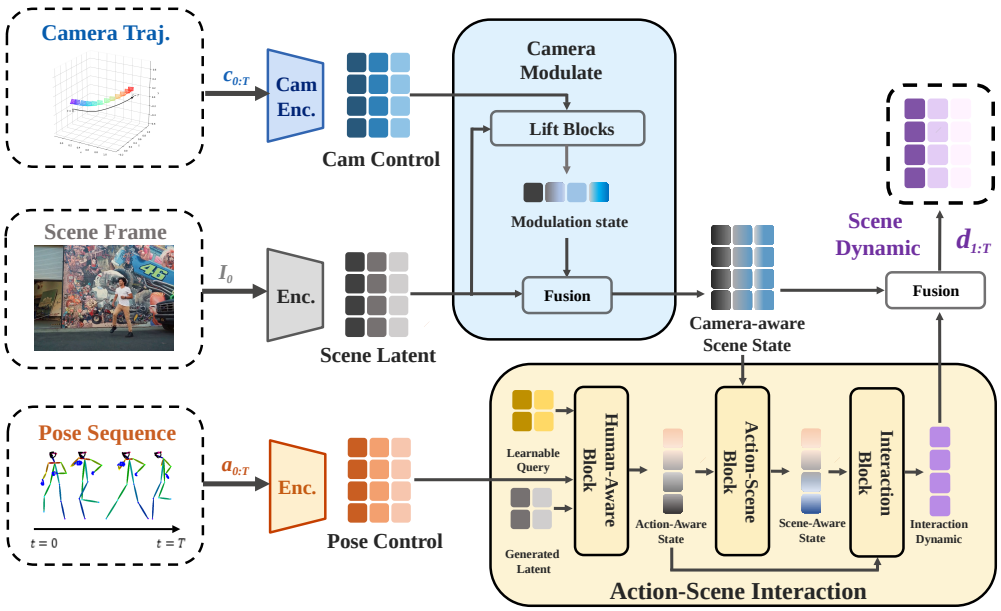


Figure 3: **Latent human–scene interaction dynamics.** Camera Modulation conditions the scene latent on the camera trajectory to form a camera-aware scene state. Human-Aware, Action–Scene, and Interaction blocks relate the pose control and current latent to this state, producing an interaction dynamic. Fusion with the camera-aware state yields the scene dynamic that conditions generation.

Thus the camera determines how the scene memory is read, while pose-conditioned interaction determines the residual that persists. These blocks predict latent features, not RGB frames.

View-Conditioned Control Encoding. To condition the video generator, the camera and pose controls are fused into an encoded control sequence, while the final scene dynamic is kept as a camera-aware interaction memory. A control encoder combines the separately encoded pose and camera controls:

$$\mathbf{k}_t = \mathcal{K}_\theta(\mathbf{e}_t^p, \mathbf{e}_t^c), \qquad \mathbf{d}_t = \mathbf{s}_t. \tag{6}$$

The encoded controls $\mathbf{k}_{1:T}$ guide the pretrained Wan video DiT. The scene dynamic $\mathbf{d}_{1:T}$ and interaction dynamic $\mathbf{r}_{1:T}$ enter a dedicated pathway linking the video latents to the initial scene.

3.3 Interaction-Aware Video Generation

The world interaction module in Section 3.2 maps the initial frame and the human and camera trajectories to encoded controls and scene conditions. $\mathbf{k}_{1:T}$ encodes the projected human pose and target camera, the scene dynamic $\mathbf{d}_{1:T}$ expresses the evolving interaction state in those views, and the interaction dynamic $\mathbf{r}_{1:T}$ provides an additional relation condition. These conditions drive a pretrained Wan video diffusion transformer that renders the interaction as a video.

Video DiT and Control Injection. The video DiT receives a noisy future video latent $\mathbf{z}_\tau$ at flow time τ and predicts the velocity needed to denoise it. The initial frame is encoded as $\mathbf{X}_0 = \mathcal{E}(I_0)$ and provides a fixed visual reference throughout generation. When available, a scene description ℓ enters through the pretrained text conditioning pathway without changing the latent state update. The generator therefore has the following inputs and output:

$$\hat{\mathbf{v}}_\tau = \mathbf{v}_\theta(\mathbf{z}_\tau, \tau \mid \mathbf{X}_0, \mathbf{k}_{1:T}, \mathbf{d}_{1:T}, \mathbf{r}_{1:T}, \ell). \tag{7}$$

Control adapters align $\mathbf{k}_t$ with the spatial and temporal positions of the DiT video tokens. This pathway directly provides the target body motion and viewpoint to the generator. The changing scene state reaches the DiT via a separate scene-dynamic pathway conditioned on $\mathbf{d}_t$ and $\mathbf{r}_t$.

Scene-Dynamic Residual Injection. Direct pose and camera conditioning specifies where the person should appear, while the scene dynamic determines how the surrounding scene responds to the motion. At selected DiT blocks, the current video features read the scene dynamic through a dedicated interaction branch. The branch uses $\mathbf{d}_t$ and $\mathbf{r}_t$ as scene conditions, applies a CrossAttn

operation followed by an FFN, and writes a gated residual back to the main DiT stream:

$$\Delta \mathbf{h}_t^b = g^b \mathcal{R}_\theta^b\big(\mathbf{h}_t^b, \mathbf{d}_t, \mathbf{r}_t\big)\,, \qquad \widetilde{\mathbf{h}}_t^b = \mathbf{h}_t^b + \Delta \mathbf{h}_t^b, \tag{8}$$

where g^b is a learned gate and $\mathcal{R}_\theta^b$ denotes the CrossAttn–FFN interaction branch. Adding this residual to the video tokens lets the model propagate the predicted scene response while retaining the appearance and layout of the initial frame. The camera controls determine where changes are rendered, while the scene dynamic determines which changes persist as the viewpoint moves. Integrating the predicted velocities yields a clean latent that the pretrained decoder maps to RGB frames.

3.4 Training Objectives and Strategy

We train HumanWorld with three complementary objectives. A flow-matching loss trains the Wan video DiT to predict velocities under the prescribed controls. Given the target video latent $\mathbf{z}_0 = \mathcal{E}(V)$, Gaussian noise $\boldsymbol{\epsilon}$, and flow time $\tau \in [0, 1]$, the noisy latent is $\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}$ and

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\left[\left\|\mathbf{v}_\theta(\mathbf{z}_\tau, \tau \mid I_0, c_0, a_{0:T}, c_{1:T}, \ell) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\right\|_2^2\right]. \tag{9}$$

A response loss $\mathcal{L}_{\text{response}}$ applies this objective, together with perceptual and semantic terms, within the human and interaction regions, which emphasizes the person and the objects affected by the action. A paired-camera loss $\mathcal{L}_{\text{multi}}$ renders the same world-space motion under two camera trajectories of a multi-camera recording, so that the observation changes while the underlying interaction stays shared. The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{response}}\mathcal{L}_{\text{response}} + \lambda_{\text{multi}}\mathcal{L}_{\text{multi}}. \tag{10}$$

The detailed formulations of all three objectives are provided in Section A.

Summary. HumanWorld separates what happens in the world from how it is observed. Human motion and camera trajectories are specified independently in a shared 3D world coordinate system. The world interaction module turns the prescribed motion into a latent scene dynamic that records how objects and human–scene relations evolve, while the camera trajectory only determines how this evolving state is read out. The video DiT renders the result through two pathways, with encoded controls fixing where the person appears and scene-dynamic residuals carrying the scene response. Trained with these objectives, the model follows the requested motion, propagates its consequences to nearby objects, and keeps these consequences consistent across viewpoints.

4 Experiments

4.1 Experimental Settings

Datasets. We train on a mixture of real-world, multi-view, and synthetic videos, including HumanVid (Wang et al., 2024), Ego-Exo4D (Grauman et al., 2024), BEHAVE (Bhatnagar et al., 2022), ABot World Explorer (Jiang et al., 2026), MultiCamVideo (Bai et al., 2025a), and WHAC-A-Mole (Yin et al., 2024). The real-world and multi-view interaction data provide full-body motion and object-interaction examples, while synthetic sequences provide calibrated camera trajectories for camera control and paired-view supervision. We cut the raw videos into fixed-length clips, filter them with temporal pose-quality checks, and attach the available camera or object trajectories.

Training Details. Built upon Wan2.2-TI2V-5B, our model uses zero-initialized control and interaction branches. The encoded pose and camera controls are added to the AdaLN-transformed video features at selected DiT blocks, while the scene dynamic and interaction relation condition the separate CrossAttn–FFN residual branch. The model is first trained for 10k steps on pose-annotated data, learning to follow full-body motion while keeping the camera-related parameters frozen. Camera-related parameters are then unfrozen and camera-conditioned data are introduced for a further 20k steps, allowing viewpoint control to be learned on top of the established pose control. We use AdamW with an effective batch size of 32 and a learning rate of 1×10^{-6}, with a warm-up over the first 1,000 steps and a constant rate thereafter. Training uses 8 NVIDIA A100 (80GB).

Table 1: **Camera control comparison on SpatialVID.** Best in **bold**, second best underlined.

Method	Subject Consistency ↑	Background Consistency ↑	Motion Smoothness ↑	Aesthetic Quality ↑	Imaging Quality ↑	ATE ↓	RPE ↓	RRE ↓
Astra (Zhu et al., 2026)	0.871	0.901	0.983	0.483	0.637	0.133	0.062	2.860
Fun-Camera (aigc apps, 2026)	0.956	0.935	0.991	**0.515**	0.669	0.028	0.016	0.222
Matrix-Game-3.5 (Qian et al., 2026)	0.953	0.936	**0.993**	0.484	0.655	0.027	0.016	**0.108**
Yume (Mao et al., 2025a)	0.881	0.931	0.963	0.490	0.608	0.057	0.023	2.066
Ours	**0.964**	**0.946**	0.992	0.514	**0.687**	**0.025**	**0.014**	0.384

Table 2: Quantitative comparison on **Fashion**. ↑ denotes higher is better, and ↓ denotes lower is better. Best and second-best results are in **bold** and underlined, respectively.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
DynamiCtrl (Haoyu et al., 2025)	17.76	0.857	0.133	275.0
MultiAnimate (Hu et al., 2026)	19.76	0.857	0.159	227.6
One-to-All-1.3B (Shi et al., 2025)	23.35	0.915	0.053	106.3
StableAnimator (Tu et al., 2025a)	**25.22**	**0.921**	0.055	67.1
Ours	24.38	0.915	**0.052**	**41.6**

4.2 Camera Control and World Exploration

Camera control turns a video generator into an explorable world model: the same scene should remain coherent while the observation trajectory changes. SpatialVID (Wang et al., 2025a) provides dynamic sequences with camera annotations. The test set has 300 sequences randomly sampled from the first 5,000 clips. Existing world models primarily use camera trajectories to explore a generated scene, whereas HumanWorld additionally conditions the scene on full-body human motion.

Perceptual quality is measured with five VBench (Huang et al., 2024) dimensions: Subject Consistency, Background Consistency, Motion Smoothness, Aesthetic Quality, and Imaging Quality. VGGT (Wang et al., 2025b) reconstructs a 3D camera trajectory from each generated video, and its agreement with the source trajectory is measured by Absolute Trajectory Error (**ATE**), Relative Pose Error (**RPE**), and Relative Rotation Error (**RRE**). Results are in Table 1. HumanWorld obtains the strongest subject consistency among the compared methods and competitive background, aesthetic, and imaging quality, while attaining the lowest ATE and RPE. These results show that the additional human-centric conditioning does not compromise the visual quality required for world exploration and provides effective control over the global and relative camera trajectory. The remaining gap in RRE indicates that precise rotational alignment remains more challenging.

4.3 Human Animation

Human animation isolates the ability to follow a full-body motion from the harder problem of changing the surrounding scene. We compare with human animation methods MultiAnimate (Hu et al., 2026), One-to-All (1.3B) (Shi et al., 2025), and StableAnimator (Tu et al., 2025a). Fashion (Zablotskaia et al., 2019) emphasizes detailed garment appearance and dynamic full-body motion. PSNR (Huynh-Thu & Ghanbari, 2008) and SSIM (Wang et al., 2004) measure pixel-level fidelity and structural preservation, LPIPS (Zhang et al., 2018) measures perceptual similarity, and Fréchet Video Distance (**FVD**) (Unterthiner et al., 2018) measures distributional and temporal quality.

On Fashion (Zablotskaia et al., 2019), HumanWorld achieves the best FVD and LPIPS, while staying competitive in reconstruction quality with the second-best PSNR and SSIM, as shown in Table 2. This confirms that the world-model formulation retains the motion following required by pose-guided animation, while leaving room for the interaction pathway to model scene responses.

4.4 Human–Scene Interaction

Following human pose alone does not guarantee a stable interaction with the surrounding scene. Existing video generators and human animation models can preserve the appearance and motion of the person, yet often lack an explicit representation of how contact, object displacement, and scene response should evolve over time. Human–object interaction videos therefore provide a direct test of whether a model can translate human motion into a coherent change of the nearby scene, rather

Table 3: Quantitative comparison on **BEHAVE**. ↑ denotes higher is better, and ↓ denotes lower is better. Best and second-best results are in **bold** and underlined, respectively.

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
Wan-Move (Chu et al., 2025)	17.72	0.727	0.326	424.9
SCAIL-2 (Yan et al., 2026)	20.75	0.780	0.238	305.8
Fun-Motion (aigc apps, 2026)	<u>23.48</u>	<u>0.856</u>	<u>0.166</u>	<u>246.8</u>
Ours	**25.82**	**0.906**	**0.114**	**232.2**

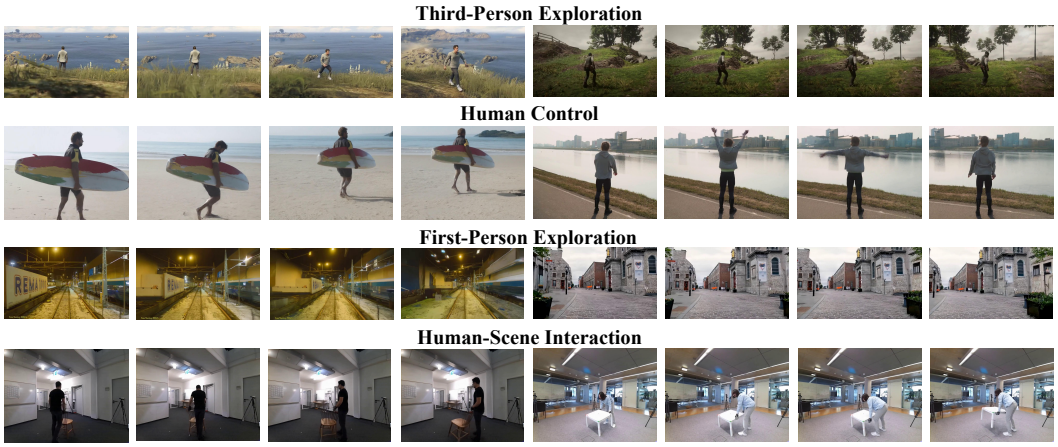


Figure 4: **Qualitative results across human-centric world modeling tasks.** Each row presents two generated sequences sampled at four time steps. HumanWorld supports third-person exploration, full-body human control, first-person exploration, and human–scene interaction while preserving scene structure, subject appearance, and action-induced object changes over time.

than merely placing a person in front of a static background. BEHAVE (Bhatnagar et al., 2022) contains full-body human motions and object interactions, providing a suitable setting for measuring the quality of generated interaction videos. The evaluation uses PSNR, SSIM, LPIPS, and FVD.

Table 3 shows that HumanWorld achieves the best results across all four metrics. The gains in PSNR and SSIM indicate closer frame-level and structural agreement with the target interaction, while the lower LPIPS and FVD reflect improved perceptual fidelity and temporal consistency. This improvement supports the role of the latent World Interaction dynamics in maintaining the relation between the moving person and the affected object throughout the video.

4.5 Qualitative Results

Figure 4 visualizes the capabilities of HumanWorld across four human-centric world modeling tasks. The third-person and first-person exploration results show that the model can reveal new regions along a specified camera trajectory while preserving the appearance and spatial structure of the scene. Under human control, the generated person follows diverse full-body motions while maintaining a consistent identity and coherent background. The human–scene interaction sequences further show that controlled actions produce corresponding changes in nearby objects, with contact and object appearance remaining stable as the interaction develops. Together, these examples illustrate how HumanWorld assigns camera trajectories to world observation and human motion to world evolution, supporting their composition within the same generative framework.

4.6 Analysis

World Interaction Module. Explicit pose and camera control specify the human motion and observation trajectory, but do not by themselves describe how the surrounding scene evolves in response. Without the latent World Interaction module, these controls reach the video generator without an intermediate representation of the action-conditioned scene state. The resulting generation can follow the prescribed body motion while relying mainly on the video prior to infer object displacement, contact, and other scene responses. The World Interaction module instead relates the

Table 4: **Effect of the latent World Interaction module on BEHAVE.** The ablated variant removes the World Interaction module while keeping the remaining settings unchanged.

Model	PSNR ↑	SSIM ↑	LPIPS ↓	FVD ↓
w/o World Interaction	23.22	0.896	0.133	280.6
Full Model	**25.82**	**0.906**	**0.114**	**232.2**

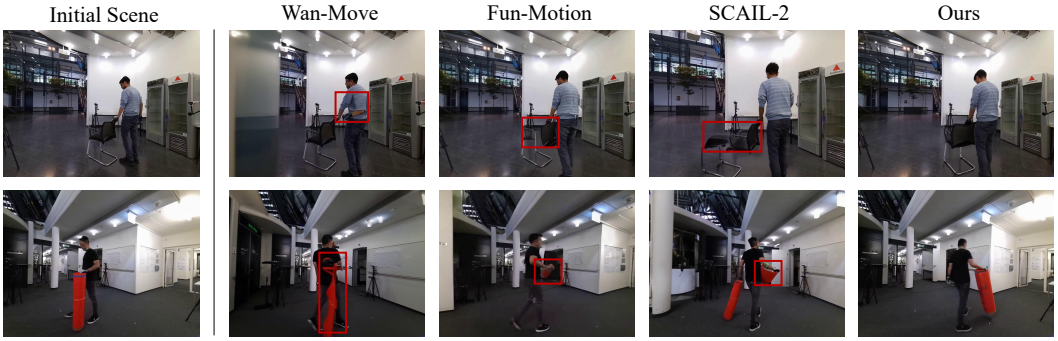


Figure 5: **Qualitative comparison of human–scene interaction.** Wan-Move, SCAIL-2, and Fun-Motion exhibit penetration or an inconsistent contact configuration when the person interacts with the object, while HumanWorld preserves the human–object relation throughout the motion.

controlled human state to the camera-aware scene representation and carries the inferred interaction dynamic into the video generator. As reported on BEHAVE in Table 4, this formulation produces higher PSNR and SSIM, showing that the generated appearance and spatial structure remain closer to the target interaction. The lower LPIPS reflects improved perceptual correspondence, while the lower FVD indicates that the gain extends from individual frames to temporal evolution. These improvements show that latent interaction dynamics translate the prescribed human motion into coherent changes in the surrounding scene, rather than merely preserving the motion of the person.

Contact-Consistent Scene Interaction. Figure 5 compares Wan-Move, SCAIL-2, Fun-Motion, and HumanWorld under the same interaction condition. As the prescribed motion develops, the baseline methods may exhibit hand–object penetration, object merging, or detached contact configurations. These artifacts suggest that the human motion is not consistently propagated into the state of the contacted object: the person can follow the intended trajectory while the contact geometry is independently reconstructed from the visual prior at each stage. HumanWorld instead maintains the contact relation and the geometry of the manipulated object throughout the sequence. This behavior is consistent with the role of the latent World Interaction dynamics, which carry the action-conditioned scene response from the human motion into the video generation process. The resulting coupling allows object displacement and contact configuration to evolve with the action, rather than treating the interaction as human motion over a largely independent background.

5 Conclusion and Discussions

In this paper, we have presented HumanWorld, a human-centric interactive world model that allows users to explore a scene through camera control and change it through the actions of a controlled human. Unlike camera-controlled world models that treat users as observers, HumanWorld specifies full-body human motion and camera trajectories independently in a shared 3D world coordinate system. A latent world interaction module predicts how the prescribed motion changes object states and human–scene relations, and a pretrained video diffusion model renders these changes from the requested viewpoint. Experiments on camera control, human animation, and human–scene interaction show that HumanWorld preserves camera controllability and motion fidelity while producing more consistent human–object interactions than existing methods. This formulation moves world models from passive exploration toward action-driven interaction.

Limitations. HumanWorld models interaction in a latent space without explicit object identities or physical constraints, so contact may be physically implausible in complex manipulations. Its interaction training data cover a limited range of objects and contact types, and generation is performed on fixed-length clips. Camera rotation is also aligned less precisely than translation.

AI USE STATEMENT

In this work, we have not used generative AI tools for any tasks with required disclosure, and all such tasks are not applicable to this work. We used generative AI tools only to polish the writing of parts of the manuscript. The authors reviewed all AI-assisted edits to ensure that they preserve the intended technical meaning and contain no unsupported claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 3 describes the model architecture and training objectives, Section 4.1 details the datasets, training schedule, and hyperparameters, and Section A gives the full objective formulations. We will release the complete source code and model weights.

REFERENCES

aigc apps. Videox-fun: A video generation pipeline for diffusion transformer, 2026. URL https://github.com/aigc-apps/VideoX-Fun.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. URL https://arxiv.org/abs/2506.09985.

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, and Di Zhang. ReCamMaster: Camera-controlled generative rendering from a single video. *arXiv preprint arXiv:2503.11647*, 2025a.

Yutong Bai, Danny Tran, Amir Bar, Yann LeCun, Trevor Darrell, and Jitendra Malik. Whole-body conditioned egocentric video prediction. *arXiv preprint arXiv:2506.21552*, 2025b. URL https://arxiv.org/abs/2506.21552.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15935–15946, 2022.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable Video Diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. URL https://arxiv.org/abs/2311.15127.

Ruihang Chu, Yefei He, Zhekai Chen, Shiwei Zhang, Xiaogang Xu, Bin Xia, Dingdong Wang, Hongwei Yi, Xihui Liu, Hengshuang Zhao, et al. Wan-move: Motion-controllable video generation via latent trajectory guidance. *arXiv preprint arXiv:2512.08765*, 2025.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J. Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina González, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey

Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbeláez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. 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 (CVPR)*, pp. 19383–19400, 2024.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 2025. doi: 10.1038/s41586-025-08744-2. URL https://www.nature.com/articles/s41586-025-08744-2.

Zhao Haoyu, Qi Zhongang, Wang Cong, Zheng Qingping, Lu Guansong, Chen Fei, Xu Hang, and Wu Zuxuan. Dynamictrl: Rethinking the basic structure and the role of text for high-quality human image animation. *arXiv:2503.21246*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-Game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025. URL https://arxiv.org/abs/2508.13009.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a. URL https://arxiv.org/abs/2210.02303.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022b. URL https://arxiv.org/abs/2204.03458.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate Anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2311.17117.

Li Hu, Guangyuan Wang, Zhen Shen, Xin Gao, Dechao Meng, Lian Zhuo, Peng Zhang, Bang Zhang, and Liefeng Bo. Animate Anyone 2: High-fidelity character image animation with environment affordance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2502.06145.

Yingcheng Hu, Haowen Gong, Chuanguang Yang, Zhulin An, Yongjun Xu, and Songhua Liu. Multianimate: Pose-guided image animation made extensible. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9306–9316, 2026.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics Letters*, 44:800–801, 2008. doi: 10.1049/el:20080522. URL https://digital-library.theiet.org/doi/abs/10.1049/el%3A20080522.

Chaonan Ji, Jinwei Qi, Peng Zhang, and Bang Zhang. Real-time human-centric world modeling for upper-body human-object interaction. *arXiv preprint arXiv:2607.23517*, 2026. URL https://arxiv.org/abs/2607.23517.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, et al. Abot-world-0: Infinite interactive world rollout on a single desktop gpu. *arXiv preprint arXiv:2607.19191*, 2026.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. URL https://arxiv.org/abs/2412.03603.

Gaojie Lin, Jianwen Jiang, Jiaqi Yang, Zerong Zheng, and Chao Liang. OmniHuman-1: Rethinking the scaling-up of one-stage conditioned human animation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://arxiv.org/abs/2502.01061.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, Tong He, Jiangmiao Pang, Yu Qiao, and Kaipeng Zhang. Yume-1.5: A text-controlled interactive world generation model, 2025a. URL https://arxiv.org/abs/2512.22096.

Xiaofeng Mao, Shaoheng Lin, Zhen Li, Chuanhao Li, Wenshuo Peng, Tong He, Jiangmiao Pang, Mingmin Chi, Yu Qiao, and Kaipeng Zhang. Yume: An interactive world generation model. *arXiv preprint arXiv:2507.17744*, 2025b. URL https://arxiv.org/abs/2507.17744.

Maxime Oquab, Timothée Darort, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Jack Parker-Holder and Shlomi Fruchter. Genie 3: A new frontier for world models. Google DeepMind, 2025. URL https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/.

Jack Parker-Holder, Philip Ball, Jake Bruce, Vibhavari Dasagi, Kristian Holsheimer, Christos Kaplanis, Alexandre Moufarek, Guy Scully, Jeremy Shar, Jimmy Shi, Stephen Spencer, Jessica Yung, Michael Dennis, Sultan Kenjeyev, Shangbang Long, Vlad Mnih, Harris Chan, Maxime Gazeau, Bonnie Li, Fabio Pardo, Luyu Wang, Lei Zhang, Frederic Besse, Tim Harley, Anna Mitenkova, Jane Wang, Jeff Clune, Demis Hassabis, Raia Hadsell, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 2: A large-scale foundation world model. Google DeepMind, 2024. URL https://deepmind.google/blog/genie-2-a-large-scale-foundation-world-model/.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory, 2026. URL https://arxiv.org/abs/2608.29910.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026. URL https://arxiv.org/abs/2601.20540.

Shijun Shi, Jing Xu, Zhihang Li, Chunli Peng, Xiaoda Yang, Lijing Lu, Kai Hu, and Jiangning Zhang. One-to-all animation: Alignment-free character animation and image pose transfer. *arXiv preprint arXiv:2511.22940*, 2025.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. WorldPlay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025. URL https://arxiv.org/abs/2512.14614.

Shuyuan Tu, Zhen Xing, Xintong Han, Zhi-Qi Cheng, Qi Dai, Chong Luo, and Zuxuan Wu. Stableanimator: High-quality identity-preserving human image animation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 21096–21106, 2025a.

Yuanpeng Tu, Hao Luo, Xi Chen, Xiang Bai, Fan Wang, and Hengshuang Zhao. PlayerOne: Egocentric world simulator. *arXiv preprint arXiv:2506.09995*, 2025b. URL https://arxiv.org/abs/2506.09995.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01784*, 2018.

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.

Haoyuan Wang, Yabo Chen, Haibin Huang, Chi Zhang, and Xuelong Li. Directing the world: Fast autoregressive video generation with compositional human-camera control. *arXiv preprint arXiv:2606.27964*, 2026. URL https://arxiv.org/abs/2606.27964.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. Spatialvid: A large-scale video dataset with spatial annotations, 2025a. URL https://arxiv.org/abs/2509.09676.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Zhenzhi Wang, Yixuan Li, Yanhong Zeng, Youqing Fang, Yuwei Guo, Wenran Liu, Jing Tan, Kai Chen, Tianfan Xue, Bo Dai, and Dahua Lin. HumanVid: Demystifying training data for camera-controllable human image animation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

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.

Linxi Xie, Lisong C. Sun, Ashley Neall, Tong Wu, Shengqu Cai, and Gordon Wetzstein. Generated reality: Human-centric world simulation using interactive video generation with hand and camera control. *arXiv preprint arXiv:2602.18422*, 2026. URL https://arxiv.org/abs/2602.18422.

Jiacong Xu, Hanwen Jiang, Zhixin Shu, Kalyan Sunkavalli, Vishal M. Patel, and Yiqun Mei. Wonder: Video world model done better. *arXiv preprint arXiv:2607.26037*, 2026. URL https://arxiv.org/abs/2607.26037.

Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023. URL https://arxiv.org/abs/2311.16498.

Wenhao Yan, Fengjia Guo, Zhuoyi Yang, and Jie Tang. Scail-2: Unifying controlled character animation with end-to-end in-context conditioning. *arXiv preprint arXiv:2606.10804*, 2026.

Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 4210–4220, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihan Wang, Yean Cheng, Ting Liu, Bin Xu, 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.

Wanqi Yin, Zhongang Cai, Ruisi Wang, Fanzhou Wang, Chen Wei, Haiyi Mei, Weiye Xiao, Zhitao Yang, Qingping Sun, Atsushi Yamashita, Lei Yang, and Ziwei Liu. WHAC: World-grounded humans and cameras. In *European Conference on Computer Vision (ECCV)*, pp. 20–37, 2024.

Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation, 2019. URL https://arxiv.org/abs/1910.09139.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. MimicMotion: High-quality human motion video generation with confidence-aware pose guidance. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2406.19680.

Shenhao Zhu et al. Champ: Controllable and consistent human image animation with 3d parametric guidance. *arXiv preprint arXiv:2403.14781*, 2024. URL https://arxiv.org/abs/2403.14781.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8UZpmrxoLG.

A Training Objectives

We train HumanWorld with three complementary objectives that match the roles of its generative and interaction pathways. Flow matching learns the target video distribution under the prescribed controls. Response supervision focuses that learning on the controlled person and the parts of the scene affected by the action. Paired camera views then require the same interaction to be rendered from different observation trajectories. These objectives use video, regional, perceptual, semantic, and multi-camera supervision without requiring a separate target for the latent state $\mathbf{s}_t$.

Video Generation. Let $\mathbf{z}_0 = \mathcal{E}(V)$ be the target video latent, $\boldsymbol{\epsilon}$ Gaussian noise, and $\tau \in [0, 1]$ the flow interpolation time. We train the velocity prediction of the Wan video DiT with

$$\begin{aligned} \mathbf{z}_\tau &= (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}, \\ \mathcal{L}_{\text{gen}} &= \mathbb{E}_{\tau,\boldsymbol{\epsilon}}\left[\left\|\mathbf{v}_\theta(\mathbf{z}_\tau, \tau \mid I_0, c_0, a_{0:T}, c_{1:T}, \ell) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\right\|_2^2\right]. \end{aligned} \tag{11}$$

The image, pose, and camera inputs produce the conditioning features $\mathbf{X}_0$, $\mathbf{k}_{1:T}$, $\mathbf{d}_{1:T}$, and $\mathbf{r}_{1:T}$ defined above. The scene description ℓ is used when available. The full video loss trains the model to preserve the observed appearance while following the human and camera trajectories.

Human and Scene Response. Full frame generation alone may underweight the person and nearby objects that express the controlled action. Let Ω_h and Ω_i denote the human and interaction regions. We apply the generation loss within their union and retain the perceptual and semantic terms of the response objective:

$$\mathcal{L}_{\text{response}} = \mathcal{L}_{\text{gen}}^{\Omega_h \cup \Omega_i} + \lambda_{\text{perc}} d_\phi\left(\hat{V}^{\Omega_h \cup \Omega_i}, V^{\Omega_h \cup \Omega_i}\right) + \lambda_{\text{sem}} d_{\text{sem}}(\mathbf{s}_{1:T}, \ell), \tag{12}$$

where $\hat{V}$ is the generated video and d_ϕ measures the distance between perceptual features of the generated and target regions. The semantic distance d_{sem} aligns the action-conditioned scene representation with the available scene description. The semantic term is applied when that description is available. The regional generation term emphasizes the visible motion of the person and the scene response, while the perceptual term helps retain their appearance.

Paired Camera Views. Multi-camera recordings show the same action and scene event under different views. For views u and v, let V^w be the target video under camera $c_{0:T}^w$, and let I_0^u be the initial frame of the reference view. We express the target camera trajectory relative to the reference camera as $c_t^{w|u_0} = (c_0^u)^{-1} c_t^w$. We train both target views with the paired generation objective:

$$\mathcal{L}_{\text{multi}} = \sum_{w \in \{u,v\}} \mathcal{L}_{\text{gen}}\left(V^w \mid I_0^u, c_0^u, a_{0:T}, c_{1:T}^{w|u_0}, \ell\right). \tag{13}$$

The same world-space human motion $a_{0:T}$ drives both views. Its projected pose and the view-aligned scene features are recomputed for each target camera. Thus the paired videos supervise observation changes while keeping the underlying action and scene event shared.

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda_{\text{response}}\mathcal{L}_{\text{response}} + \lambda_{\text{multi}}\mathcal{L}_{\text{multi}}. \tag{14}$$

The three terms respectively train video synthesis, emphasize the human and scene response, and align different views of the same interaction. Together, they train controlled human motion, its effect on the scene, and observation from the requested camera.

B Dataset Details

Table 5 summarizes the training data. The six sources play complementary roles in learning motion following, viewpoint control, and human–scene interaction. BEHAVE (Bhatnagar et al., 2022) records indoor human–object interactions with object poses and masks in addition to human pose, and serves as the primary source of interaction supervision. HumanVid (Wang et al., 2024) contributes in-the-wild human videos with estimated cameras, which increases the diversity of subject appearance and joint human and camera motion. Ego-Exo4D (Grauman et al., 2024) covers a wide range of skilled activities and provides large-scale full-body motion. MultiCamVideo (Bai et al., 2025a) is

Table 5: **Overview of training datasets.** Summary of the scene domains, annotations, and sample counts across all data sources, totaling 204.7 hours of video.

Dataset	Scenes	Annotations	Size
BEHAVE (Bhatnagar et al., 2022)	Indoor HOI	2D & object pose, mask	8,617
HumanVid (Wang et al., 2024)	In-the-wild	2D pose, est. camera	7,290
Ego-Exo4D (Grauman et al., 2024)	Skilled activity	2D pose	71.2K
MultiCamVideo (Bai et al., 2025a)	Synthetic	2D pose, native camera	74.2K
WHAC (Yin et al., 2024)	Cinematic dance	2D pose, native camera	4,360
ABot World Explorer (Jiang et al., 2026)	AAA gameplay	2D pose, COLMAP camera	18.6K
Total	–	–	$\sim$ 184K

Table 6: **Implementation details of HumanWorld.** Entries in red are to be completed.

Item	Setting
Backbone	Wan2.2-TI2V-5B
Output resolution and length	Processed resolution, 81 frames
Learnable queries $\lvert\mathbf{Q}^h\rvert$, $\lvert\mathbf{Q}^s\rvert$	6, 4
Interaction module width	512
Loss weights $\lambda_{\text{response}}$, λ_{multi}	0.3, 1
Loss weights λ_{perc}, λ_{sem}	1.2, 0.8
Stage 1	10k steps, pose-annotated data, camera parameters frozen
Stage 2	20k steps, camera-conditioned data added
Optimizer	AdamW
Learning rate	1×10^{-6}, 1,000-step warm-up, then constant
Effective batch size	32
Hardware	8 NVIDIA A100 (80GB)
Sampling steps	50
Classifier-free guidance scale	1.0

rendered with native cameras and multiple synchronized views of the same event, which supplies the paired views used by $\mathcal{L}_{\text{multi}}$. WHAC-A-Mole (Yin et al., 2024) adds large-amplitude dance motion under cinematic camera movement. ABot World Explorer (Jiang et al., 2026) provides gameplay sequences with COLMAP cameras and exploration-style trajectories through diverse environments.

Preprocessing. All videos are cut into clips of 81 RGB frames at the processed dataset resolution and 20 fps. Human poses are extracted with DWPose(Yang et al., 2023), and clips with missing, truncated, or temporally unstable poses are removed by the pose-quality checks described in Section 4.1. For sources with camera annotations, each trajectory is expressed relative to the first frame as $\bar{c}_t = c_0^{-1} c_t$, which matches the camera input of the model.

Data usage across stages. The first training stage uses all pose-annotated clips with the camera-related parameters frozen. The second stage adds the clips with native, estimated, or COLMAP cameras, and paired views from MultiCamVideo enable $\mathcal{L}_{\text{multi}}$.

C Implementation Details

Table 6 lists the main settings. HumanWorld is built on Wan2.2-TI2V-5B. The control and interaction branches are zero-initialized, so training starts from the behavior of the pretrained generator.

Control encoding. The pose encoder $\mathcal{C}_\theta^{\text{pose}}$ takes the world-space pose projected into the target view, $\Pi_{c_t}(a_t)$, rendered as a 134-keypoint DWPose RGB sketch. The camera encoder $\mathcal{C}_\theta^{\text{cam}}$ takes the relative camera $\bar{c}_t$ represented as a six-channel C2W Plücker ray map. The control encoder $\mathcal{K}_\theta$ fuses the two encodings, and control adapters add the result to the AdaLN-transformed video features at the blocks listed in Tab. 6.

World interaction module. The Human-Aware, Action–Scene, and Interaction blocks use the relational-attention layers listed in Tab. 6, with hidden width 512. The interaction dynamic $\mathbf{r}_t$ and the scene dynamic $\mathbf{d}_t$ are passed to the CrossAttn–FFN interaction branch.

Losses. The perceptual distance d_ϕ is computed from frozen ArcFace IR50 (Deng et al., 2019) and DINOv2-Base (Oquab et al., 2023) features. The semantic distance d_{sem} uses text embeddings produced by the frozen UMT5-XXL

D Evaluation Protocol

Camera control. The SpatialVID test set contains 300 sequences randomly sampled from the first 5,000 clips. Every method receives the same initial frame and source camera trajectory, and 81 frames are generated per sequence. VGGT (Wang et al., 2025b) estimates camera poses from the generated frames, which are aligned to the source trajectory with a Sim(3) transformation estimated by Umeyama alignment before computing the errors. ATE is the root mean square translation error of the aligned trajectory, RPE is the translation error between consecutive relative poses, and RRE is the rotation error between consecutive relative poses in degrees. The five VBench (Huang et al., 2024) dimensions are computed with the official implementation.

Human animation and interaction. On the Fashion (Zablotskaia et al., 2019) and BEHAVE (Bhatnagar et al., 2022) datasets, we evaluate on 100 test videos from each. Each method receives the first frame and the ground-truth pose sequence. PSNR, SSIM, and LPIPS are averaged over all generated frames, and FVD is computed over the set of generated videos.

Baselines. We follow each baseline's official implementation and recommended inference configuration as closely as possible. Metric-sensitive inputs—including the text prompt, spatial resolution, and number of frames—are handled under a shared evaluation protocol.