AnchorWorld: Embodied Egocentric World Simulation with View-based Evolution Customization

Yu Li[1]* **Menghan Xia**[2]† **Gongye Liu**[4] **Xintao Wang**[3] **Conglang Zhang**[5]
Lei Ke[1] **Yuxuan Lin**[1] **Ruihang Chu**[1] **Pengfei Wan**[3] **Kun Gai**[3] **Yujiu Yang**[1]†

[1]Tsinghua University [2]HUST [3]Kling Team, Kuaishou Technology [4]HKUST [5]WHU

https://yuli0103.github.io/AdaViewPlanner/

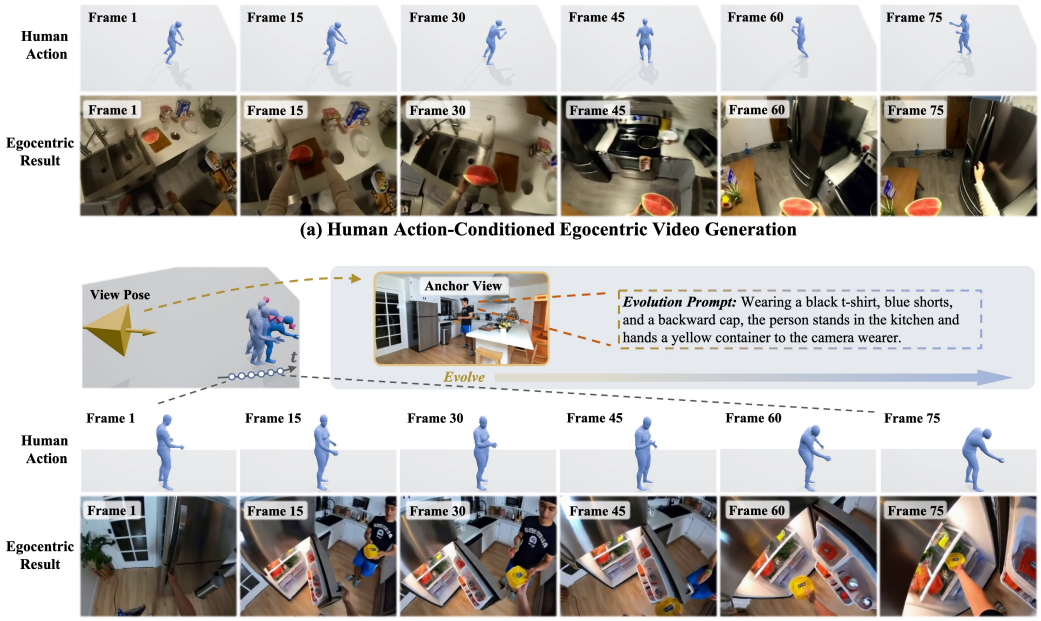


Figure 1: Showcasing AnchorWorld. (a) AnchorWorld synthesizes egocentric videos conditioned on human action and initial ego-view frame. (b) It further enables world customization with conditional anchor views, which provide local appearance, 3D pose, and evolution prompts for scene evolution.

Abstract

Despite being a pivotal frontier, interactive world modeling remains underexplored in terms of the versatile controllability required by practical scenarios. To bridge this gap, we present AnchorWorld, a framework that advances egocentric simulation through enhanced interaction integrity and a flexible mechanism for world customization. First, we utilize 3D human motion as the primary interaction modality. To complement the out-of-view or truncated body parts in egocentric views, we introduce an auxiliary training supervision that incorporates exogenous viewpoints decoupled from the agent's first-person sensorium. It allows the model to observe the agent's full-body positioning relative to the environment, facilitating a more robust spatial grounding of human-world interactions. Furthermore, we propose a simple yet effective mechanism for customizing self-evolving worlds.

*This work was conducted during the author's internship at Kling Team, Kuaishou Technology.
†Corresponding authors.

Preprint.

This is achieved by defining anchor views within a unified world coordinate system, coupled with textual descriptions dictating the dynamic evolution of local scenes. Experimental results show that AnchorWorld significantly outperforms state-of-the-art baselines, while ablation studies validate the effectiveness of our key designs. Notably, our customization scheme exhibits promising spatio-temporal geometric consistency and adheres strictly to the prescribed evolutionary dynamics.

1 Introduction

Interactive world models aim to simulate dynamic visual environments that respond to user intervention. For first-person applications such as virtual reality and embodied AI [5, 9, 14], this response is not merely a matter of predicting visually plausible continuations. The simulator must account for how the user moves and acts: head motion determines where the camera looks, body motion drives navigation, and coordinated actions shape how the user interacts with nearby objects. Meanwhile, the simulated world should not be treated as an unconstrained visual continuation: it should contain local states that can be specified, preserved, and evolved as the user moves through the environment. Together, these requirements call for an egocentric world simulator with two complementary forms of control: *embodied action control* and *localized world-state customization*.

Existing interactive world models only partially satisfy these requirements. Many approaches [6, 43, 44, 53] rely on simplified control signals such as keyboard inputs, camera trajectories, or text prompts, which are convenient for navigation but do not reflect how humans act from a first-person perspective. Recent egocentric methods move toward more natural control by incorporating hand actions [48, 52] or full-body motion [4, 45]. However, learning such control from egocentric videos remains challenging. The motion condition describes the body in 3D, while most of the body is absent from the egocentric frame to be predicted. Therefore, the model observes the visual consequences of body motion only indirectly, making motion supervision sparse and weakly aligned. A second challenge lies in how the "world" itself is defined. Existing methods [45, 57] typically determine the environment implicitly through an initial frame, a global prompt, or historical context; newly observed regions are therefore weakly constrained. This makes it difficult to specify what should exist at particular 3D locations or how local scene states should evolve over time.

The two limitations above motivate **AnchorWorld**, a framework for world-customizable embodied egocentric simulation. AnchorWorld provides two complementary forms of control: human body motion for egocentric navigation and interaction, and pose-associated anchor views for explicit world customization. For egocentric action control, the supervision missing from first-person videos is precisely what third-person videos provide, since the body and its interaction with the scene are visible from outside. We thus pair 3D human motion with camera viewpoint and formulate action conditioning in a projection-based manner, where the camera viewpoint can correspond to either an external observation view or the head-mounted view, enabling hybrid-view training. This *hybrid-view human action control* lets the model learn how full-body motion shapes first-person visual observations. For world customization, we represent local world states with *pose-associated anchor views*. Each anchor view consists of an RGB image specifying local visual appearance, a 3D pose that grounds the anchor, and an evolution prompt that describes its dynamic changes. These anchors allow users to specify local states at chosen 3D locations, preserve them across changing viewpoints, and guide their evolution, including in regions initially out of sight.

We train AnchorWorld with a progressive strategy that introduces hybrid-view human action control, anchor-view scene consistency, and dynamic evolution in successive stages so each component builds on a stable base. Across egocentric, synthetic UE, and captured real-world scenarios, AnchorWorld improves over adapted baselines on action accuracy, scene consistency, and dynamic evolution. The results further reveal remarkable generalization to out-of-distribution scenarios, especially under large viewpoint changes and limited overlap between the initial ego-view and anchor views. Additional analyses show two key capabilities for localized world customization: out-of-sight scene evolution and pose-consistent anchoring under spatial transformations. Our contributions are summarized:

- We formulate *world-customizable embodied egocentric simulation*, a task that enables human-motion-driven exploration and interaction within customizable, self-evolving worlds.
- We propose *AnchorWorld*, a unified framework that combines embodied egocentric action control with pose-associated anchor-view customization.

- We validate AnchorWorld through extensive experiments, demonstrating accurate egocentric human action control, strong spatial awareness, and controllable scene evolution.

2 Related Work

Interactive World Models. The core pursuit of interactive world models is to synthesize visual environments conditioned on user input actions. A large body of early research adopts keyboard and mouse operations to control viewpoints and navigate simulated worlds [6, 17, 42, 44, 50, 54, 63]. Concurrently, another line of work employs text prompts as interaction signals, enabling users to trigger specific world events and drive environmental transitions [1, 7, 31, 39, 43, 51, 53]. To support more fine-grained and embodied interactions, recent studies introduce hand poses as control signals [13, 16, 26, 48, 52, 60]. However, they are often limited to egocentric scenarios with restricted camera motion. DWM [24] performs interaction within static 3D scenes and achieves embodied simulation conditioned on rendered first-person videos and rendered hand meshes. PlayerOne [45] uses full-body human motion to build egocentric world simulators. It introduces a part-disentangled motion injection scheme, allowing the model to perceive the roles of different body parts. Similarly, PEVA [4] adopts human motion as the action condition and generates videos without text input, encouraging intention inference from first-person videos and motion cues.

Scene-Consistent Video Generation. ReCamMaster [2] tackles novel camera trajectory synthesis by enforcing scene consistency through source-video conditioning via in-context learning. It further constructs paired training data with different camera trajectories using synthetic Unreal Engine data. CineScene [19] represents a scene with a dense sequence of images captured at regular angular intervals, and leverages implicit 3D features [47] to build scene understanding for camera-controlled cinematic video generation. SWM [37] grounds its world model in real-world urban environments by retrieving nearby street-view images during navigation, and uses geometric and semantic references to improve spatial realism. Context-as-Memory [57] maintains scene consistency in long video navigation by retrieving field-of-view-relevant historical frames and injecting both scene and viewpoint cues into generation. Additionally, another line of work incorporates explicit 3D representations to improve view consistency across generated frames [10, 18, 20, 32, 36, 56, 58, 59]. These methods typically reconstruct or maintain intermediate 3D scene representations, such as depth maps or point clouds, and use them to guide novel-view or trajectory-conditioned video generation.

3 Method

Given a sequence of human actions and a customizable world specification, our goal is to synthesize an egocentric video that reflects how a user navigates and interacts within the defined environment. To this end, AnchorWorld takes two types of control signals as input: embodied human motion for action control, and pose-associated anchor views for world customization. We instantiate AnchorWorld with Wan [46], a flow-matching-based [29] DiT [34] video generation model, and condition its video synthesis on the action and anchor-view signals. The human motion is represented as a sequence of body actions derived from the SMPL-X parametric body model [33], denoted as $M \in \mathbb{R}^{f \times k \times 6}$, where f is the number of frames and k is the number of joints. Each joint state consists of its 3D position and 3D axis-angle rotation vector. The customizable world is defined by an initial egocentric view I_0 and a set of localized anchor views $\mathcal{S} = \{(\boldsymbol{I}_i, \boldsymbol{c}_i, \boldsymbol{t}_i)\}_{i=1}^{n}$. Each anchor view contains an RGB image $\boldsymbol{I}_i$, a 6-DoF viewpoint pose $\boldsymbol{c}_i = [\boldsymbol{R}_i \mid \boldsymbol{p}_i] \in \mathbb{R}^{3\times4}$, and an evolution prompt $\boldsymbol{t}_i$ that describes the temporal change of local scene states. Figure 2 provides an overview of the proposed framework. We detail each component of our approach in the following subsections.

3.1 Hybrid-View Human Action Control

Enhanced Egocentric Action Control via hybrid views. Human action contains rich spatial and interaction cues: the root trajectory determines global navigation, the limbs indicate potential interactions with the surrounding scene, and the head motion induces the egocentric viewpoint. However, in first-person videos, most body parts are often outside the camera field of view, making direct supervision of full-body action control sparse and incomplete. To overcome this limitation, we introduce third-person view (TPV) videos as auxiliary training data, where the full human body and its interactions with the surrounding scene are explicitly visible. These videos provide rich interaction

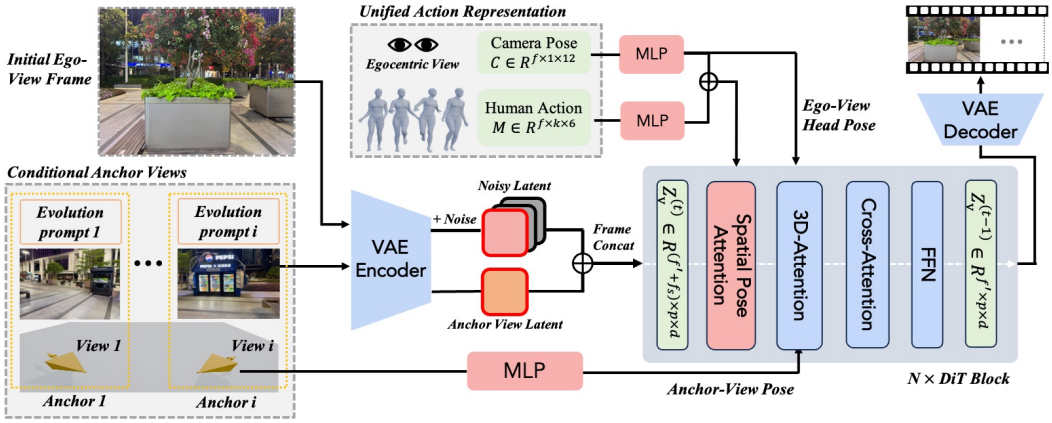


Figure 2: AnchorWorld synthesizes egocentric videos conditioned on embodied human actions and anchor views. For action control, full-body motion and ego-view pose are concatenated as a unified action representation and injected via spatial pose attention. For world customization, each anchor view includes an RGB image, a 3D pose, and an evolution prompt, enabling spatially grounded and temporally evolvable world simulation. Evolution prompts are incorporated via cross-attention layers.

context and complete motion supervision, helping the model learn stronger spatial grounding between human motion and scene responses. To support joint training on both TPV and first-person view (FPV) data within a unified framework, we formulate action conditioning in a projection-based manner. Specifically, we represent the action condition by combining the full-body motion sequence with the camera trajectory, allowing the model to project 3D human motion into 2D visual observations under arbitrary viewpoints. We first pre-train the model on large-scale and diverse TPV videos, where the camera parameters correspond to the external observation viewpoint, enabling the model to acquire projection knowledge and human-scene interaction priors. Then, we adapt the model to egocentric simulation by aligning the camera parameters with the human head perspective in FPV data. This design enables more accurate human-action control and stronger spatial pose awareness.

Spatial Pose Attention. Inspired by prior work [12, 27], we inject the pose conditions through a spatial pose attention mechanism. Specifically, a motion encoder first projects the input motion sequence $M \in \mathbb{R}^{f\times k\times 6}$ into a latent embedding $\boldsymbol{z}_m \in \mathbb{R}^{f'\times k\times d}$, where d is the model's hidden dimension. To ensure temporal alignment with the VAE-encoded [25] video latents, we employ temporal downsampling to match the temporal resolution f'. Analogously, a camera encoder processes the camera pose sequence $C \in \mathbb{R}^{f\times 3\times 4}$ into $\boldsymbol{z}_c \in \mathbb{R}^{f'\times 1\times d}$, where the camera pose can represent either a third-person observation viewpoint or the first-person head viewpoint.

To exploit the inherent frame-wise correspondence between motion and video tokens, we concatenate the video tokens $\boldsymbol{z}_v^{(t)}$ with the human motion tokens $\boldsymbol{z}_m$ and camera pose tokens $\boldsymbol{z}_c$ along the spatial dimension. This unified sequence is then processed by the spatial self-attention block:

$$\begin{aligned} \boldsymbol{T} &= [\boldsymbol{z}_v^{(t)}; \boldsymbol{z}_m; \boldsymbol{z}_c] \in \mathbb{R}^{f'\times(h\cdot w+k+1)\times d}, \\ \boldsymbol{z}_v^{(t)} &= \boldsymbol{z}_v^{(t)} + \text{Truncate}\left(\text{Attn}(\boldsymbol{W}_Q\cdot\boldsymbol{T}, \boldsymbol{W}_K\cdot\boldsymbol{T}, \boldsymbol{W}_V\cdot\boldsymbol{T})\right) \end{aligned} \tag{1}$$

The Truncate operator discards the auxiliary pose tokens, retaining only the updated video features.

3.2 Evolvable Anchor-View Customization

To enable evolvable world customization, we represent the environment with a set of anchor views. Each anchor view provides three types of localized world priors: an RGB image for visual appearance, a 3D pose for spatial grounding, and an evolution prompt for temporal state evolution.

In-Context Anchor-View Priors. To incorporate anchor-view image priors while preserving the generative capability of the pre-trained video model, we adopt an in-context conditioning strategy [19,

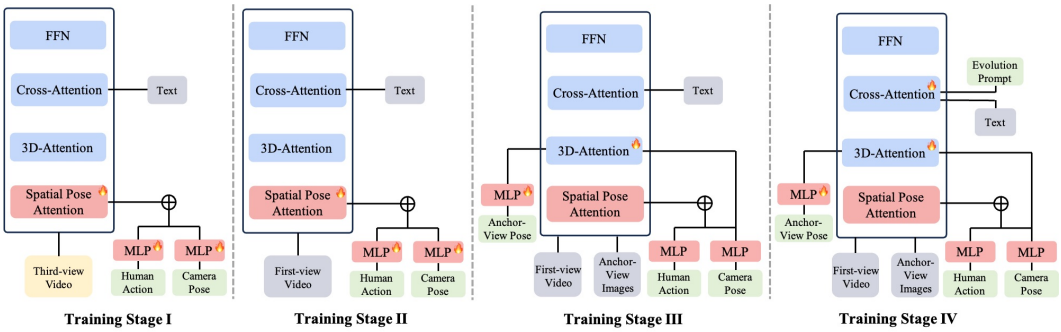


Figure 3: Progressive multi-stage training strategy. Stage I: TPV action training; Stage II: FPV action training; Stage III: static anchor-view customization; Stage IV: dynamic anchor-view evolution.

23, 55]. Specifically, the images of anchor views are encoded into latent tokens $\boldsymbol{z}_s \in \mathbb{R}^{f_s \times h \cdot w \times d}$, which are concatenated with the video latent tokens $\boldsymbol{z}_v^{(t)} \in \mathbb{R}^{f' \times h \cdot w \times d}$ along the frame dimension:

$$\mathcal{T}_{total} = [\boldsymbol{z}_v^{(t)}; \boldsymbol{z}_s] \in \mathbb{R}^{(f'+f_s) \times h \cdot w \times d}. \tag{2}$$

This design enables anchor views to guide world synthesis in-context, without requiring architectural modifications to the base model. We further employ 3D RoPE [41] to differentiate anchor views by assigning them distinct frame-axis positions in the positional embedding space.

View Pose Injection. Since each view corresponds to a specific 3D location in the world, its spatial pose is essential for grounding the customized content. We therefore inject pose information for both generated video frames and anchor views. The camera poses are encoded into embeddings $\boldsymbol{z}_{pose} \in \mathbb{R}^{(f'+f_s) \times 1 \times d}$ and spatially broadcast to match the latent resolution, yielding $\boldsymbol{z}_{pose} \in \mathbb{R}^{(f'+f_s) \times h \cdot w \times d}$. Before the self-attention layers, the pose embeddings are added to the visual tokens:

$$\mathcal{T}_{total} = \mathcal{T}_{total} + \boldsymbol{z}_{pose}. \tag{3}$$

By coupling visual tokens with spatial poses, the model can distinguish anchor views located at different positions and associate the generated egocentric trajectory with the correct local constraints.

Text-Driven Anchor-View Evolution. To enable dynamic world customization, each anchor view is paired with a localized evolution description $\boldsymbol{t}_i$ that specifies its temporal scene changes. We inject these descriptions through cross-attention, leveraging the semantic priors of the pre-trained video model. To preserve the locality of dynamic instructions, we restrict the interaction between text prompts and visual tokens using an attention mask. For a text prompt $\boldsymbol{t}_j$, its text keys are visible only to the generated video tokens and the corresponding anchor-view tokens $\boldsymbol{z}_s^{(j)}$:

$$\mathcal{M}(q, k_j) = \begin{cases} 0, & \text{if } q \in \boldsymbol{z}_v \text{ or } q \in \boldsymbol{z}_s^{(j)}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{4}$$

This masked cross-attention enables anchor-specific text control, allowing local scene states to evolve over time while reducing interference across different anchor views.

3.3 Progressive Multi-Stage Training Strategy

To progressively equip the model with egocentric human action control and evolvable anchor-view customization, we adopt a multi-stage training strategy, as illustrated in Figure 3. **Stage I & II: Hybrid-View Action Control Training.** We train the model to learn action-conditioned generation from hybrid viewpoints, where TPV videos provide complete full-body motion supervision. In Stage I, the model is trained on large-scale third-person videos, where the camera parameters represent external observation viewpoints. In Stage II, we then adapt the model to first-person videos by aligning the camera trajectory with the head pose of the character. **Stage III & IV: Evolvable Anchor-View Customization Training.** After establishing action controllability, we train the model

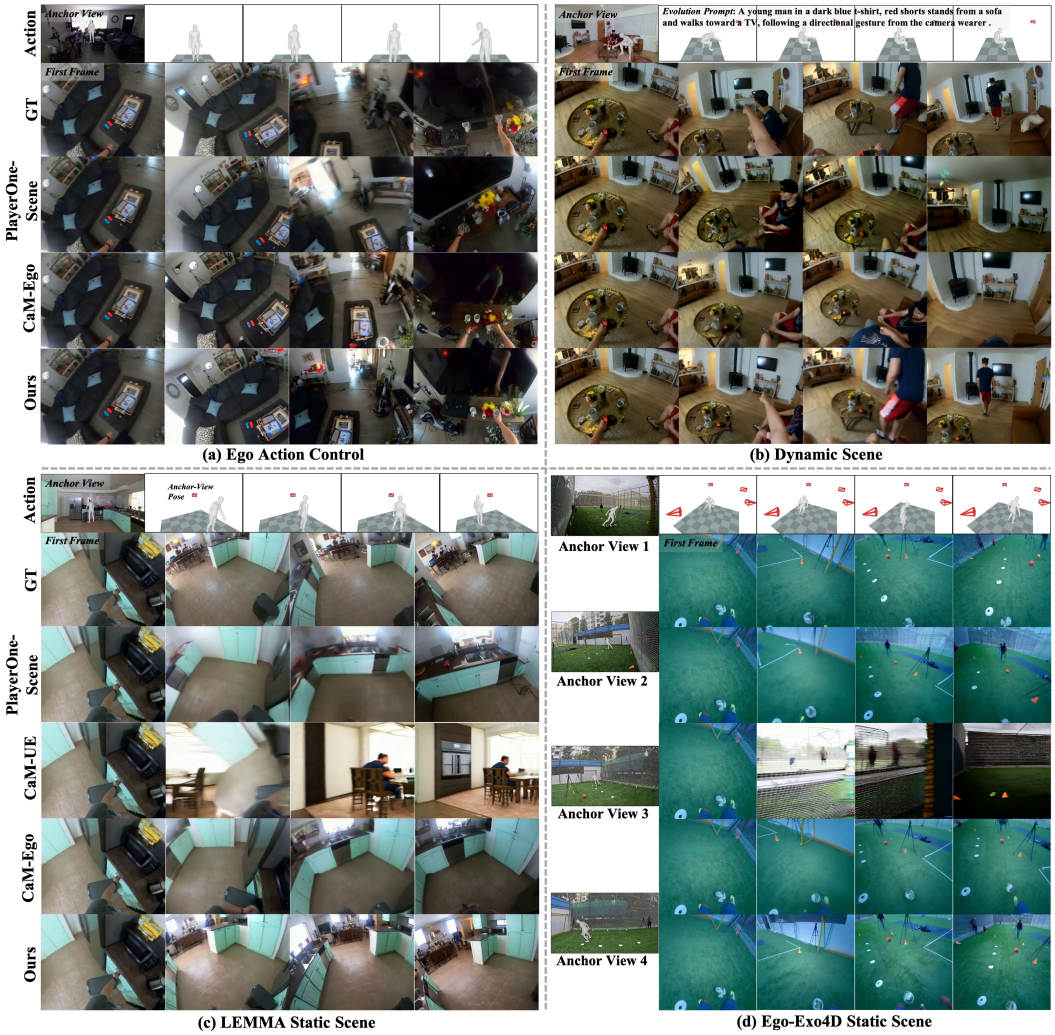


Figure 4: Qualitative Comparison. The gray mask denotes the human action and its location in the anchor view. During inference, the gray-masked region in the anchor view is inpainted. Red wireframes visualize the 3D anchor-view poses. Our method achieves better egocentric action control, scene consistency under large viewpoint changes, and dynamic scene evolution.

to incorporate anchor-view priors for world customization. In Stage III, we train the model on static scenes to learn pose-aware anchor-view conditioning for consistent egocentric roaming. In Stage IV, we mix in dynamic data with evolution descriptions to model text-driven local state changes.

4 Experimental Results

4.1 Experiment Settings

Implementation Details. We adopt Wan2.2 TI2V 5B [46] as the base model and synthesize 77-frame videos at 480p resolution under an image-to-video formulation. For exocentric training, we use an internally curated dataset of 200K single-person action videos and 101K videos from the UE-based MultiCamVideo dataset [2]. For egocentric training, we require synchronized third-person and first-person views to extract the camera wearer's human pose and anchor-view information; therefore, we use Ego-Exo4D [15] and LEMMA [22], which provide synchronized cross-view pairings, diverse egocentric interactions, and dynamic scenes. We use GVHMR [40] to estimate both 3D human motion and anchor-view poses in a shared 3D global coordinate system. The estimated motion

Table 1: Quantitative results. For static scenes, we evaluate scene consistency, camera accuracy, and video quality. For dynamic scenes, we additionally report text alignment with the evolution prompts. For the CineScene test set, due to its data characteristics, we report only the applicable metrics.

Method	Scene Consistency					Camera Accuracy			Text Alignment	Video Quality
	Mat. Pix.(K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓	ATE↓	RTE↓	RRE↓	VideoAlign-TA↑	VBench↑
EGO STATIC SCENE										
PlayerOne	3961.6	0.845	13.26	0.459	0.596	0.131	0.037	3.741	–	0.734
PlayerOne-Scene	4334.8	0.864	14.38	0.500	0.545	0.142	0.032	3.353	–	0.735
CaM-UE	3706.9	0.804	11.57	0.448	0.686	0.163	0.040	3.590	–	0.729
CaM-Ego	4379.4	0.872	15.16	0.554	0.515	0.125	0.032	3.207	–	**0.748**
Ours	**4493.4**	**0.885**	**16.06**	**0.578**	**0.470**	**0.112**	**0.029**	**3.145**	–	0.748
UE STATIC CINESCENE										
PlayerOne	3947.0	0.787	–	–	–	–	–	2.438	–	0.736
PlayerOne-Scene	4413.5	0.802	–	–	–	–	–	2.401	–	0.737
CaM-UE	4301.1	**0.852**	–	–	–	–	–	1.722	–	0.750
CaM-Ego	4429.1	0.842	–	–	–	–	–	2.009	–	**0.770**
Ours	**4555.1**	0.851	–	–	–	–	–	**1.656**	–	0.769
EGO DYNAMIC SCENE										
PlayerOne-Scene	4455.4	0.864	14.24	0.454	0.583	0.067	0.017	1.784	0.449	0.756
CaM-UE	4466.5	0.856	12.82	0.462	0.627	0.083	0.018	**1.230**	0.115	0.770
CaM-Ego	4459.0	0.871	14.57	0.501	0.574	0.083	0.016	1.636	0.385	0.770
Ours	**4634.6**	**0.899**	**16.37**	**0.555**	**0.486**	**0.048**	**0.013**	1.346	**0.717**	**0.774**

contains 22 major body joints, excluding hand poses due to unreliable estimation in current egocentric data, as also noted in GigaHands [11]. More details are provided in Appendix A.

Baselines. We compare with baselines across three tasks: (1) **Egocentric Action Control**: We use PlayerOne [45] as the main baseline, which decomposes human pose into body-part controls. Since its official code is unavailable, we re-implement it on Wan2.2 TI2V 5B using our training data, excluding hand poses for fairness due to unreliable estimation. (2) **Static Scene Consistency**: We evaluate PlayerOne with our anchor-view injection mechanism, denoted as PlayerOne-Scene. We also compare with CaM [57], which takes camera poses, scene context, and scene viewpoints as inputs, training two variants on our egocentric data and the official UE dataset. CineScene [19] and SWM [37] are excluded due to FOV issues and unavailable code, respectively. (3) **Dynamic Scene Evolution**: Since no prior work shares the same setting, we adapt the static-scene baselines by appending evolution prompts to their global text prompts.

Evaluation. We evaluate the generated results from four aspects: (1) **Action Accuracy**: As most body parts are out of view in egocentric videos, we quantify action controllability through camera-view control and qualitatively assess limb motion. Following MegaSaM [28], we use camera pose error metrics, including Absolute Translation Error (ATE), Relative Translation Error (RTE), and Relative Rotation Error (RRE). We estimate camera trajectories from synthesized videos using MegaSaM. (2) **Scene Consistency**: Following prior works [2, 19], we report GIM-based Mat. Pix. [38] to measure the ratio of matched pixels, CLIP-V [35] for semantic similarity, pixel-aligned metrics including PSNR and SSIM [49], and the perceptual metric LPIPS [62]. (3) **Dynamic Evolution**: We adopt the Text Alignment (TA) metric from VideoAlign [30] to measure semantic consistency with anchor-view evolution prompts. (4) **Video Quality**: We evaluate visual quality using VBench [21]. Averaged results are reported in Table 1, with detailed results for each evaluation dimension provided in Table 6.

Test Sets. To evaluate performance and generalization, we construct four test sets: (1) **Egocentric Static Test Set**: 100 held-out sequences from the same data sources as the training set, featuring significant motion and viewpoint variations. (2) **UE Test Set**: 100 Unreal Engine (UE) [8] sequences filtered from CineScene [19], whose initial frames do not overlap with the provided anchor views. Since CineScene camera trajectories are repurposed as character head poses, we retain only in-place rotational motions to avoid unnatural poses while preserving large viewpoint changes. Thus, viewpoint accuracy is evaluated only by RRE, and scene consistency is assessed only using CLIP-V and GIM due to inconsistent camera intrinsics. (3) **Real-World Test Set**: sequences captured from diverse real-world scenes with anchor views and human actions under large viewpoint changes, used only for qualitative evaluation due to unavailable ground-truth references. (4) **Egocentric Dynamic Test Set**: 100 held-out sequences from training data with pronounced dynamic human activities. We do not include out-of-domain dynamic data, as collecting such data remains challenging.

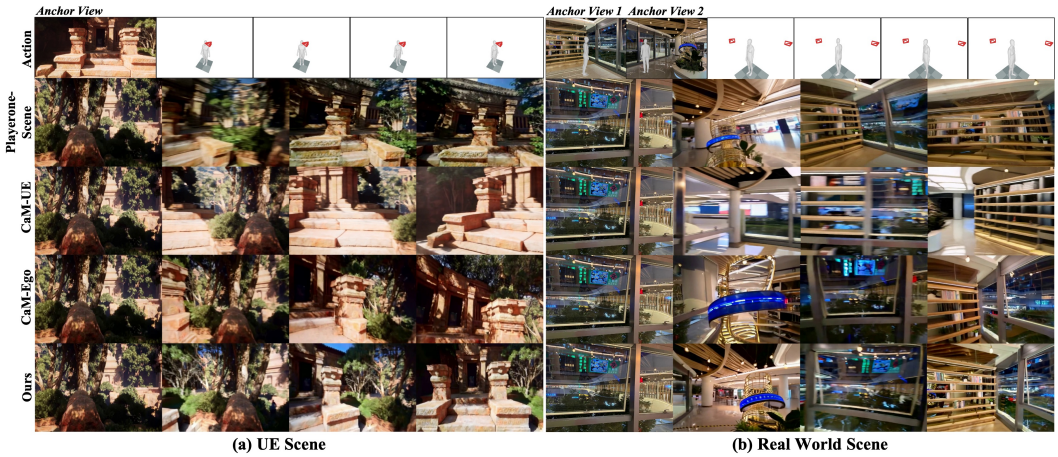


Figure 5: Qualitative comparison on rendered UE scenes and real-world captured scenes.

4.2 Comparisons

Quantitative Results. As shown in Table 1, our method achieves the best results in scene consistency, camera accuracy, and text alignment across all test scenarios, while maintaining comparable visual quality. PlayerOne learns from incomplete supervision targets captured from first-person videos, leading to weaker motion control. PlayerOne lacks scene information as input, and CaM-UE is trained only on UE data with slow camera motion; therefore, both methods perform poorly in scene consistency. Although PlayerOne-Scene and CaM-Ego are trained on the same data as ours, both exhibit weaker spatial perception than our projection-based control learning scheme. PlayerOne-Scene is limited by its part-wise learning scheme in modeling spatial pose variations, while CaM-Ego only takes viewpoint information as input. Notably, since our method supports state evolution control via evolution prompts, its advantage becomes more pronounced in dynamic scenes.

Qualitative Results. Figure 4 presents visual comparisons across multiple test tasks. Our method shows superior performance in egocentric human motion control, scene consistency under large viewpoint changes, and dynamic scene evolution driven by evolution prompts. CaM-Ego only controls viewpoint changes without body motion input, while PlayerOne-Scene suffers from limited motion accuracy due to its part-wise control scheme. Additional action control results are shown in Figures 9 and 10. In addition, more results on evolution prompt control are shown in Figure 8. We further evaluate our method on out-of-distribution UE scenes and real-world scenes in Figure 5, where there is limited or no overlap between the anchor view and the initial ego-view frame. The results show that our method exhibits strong generalization ability.

4.3 Ablation Studies

Ablations of Design Strategies. We conduct ablation studies on key design choices in Table 2. Under the action control setting, Stage-I third-person video training and the projection-based control design are essential, as shown quantitatively and visually in Figure 9. Removing them also weakens scene consistency in later stages by degrading spatial perception. In addition, removing anchor-view pose or anchor-view RoPE leads to worse scene consistency, confirming their roles in providing pose-aware view conditioning and distinguishing multiple anchor views. Finally, we validate the effectiveness of the multi-stage training strategy through mixed-training variants across stages.

Out-of-Sight Scene Evolution. As shown in Figure 6, we evaluate whether our model can infer scene dynamics beyond the initial egocentric view. We construct cases where another person appears in the anchor view but is initially invisible from the egocentric perspective, becoming visible only after a viewpoint change by the first-person player. We vary the timing of this viewpoint change by modifying the egocentric human motion. For example, when the caption describes a person standing up from a sofa, an earlier viewpoint change reveals the person still sitting at frame 25 and subsequently standing up, whereas a delayed change first reveals them already standing at frame 60.

Table 2: Ablations on design choices. We compare design strategies across egocentric action control and evolvable anchor-view customization. TA denotes the VideoAlign text-alignment score. "Joint" denotes joint training of the corresponding stages.

Variant	Scene Consistency					Camera Accuracy			Text Alignment
	Mat. Pix.(K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓	ATE↓	RTE↓	RRE↓	TA↑
Egocentric Action Control *(Camera Accuracy)*									
w/o Stage I	–	–	–	–	–	0.125	0.033	3.532	–
w/o Head Pose	–	–	–	–	–	0.123	0.032	3.806	–
Joint Stage I & II	–	–	–	–	–	0.123	**0.030**	3.372	–
Ours	–	–	–	–	–	**0.112**	0.030	**3.187**	–
Anchor-View Customization *(Ego Static Scene)*									
w/o Stage I	4438.3	0.877	15.50	0.557	0.492	0.116	0.031	3.351	–
w/o Head Pose	4425.4	0.877	15.42	0.561	0.502	0.119	0.032	3.395	–
w/o Stage II	4416.1	0.879	15.68	0.571	0.487	0.115	0.031	3.234	–
w/o Anchor-View Pose	4401.7	0.879	15.59	0.568	0.493	0.112	0.033	3.184	–
w/o Anchor-View RoPE	4395.2	0.878	15.40	0.564	0.498	0.110	0.031	3.162	–
Joint Stage III & IV	4442.6	0.877	15.59	0.570	0.489	**0.109**	0.031	3.180	–
Ours	**4493.4**	**0.885**	**16.06**	**0.578**	**0.470**	0.112	**0.029**	**3.145**	–
Anchor-View Evolution *(Ego Dynamic Scene)*									
Joint Stage III & IV	4573.4	0.893	15.67	0.522	0.502	0.050	0.014	1.362	0.703
Ours	**4634.6**	**0.899**	**16.37**	**0.555**	**0.486**	**0.048**	**0.013**	**1.346**	**0.717**

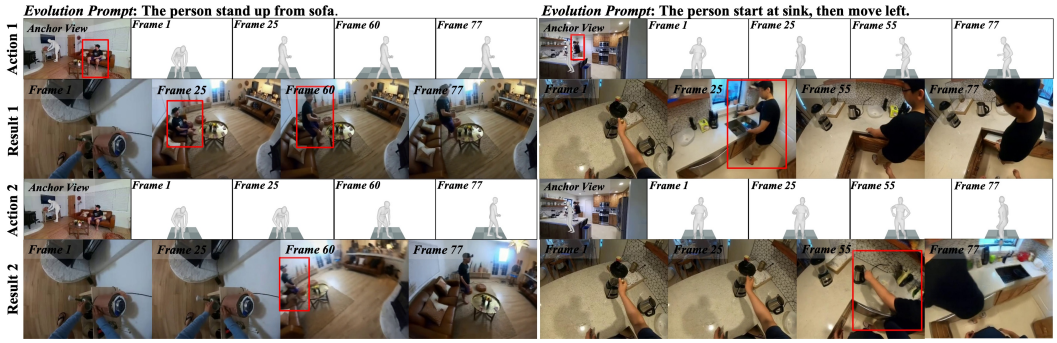


Figure 6: Out-of-Sight Scene Evolution. We show that our model can infer scene evolution beyond the observed view by varying the timing of the action-induced viewpoint transition. Even when dynamic scene elements are not visible, our model can still reason about their state changes.

Figure 7: Spatial Pose Awareness. We horizontally flip the human pose and anchor-view pose while keeping the anchor-view image fixed, creating overlapping and non-overlapping view settings.

Spatial Pose Awareness. As shown in Figure 7, we flip the human and anchor-view poses, forming overlapping and non-overlapping settings. The results show that our method understands spatial pose relationships and retrieves appearance details when the poses overlap.

5 Conclusion and Limitations

In this work, we introduced AnchorWorld, a framework for world-customizable embodied egocentric simulation that integrates natural embodied action control with localized world-state customization. Specifically, AnchorWorld leverages third-person videos to provide rich interaction context and complete human motion supervision, and employs projection-based action control to support hybrid-view training, while pose-associated anchor views provide spatially grounded appearance priors and text-driven local scene evolution. Extensive experiments demonstrate that AnchorWorld consistently surpasses existing methods, while ablations validate each key design. Additionally, AnchorWorld still has several limitations, including challenges in long-term exploration, open-world generalization, and diverse dynamic scenario modeling, which are discussed in detail in Appendix B.

References

[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

[2] Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14834–14844, 2025.

[3] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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

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

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

[7] Xiaowei Chi, Peidong Jia, Chun-Kai Fan, Xiaozhu Ju, Weishi Mi, Kevin Zhang, Zhiyuan Qin, Wanxin Tian, Kuangzhi Ge, Hao Li, et al. Wow: Towards a world omniscient world model through embodied interaction. *arXiv preprint arXiv:2509.22642*, 2025.

[8] Epic Games. Unreal engine 5. https://www.unrealengine.com/en-US/unreal-engine-5, 2022. Accessed: 2025-09-25.

[9] Yao Feng, Chendong Xiang, Xinyi Mao, Hengkai Tan, Zuyue Zhang, Shuhe Huang, Kaiwen Zheng, Haitian Liu, Hang Su, and Jun Zhu. Vidarc: Embodied video diffusion model for closed-loop control. *arXiv preprint arXiv:2512.17661*, 2025.

[10] Rafail Fridman, Amit Abecasis, Yoni Kasten, and Tali Dekel. Scenescape: Text-driven consistent scene generation. *Advances in Neural Information Processing Systems*, 36:39897–39914, 2023.

[11] Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17461–17474, 2025.

[12] Xiao Fu, Xian Liu, Xintao Wang, Sida Peng, Menghan Xia, Xiaoyu Shi, Ziyang Yuan, Pengfei Wan, Di Zhang, and Dahua Lin. 3dtrajmaster: Mastering 3d trajectory for multi-entity motion in video generation. *arXiv preprint arXiv:2412.07759*, 2024.

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

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

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

[16] Jinkun Hao, Mingda Jia, Ruiyan Wang, Xihui Liu, Ran Yi, Lizhuang Ma, Jiangmiao Pang, and Xudong Xu. Egosim: Egocentric world simulator for embodied interaction generation. *arXiv preprint arXiv:2604.01001*, 2026.

[17] Yicong Hong, Yiqun Mei, Chongjian Ge, Yiran Xu, Yang Zhou, Sai Bi, Yannick Hold-Geoffroy, Mike Roberts, Matthew Fisher, Eli Shechtman, et al. Relic: Interactive video world model with long-horizon memory. *arXiv preprint arXiv:2512.04040*, 2025.

[18] Jiaxin Huang, Yuanbo Yang, Bangbang Yang, Lin Ma, Yuewen Ma, and Yiyi Liao. Gen3r: 3d scene generation meets feed-forward reconstruction. *arXiv preprint arXiv:2601.04090*, 2026.

[19] Kaiyi Huang, Yukun Huang, Yu Li, Jianhong Bai, Xintao Wang, Zinan Lin, Xuefei Ning, Jiwen Yu, Pengfei Wan, Yu Wang, et al. Cinescene: Implicit 3d as effective scene representation for cinematic video generation. *arXiv preprint arXiv:2602.06959*, 2026.

[20] Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44(6):1–15, 2025.

[21] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

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

[23] Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. *arXiv preprint arXiv:2503.19907*, 2025.

[24] Byungjun Kim, Taeksoo Kim, Junyoung Lee, and Hanbyul Joo. Dexterous world models. *arXiv preprint arXiv:2512.17907*, 2025.

[25] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[26] Dayou Li, Lulin Liu, Bangya Liu, Shijie Zhou, Jiu Feng, Ziqi Lu, Minghui Zheng, Chenyu You, and Zhiwen Fan. Egocentric world model for photorealistic hand-object interaction synthesis. *arXiv preprint arXiv:2603.13615*, 2026.

[27] Yu Li, Menghan Xia, Gongye Liu, Jianhong Bai, Xintao Wang, Conglang Zhang, Yuxuan Lin, Ruihang Chu, Pengfei Wan, and Yujiu Yang. Adaviewplanner: Adapting video diffusion models for viewpoint planning in 4d scenes. *arXiv preprint arXiv:2510.10670*, 2025.

[28] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. Megasam: Accurate, fast and robust structure and motion from casual dynamic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10486–10496, 2025.

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

[30] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025.

[31] 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. *arXiv preprint arXiv:2512.22096*, 2025.

[32] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025.

[33] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[36] Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6132, 2025.

[37] Junyoung Seo, Hyunwook Choi, Minkyung Kwon, Jinhyeok Choi, Siyoon Jin, Gayoung Lee, Junho Kim, JoungBin Lee, Geonmo Gu, Dongyoon Han, et al. Grounding world simulation models in a real-world metropolis. *arXiv preprint arXiv:2603.15583*, 2026.

[38] Xuelun Shen, Zhipeng Cai, Wei Yin, Matthias Müller, Zijun Li, Kaixuan Wang, Xiaozhi Chen, and Cheng Wang. Gim: Learning generalizable image matcher from internet videos. *arXiv preprint arXiv:2402.11095*, 2024.

[39] Yifan Shen, Jiateng Liu, Xinzhuo Li, Yuanzhe Liu, Bingxuan Li, Houze Yang, Wenqi Jia, Yijiang Li, Tianjiao Yu, James Matthew Rehg, et al. Egoforge: Goal-directed egocentric world simulator. *arXiv preprint arXiv:2603.20169*, 2026.

[40] 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 2024 Conference Papers*, pages 1–11, 2024.

[41] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[42] 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.

[43] Junshu Tang, Jiacheng Liu, Jiaqi Li, Longhuang Wu, Haoyu Yang, Penghao Zhao, Siruis Gong, Xiang Yuan, Shuai Shao, Linfeng Zhang, et al. Hunyuan-gamecraft-2: Instruction-following interactive game world model. *arXiv preprint arXiv:2511.23429*, 2025.

[44] 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.

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

[46] 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.

[47] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[48] Yuxi Wang, Wenqi Ouyang, Tianyi Wei, Yi Dong, Zhiqi Shen, and Xingang Pan. Hand2world: Autoregressive egocentric interaction generation via free-space hand gestures. *arXiv preprint arXiv:2602.09600*, 2026.

[49] 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.

[50] Zile Wang, Zexiang Liu, Jaixing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026.

[51] Jiannan Xiang, Yi Gu, Zihan Liu, Zeyu Feng, Qiyue Gao, Yiyan Hu, Benhao Huang, Guangyi Liu, Yichi Yang, Kun Zhou, et al. Pan: A world model for general, interactable, and long-horizon world simulation. *arXiv preprint arXiv:2511.09057*, 2025.

[52] 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.

[53] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025.

[54] Deheng Ye, Fangyun Zhou, Jiacheng Lv, Jianqi Ma, Jun Zhang, Junyan Lv, Junyou Li, Minwen Deng, Mingyu Yang, Qiang Fu, et al. Yan: Foundational interactive video generation. *arXiv preprint arXiv:2508.08601*, 2025.

[55] Zixuan Ye, Xuanhua He, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Qifeng Chen, and Wenhan Luo. Unic: Unified in-context video editing. *arXiv preprint arXiv:2506.04216*, 2025.

[56] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5916–5926, 2025.

[57] Jiwen Yu, Jianhong Bai, Yiran Qin, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Context as memory: Scene-consistent interactive long video generation with memory retrieval. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

[58] Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 100–111, 2025.

[59] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024.

[60] Chenyangguang Zhang, Botao Ye, Boqi Chen, Alexandros Delitzas, Fangjinhua Wang, Marc Pollefeys, and Xi Wang. Controllable egocentric video generation via occlusion-aware sparse 3d hand joints. *arXiv preprint arXiv:2603.11755*, 2026.

[61] Lvmin Zhang, Shengqu Cai, Muyang Li, Chong Zeng, Beijia Lu, Anyi Rao, Song Han, Gordon Wetzstein, and Maneesh Agrawala. Pretraining frame preservation in autoregressive video memory compression. *arXiv preprint arXiv:2512.23851*, 2025.

[62] 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*, pages 586–595, 2018.

[63] 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. *arXiv preprint arXiv:2512.08931*, 2025.

Appendix

The appendix consists of four sections. Readers can click on each section number to navigate to the corresponding section:

- Section **A** provides more implementation details.
- Section **B** describes the limitations.
- Section **C** describes additional analyses and results, including dynamic evolution prompt control, egocentric and exocentric action control, and real-world hard scenes.
- Section **D** describes the failure cases.

A Implementation Details

Table 3: Overview of the progressive training stages.

Setting	Stage I	Stage II	Stage III	Stage IV
Objective	Exocentric Motion	Egocentric Motion	Static Scene	Dynamic Scene
Training Data	Internal videos; MultiCamVideo [2]	Ego-Exo4D [15]; LEMMA [22]	Ego-Exo4D [15]; LEMMA [22]	Ego-Exo4D [15]; LEMMA [22]
Data Scale	200K+101K	100K	25K	25K+10K
Iterations	30K	15K	10K	10K
Batch Size	16	16	16	16
Learning Rate	1×10^{-4}	1×10^{-4}	1×10^{-4}	1×10^{-4}
Optimizer	AdamW	AdamW	AdamW	AdamW
Compute Resources	16 NVIDIA GPUs@80G	16 NVIDIA GPUs@80G	16 NVIDIA GPUs@80G	16 NVIDIA GPUs@80G
GPU Hours	600	300	253	253

We adopt Wan2.2 TI2V 5B [46] as our base video generation model and train it in an image-to-video manner. Our training follows a progressive strategy, as summarized in Table 3. In the data scale row, Stage I uses 200K internally curated real single-person action videos and 101K synthetic UE videos from MultiCamVideo [2]; Stage II uses 100K egocentric action samples; Stage III uses 25K filtered samples with large viewpoint changes; and Stage IV jointly trains on the 25K static-scene samples from Stage III and 10K filtered dynamic-scene samples with noticeable human activities.

All videos are processed at 480p resolution while preserving their original aspect ratios, which retains visual content and avoids geometric distortion. All training stages are conducted on 16 NVIDIA GPUs with a total batch size of 16, a learning rate of 1×10^{-4}, a timestep shift of 15, and the AdamW optimizer. During training, pose conditions and anchor-view information are independently dropped with a probability of 5%. During inference, we use 50 denoising steps and set the classifier-free guidance scale to 5.

For egocentric video data, LEMMA [22] provides one anchor view for each sample, whereas Ego-Exo4D [15] contains one to six anchor views captured from different viewpoints. Due to the construction procedure of these paired third-person-to-first-person datasets, the anchor view images may contain the first-person player. Ideally, an anchor view should be defined independently of the player and thus should not include the player itself. However, given the relatively low data resolution, directly masking the player and applying inpainting would introduce visible artifacts and degrade image quality. Therefore, we do not apply inpainting during training. Importantly, using clean anchor-view images at inference time does not adversely affect the results. This can be attributed to two factors: (1) supervision from first-person videos enables the model to learn to ignore the player when interpreting anchor views; and (2) our input conditions include both human pose and view pose information, which allows the model to determine spatial relationships based primarily on pose cues.

For Ego-Exo4D, we undistort the egocentric fisheye videos and apply moderate brightness enhancement due to their low illumination. In addition, Ego-Exo4D exhibits noticeable color discrepancies between third-person and first-person videos, as these videos are captured by the different cameras. Nevertheless, our model can leverage valuable scene information from the anchor view while maintaining a color tone consistent with the initial ego-view frame.

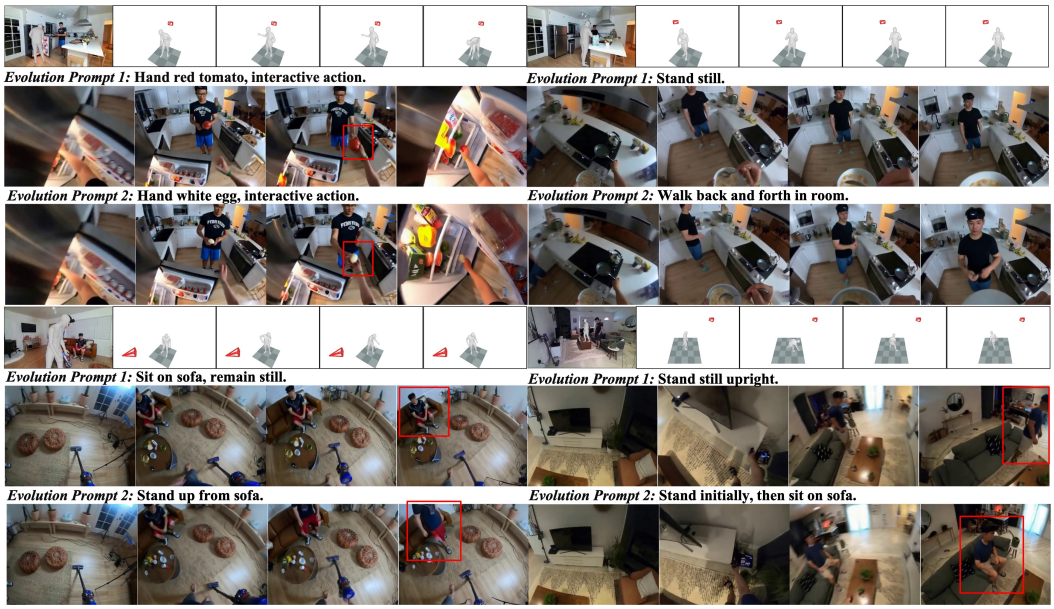


Figure 8: Evolution prompt control. We demonstrate that, within the same anchor-view image, modifying the evolution prompt enables control over different state changes.

For human motion and anchor-view pose estimation, we use GVHMR [40]. Specifically, we estimate 3D human motion from third-person views and canonicalize each sequence by placing the initial pose at the origin and aligning its horizontal orientation. The estimated motion contains 22 major body joints, excluding hand poses because hand estimation is unreliable in current egocentric data, due to frequent out-of-view hands, occlusions, and multi-person interference, as also noted in GigaHands [11]. We further use GVHMR to estimate anchor-view poses relative to the target subject, thereby unifying human motion and anchor viewpoints in a shared 3D global coordinate system. In multi-person scenes, the estimated human motion may correspond to a non-egocentric subject. We therefore manually inspect the annotations, correct subject assignments when necessary, and discard samples with low-quality motion estimation.

For evolution prompts, they are annotated by Qwen3-VL-32B-Instruct [3] using carefully designed prompt templates, as shown in Table 7.

B Limitation

Long-Term Exploration. In this work, we primarily focus on scenarios involving short video clips. However, enabling longer-horizon world exploration and interaction is essential for future progress. To this end, we plan to extend our framework toward real-time autoregressive interaction. We note that, in first-person settings, an embodied agent may continuously interact with the environment and explore larger-scale scenes. During this process, the model must update environmental state changes induced by its own actions in real time. Addressing this challenge requires a stronger emphasis on long-term memory mechanisms [61] within the model.

Open World. In this work, the training data primarily focuses on a constrained set of scenarios. In the future, collecting open-world data to construct broader environments and support longer-horizon world exploration will be an important direction.

Diverse Dynamic Scenarios. Due to limitations in current egocentric training data, which typically provide multiple viewpoints of the same dynamic human activity, our empirical implementation uses a globally consistent evolution description for all anchor-view priors, i.e., $t_1 = \cdots = t_n$, and mainly focuses on human-related activities rather than diverse dynamic scenarios. Future work can extend our framework to more diverse scenarios and anchor-specific dynamic controls, while incorporating the natural dynamic evolution of the world, thereby enabling the construction of more realistic and temporally rich worlds.

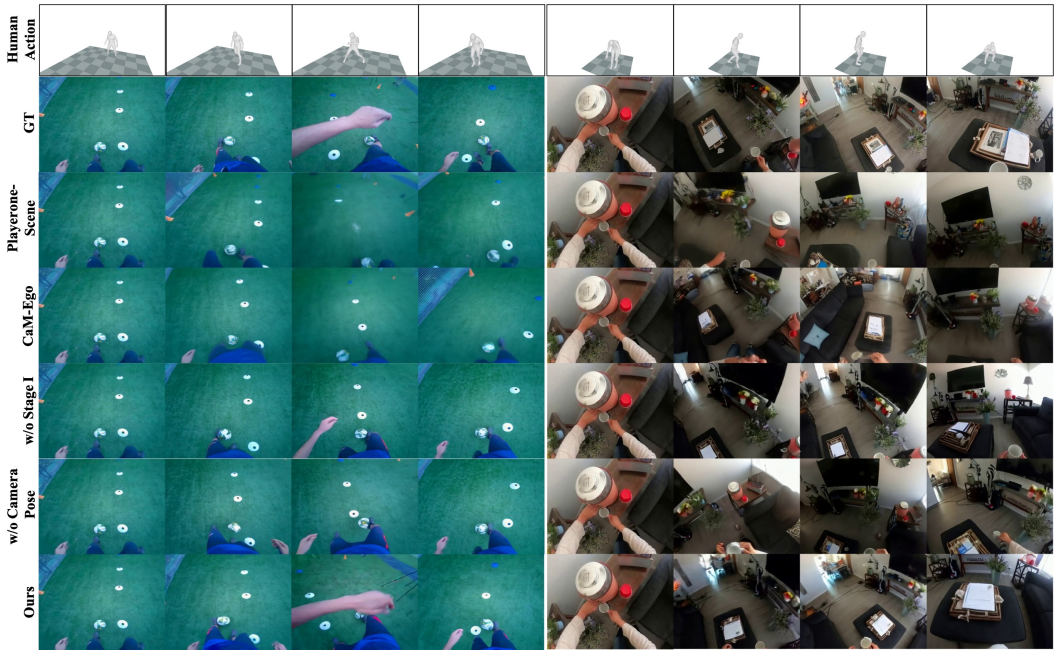


Figure 9: Visualization results of egocentric action control. We show the results compared with baseline methods and our ablation settings.

C Additional Analyses and Results

C.1 Evolution Prompt Control

As shown in Figure 8, we achieve different dynamic evolutions of the scene by modifying the evolution prompt. This demonstrates that our method provides flexible support for diverse dynamic evolutions, allowing users to describe anchor-specific dynamic evolution.

C.2 Egocentric Action Control

Since most body motions are not visible in egocentric videos, we conduct qualitative comparisons to evaluate the performance of different methods on ego human action control. Figure 9 shows the results on the in-domain test set, while Figure 10 further compares the results in real-world scenarios. The results demonstrate the superior performance of our projection-based control method. PlayerOne suffers from inaccurate body-motion control, whereas CaM-Ego only supports viewpoint control.

In addition, Figure 9 presents a qualitative comparison of the ablation study on the design of egocentric action control. The results show that the absence of motion knowledge from third-person video data, as well as the use of non-projection-based action control, leads to reduced accuracy in body-motion control.

C.3 Real World Scene

To evaluate the generalization ability of our method, we construct test data through real-world capture, as shown in Figure 11. In addition to the single-anchor-view setting, we construct a multi-anchor-view setting by capturing multiple scene images, where the subject undergoes continuous viewpoint changes that overlap with different anchor views. Furthermore, to verify that our method infers spatial locations from spatial poses rather than relying on overlapping RGB information, we construct test data in which the anchor-view image and the first ego-view frame have no visual overlap by performing coordinate transfer through multiple captures, as illustrated in Figure 11 (a). The results show that our method can still generate correct outputs under this challenging setting, demonstrating its spatial awareness.

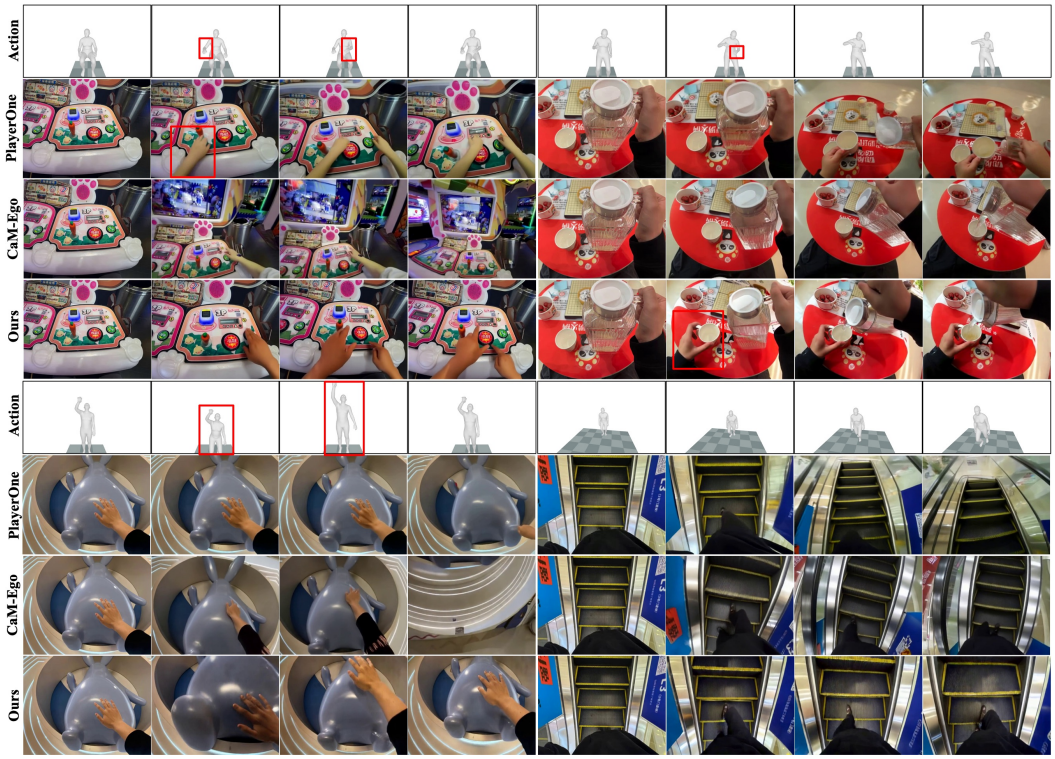


Figure 10: Visualization results of egocentric action control in real-world scenes. Our method generates stable results in response to diverse body motions, such as pouring water, squatting and jumping, and walking up stairs.

C.4 Scene Coherence.

As shown in Figure 12, we consider two challenging settings: (i) replacing the anchor-view image with another image of a different style, such that the first ego-view frame and the provided anchor scene no longer describe the same underlying world; and (ii) using the same anchor-view image while simultaneously flipping both the anchor-view pose and the human pose, which mirrors the world space horizontally. In both settings, the human pose and the anchor-view pose still exhibit apparent view overlap. However, the generated videos may become inconsistent or visually incoherent. For setting (i), this is because the model is forced to refer to an anchor scene that is incompatible with the ego-view observation. For setting (ii), when the world-space geometry becomes inconsistent or physically implausible, the model struggles to generate reasonable results, as can be observed from the wall surface in the first row of Figure 12. These results indicate that video generation models internally require a continuous and complete world representation with spatially consistent geometry.

C.5 Exocentric Action Control

We report the ablation results of Stage I third-person human action control in Table 4. We use GVHMR [40] to estimate the 3D human poses of the generated videos, and compute MPJPE-related metrics against the ground-truth poses to measure control accuracy. The first row corresponds to using only 3D joint positions to represent joint information, instead of our 6D pose representation. The results show that, due to the lack of orientation information, this design hinders the model from fully understanding the 3D human pose, and may lead to incorrect human orientations in the generated results.

In addition, we explore different pose-condition injection strategies. The results demonstrate that our proposed spatial pose attention achieves the best performance and enables the model to correctly interpret the pose condition. This is because this design explicitly informs the model of the frame-level alignment between video tokens and pose tokens, and drops the pose tokens after attention, since

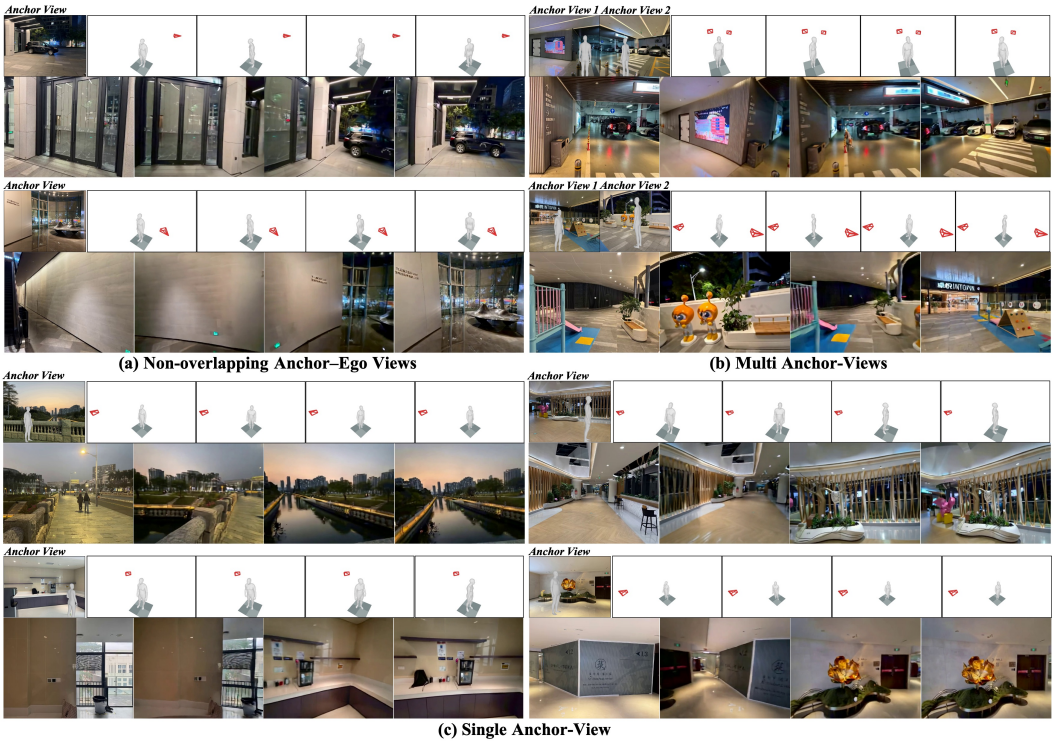


Figure 11: Visualization results in real-world scenes. We show that our method can generate stable results in scenes with non-overlapping viewpoints, as well as in both multi-anchor-view and single-anchor-view settings.

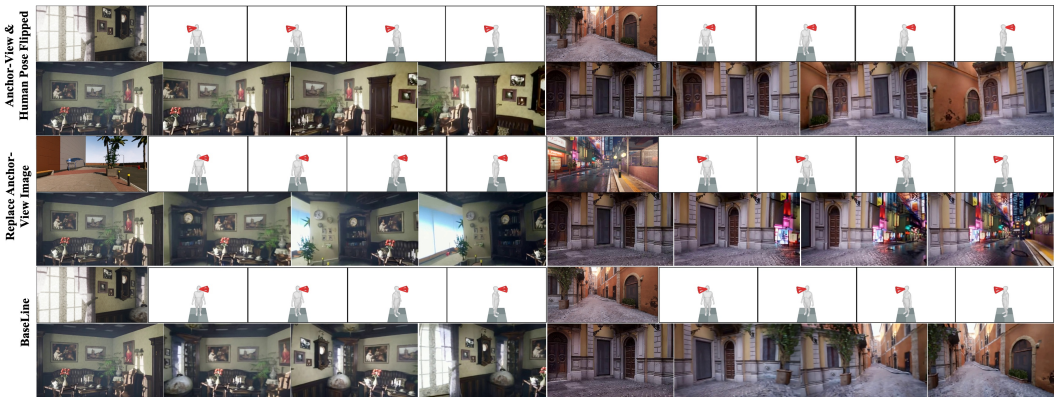


Figure 12: Visualization of scene coherence cases. We replace the anchor view with a style-mismatched image or mirror the world by flipping both the anchor-view and human poses. The results indicate that video generation models internally require a continuous and complete world representation with spatially consistent geometry.

there exists a distribution gap between pose features and VAE latents. Figure 13 shows visualization results of third-person action control.

C.6 Additional Quantitative Results

We show in Table 5 how the number of anchor views surrounding a world scene helps improve scene consistency performance. We also present in Table 6 the detailed per-dimension results of the average VBench metrics reported in Table 1.

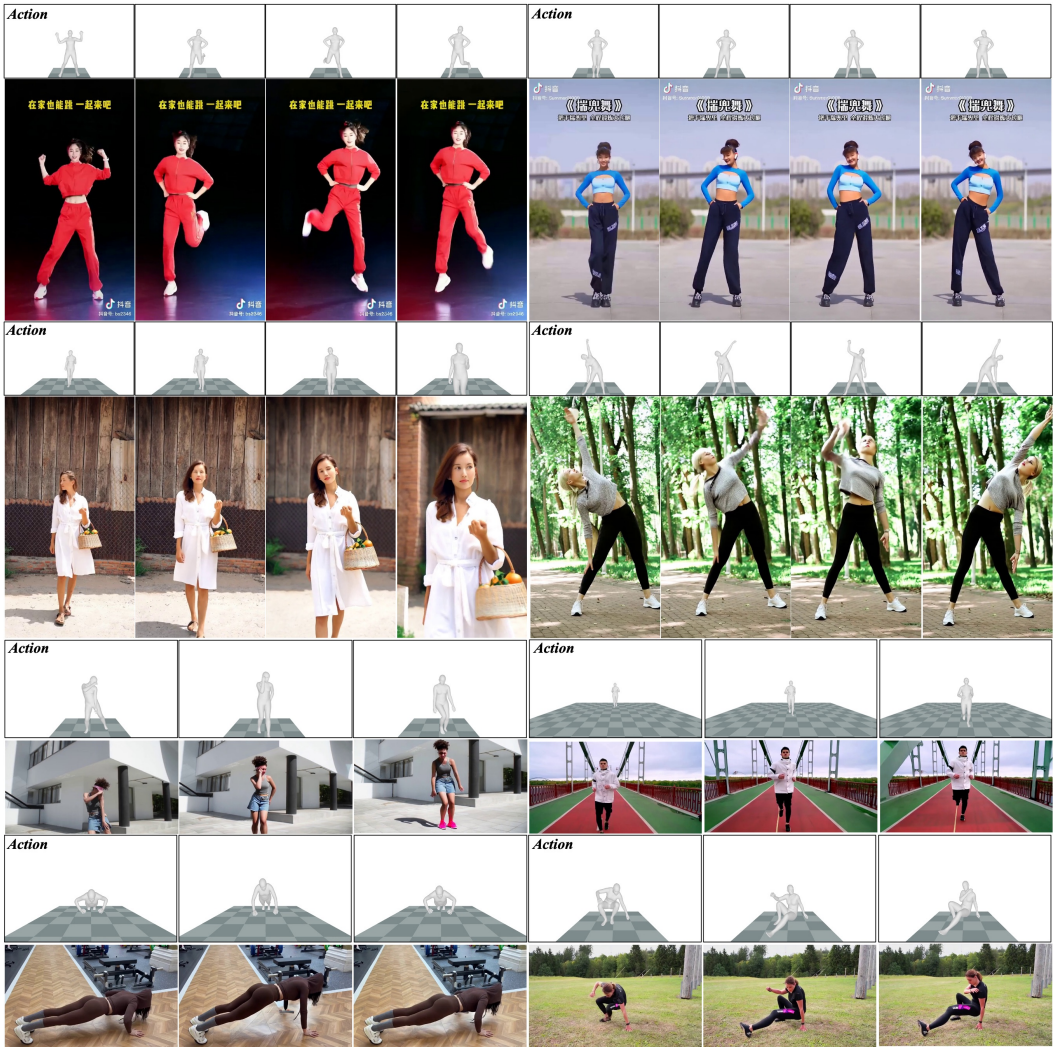


Figure 13: Visualization results of third-person human action control.

Table 4: Quantitative ablation results on **third-person action control**. We report WA-MPJPE and PA-MPJPE, where lower values indicate better performance.

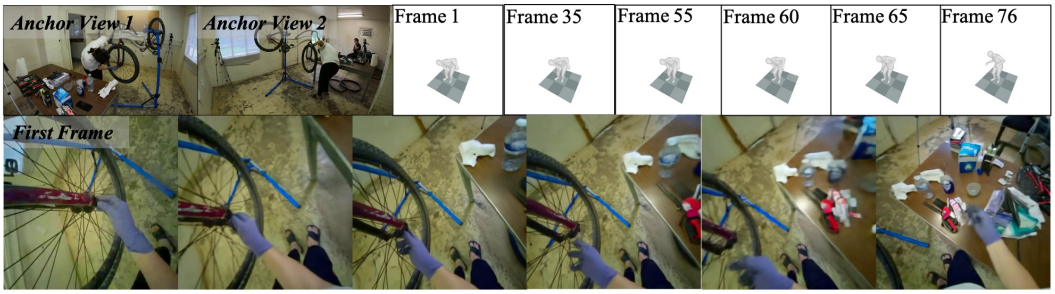
Method	**Motion Accuracy**	
	WA-MPJPE↓	PA-MPJPE↓
Joint Position Only	90.47	38.71
3D Pose Attention	188.17	82.37
Cross-Attention Fusion	187.55	88.23
In-Context Frame Concat	161.67	74.64
Ours	**74.57**	**28.01**

Table 5: Quantitative ablation results on the number of anchor views.

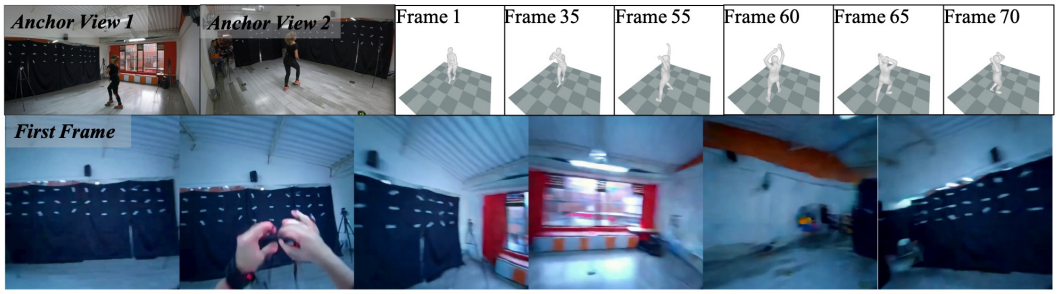
# Anchor Views	**Scene Consistency**				
	Mat. Pix.(K)↑	CLIP-V↑	PSNR↑	SSIM↑	LPIPS↓
1	4074.94	0.8605	14.9740	0.5600	0.5174
2	4152.91	0.8645	15.0294	0.5585	0.5178
3	**4233.59**	**0.8667**	**15.1877**	**0.5622**	**0.5104**

Table 6: Evaluation metrics cover the fine-grained dimensions of **VBench**: Subject Consistency (Sub. Cons.), Background Consistency (Bg. Cons.), Temporal Flickering (Temp. Flick.), Motion Smoothness (Mot. Smooth.), Imaging Quality (Img. Qual.), and Aesthetic Quality (Aes. Qual.).

Method	**VBench Dimensions**					
	Sub. Cons.↑	Bg. Cons.↑	Temp. Flick.↑	Mot. Smooth.↑	Img. Qual.↑	Aes. Qual.↑
Ego Static Scene						
PlayerOne	0.7956	0.8964	0.9474	0.9821	0.3945	0.3800
PlayerOne-Scene	0.8071	0.8974	0.9498	0.9820	0.3940	0.3803
CaM-UE	0.7694	0.8899	0.9357	0.9811	0.4099	0.3903
CaM-Ego	0.8142	0.9040	**0.9533**	**0.9851**	**0.4172**	0.4155
Ours	**0.8167**	**0.9041**	0.9523	0.9832	0.4140	**0.4171**
UE CineScene						
PlayerOne	0.7920	0.8600	0.9361	0.9818	0.4309	0.4125
PlayerOne-Scene	0.8147	0.8699	0.9444	0.9856	0.4026	0.4052
CaM-UE	0.8004	0.8961	0.9289	0.9903	0.4214	0.4631
CaM-Ego	0.8496	**0.9035**	**0.9426**	**0.9911**	0.4566	**0.4789**
Ours	**0.8522**	0.8986	0.9382	**0.9911**	**0.4571**	0.4781
Ego Dynamic Scene						
PlayerOne-Scene	0.8743	0.9140	0.9649	0.9889	0.4015	0.3941
CaM-UE	0.8824	0.9230	0.9586	**0.9921**	**0.4508**	0.4156
CaM-Ego	0.8751	0.9266	0.9669	0.9913	0.4388	0.4204
Ours	**0.8937**	**0.9295**	**0.9689**	0.9901	0.4371	**0.4272**



(a) Inconsistent fine-grained details



(b) Blurry frames from rapid panning

Figure 14: Failure cases. (a) Due to the limited capability of the base model, our method may struggle to preserve highly fine-grained texture details in scenes with complex local structures, leading to inconsistent scene details. (b) Since egocentric videos often involve rapid viewpoint changes, the training data contains blurry frames, which may result in blurry generation artifacts.

D Failure Cases

Inconsistent Scene Details. We observe that when local regions of a scene contain complex structures and rich texture details, our method may produce results with inconsistent fine-grained details, as shown in Figure 14 (a). We believe that this limitation is largely constrained by the capability of the base model. Specifically, the VAE of Wan TI2V 2.2 5B used in our experiments has a

spatial downsampling factor of 16, leading to a relatively high compression ratio in the latent spatial dimensions and thus the loss of fine-detail information. In the future, adopting more powerful base models is expected to alleviate this issue.

Blurry Results. Our first-person training data contains a large number of videos with rapid viewpoint changes, which often leads to motion blur in the frames. Consequently, the generated results may also exhibit similar blurring artifacts, as shown in Figure 14 (b). In addition, due to the limitations of the base model and the fast motion commonly present in first-person data, the generated hands may suffer from degraded visual quality.

Table 7: Instruction Template for Evolution Prompt

Role
You are an expert video analyst specializing in First-Person Perspective (FPV) footage.
Objective
Your primary task is to detect and describe the **external character(s)** visible in the video. Do not describe the first-person observer, the camera wearer, or their body parts.
Detection Phase
Scan the video for distinct people other than the camera wearer.
If no external person is present, or if people are too obscured to be identifiable, strictly output: `False`.
If distinct external people are visible, proceed to the description phase.
Description Strategy
Select the subject to describe according to the following rules:
1. **Main subject:** If one person is the clear focal point, such as interacting with the camera wearer, being closest to the observer, or performing a distinctive action, describe only this main person and ignore background extras.
2. **Collective group:** If multiple people are present without a clear protagonist, describe them as a single collective unit, such as "a group of students" or "a crowd of pedestrians".
3. **Constraint:** Do not enumerate individuals as "Character 1", "Character 2", or "Character 3". Describe either the main subject or the collective group state.
Description Aspects
Provide a concise and objective description covering the following four aspects:
a) Location & Position: Describe the subject's spatial location in the physical 3D environment. Relate the position to physical objects or spaces, such as "sitting on the sofa", "leaning against the wall", "standing in the doorway", or "walking down the hallway". Do not describe positions relative to the video frame, such as "in the center of the screen", "on the left side", or "bottom right".
b) Appearance: Describe visible physical traits and clothing. For groups, describe the shared appearance, such as suits or casual clothes.
c) Dynamic Actions: Describe the subject's movement or activity, focusing on the flow of motion.
d) Interaction with "I": Briefly state the interaction type if it exists, such as "talking to me", "handing an object", or "blocking my path". If no direct interaction is observed, state "None".
Constraints
Describe only visually confirmed content and avoid hallucinations. Keep the description concise and do not list multiple individuals separately.
Output example:
a) Location & Position: The person is standing in the living room, positioned between a coffee table and a TV console.
b) Appearance: The person is wearing a black short-sleeved T-shirt with a white graphic, gray knee-length shorts, flip-flops, and a backward black cap.
c) Dynamic Actions: The person is vacuuming the floor, pushing the upright vacuum cleaner forward and then bending down.
d) Interaction with "I": None. The person is engaged in cleaning and does not appear to interact with the camera wearer.