DISCO: WORLD MODELS WITH DISCRETE CAMERA MOTION CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Controllable video world models target interactive world exploration, where models must faithfully execute explicit action commands while preserving visual quality and temporal coherence. However, most existing approaches rely on continuous camera trajectories as action conditions, which often lead to unreliable action following, especially under complex motion sequences. In this work, we identify action representation entanglement as a key bottleneck in controllable video generation, and show that continuous camera representations lead to high feature similarity across distinct motion patterns, degrading action controllability. Based on this insight, we propose DisCo, a controllable video world model that conditions generation on a compact set of discrete action primitives to improve action separability. We further introduce DisCoBench, a comprehensive benchmark for evaluating the ability of models in short-term, long-horizon, and highly dynamic exploration scenarios. Extensive experiments demonstrate that DisCo achieves significantly more reliable action following while preserving visual quality.

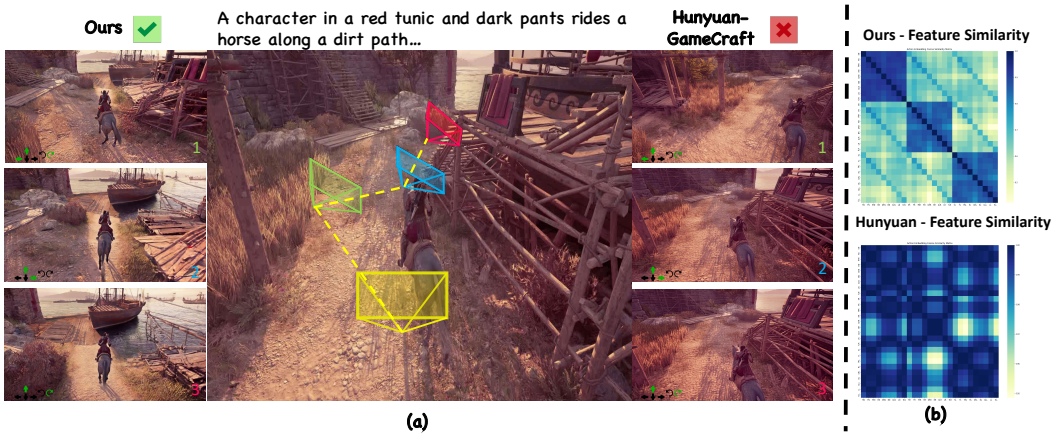

Figure 1: (a) Comparison between the continuous-action-based Hunyuan-GameCraft and our discrete-action-based approach. Our method successfully executes complex action commands, while Hunyuan-GameCraft fails under the same settings. (b) Heatmaps of pairwise action feature similarity, where darker colors correspond to higher similarity.

1 INTRODUCTION

World models learn the dynamics of an environment to forecast future states from past observations and actions (Ha & Schmidhuber, 2018; Hafner et al., 2019; Koh et al., 2021). They support model-based decision making by enabling planning and control through imagined rollouts, where an agent evaluates candidate action sequences before execution to improve data efficiency and safety (Bar et al., 2025; Jiang et al., 2025). Recently, advances in video diffusion models have made pixel-space world modeling increasingly practical (Blattmann et al., 2023; Wang et al., 2024b; Yang et al., 2025b; Kong et al., 2024; Wan et al., 2025; Mao et al., 2025; He et al., 2025b), with high-fidelity video generators serving as the backbone for high-fidelity simulation.

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.

Despite substantial progress in visual quality and temporal consistency, reliable action controllability remains a critical bottleneck in video world models (Li et al., 2025b;a; He et al., 2025b; Sun et al., 2025; Shin et al., 2025; Gao et al., 2025; Mao et al., 2025). Most existing approaches (Li et al., 2025b; Shin et al., 2025; Gao et al., 2025; Sun et al., 2025) rely on continuous camera trajectories as action conditions. While effective for coarse motion control, this formulation may collapse distinct action patterns into highly similar conditioning features, thus producing ambiguous guidance and degraded adherence when actions change frequently, as shown in Fig.1.

In this work, we posit that carefully designed discrete action representations provide a stronger foundation for controllable world modeling, as they naturally induce higher separability in the conditional feature space. By partitioning control signals into a compact set of geometrically meaningful action primitives, such representations could offer cleaner supervision for learning complex action compositions and action-conditioned dynamics. Building on this principle, we propose DisCo, a discrete action-conditioned video diffusion model for highly controllable world modeling. To support robust controllability under dynamic and complex scenarios, we first curate a camera-controlled video dataset that covers a diverse range of dynamic scenes and camera motion patterns. Starting from a pretrained video diffusion model, *i.e.*, Wan 2.2 (Wan et al., 2025), DisCo is fine-tuned for action-conditioned generation with structured motion representations derived from discretized camera controls. To enable efficient long-horizon video generation, we distill DisCo into a four-steps causal model, incorporating windowed attention and attention sinks with rolling KV caches to preserve temporal coherence and long-term consistency.

To comprehensively evaluate the proposed approach, we further introduce DisCoBench, a carefully curated benchmark that assesses controllable video generation across various scenarios. Unlike existing evaluations that primarily focus on visual and temporal quality, DisCoBench systematically assesses the ability of models to (i) accurately follow complex and compositional action commands in various scenes, (ii) maintain temporal coherence and visual consistency over long horizons, and (iii) handle scenarios involving highly dynamic elements. The results on DisCoBench show that DisCo exhibits superior controllability on challenging camera conditions, achieving an action following score of 0.633 compared to 0.588 for Matrix-Game2.

2 Related Work

2.1 Camera Controllable Video Generation

Camera control plays a central role in interactive video generation as it directly determines the viewpoint dynamics. Early methods (Wang et al., 2024b; He et al., 2025a; Wan et al., 2025) typically encode explicit camera intrinsics and extrinsics as control signals, which are then injected into diffusion models via ControlNet-style (Zhang et al., 2023) adapters. Recent approaches (Ren et al., 2025; Bahmani et al., 2025; Li et al., 2025a) further incorporate 3D geometric priors to improve scene consistency, while ReCamMaster (Bai et al., 2025a) and ReCapture (Zhang et al., 2025) focus on novel view synthesis by conditioning on relative camera pose trajectories. Despite their impressive visual quality, these approaches have two key limitations: (i) their controllability normally degrades when faced with complicated actions. (ii) raw numerical camera parameters are cumbersome and unintuitive for human interaction, limiting practical usability in interactive settings.

2.2 Auto-regressive Video Generation

Autoregressive generation (Kondratyuk et al., 2023; Sun et al., 2024; Wang et al., 2025a) has emerged as a powerful paradigm for video generation, typically utilizing next-token prediction to capture temporal dynamics (Yan et al., 2021; Wang et al., 2024a; 2025b). Some works have explored hybrid AR-diffusion architectures (Chen et al., 2024; Gao et al., 2024; Jin et al., 2024; Li et al., 2024; Guo et al., 2025), integrating efficiency of AR with high quality of diffusion. Most recently, Self-forcing (Huang et al., 2025b), and its following works (Cui et al., 2025; Liu et al., 2025; Lu et al., 2025) address the train-test gap by simulating inference conditions during training, and distill slow, multi-step teacher model into a fast, few-step AR student. Inspired by this, we develop DisCo and enhance it on controllability and consistency.

2.3 Video World Models

Video world models (Li et al., 2025b; Mao et al., 2025; He et al., 2025b; Yu et al., 2025b; Li et al., 2025c; Shin et al., 2025; Gao et al., 2025) have recently made significant strides by leveraging large-scale transformer architectures. Conditioning generation on actions or camera trajectories, these systems learn to synthesize high-fidelity, temporally coherent environments for interactive exploration. However, their controllability often degrades under complex and compositional action sequences, and the underlying causes of this degradation remain largely unexplored. Both discrete-action-based models (Mao et al., 2025; He et al., 2025b) and continuous-action-based models (Li et al., 2025b; Sun et al., 2025; Shin et al., 2025; Gao et al., 2025) primarily treat action conditioning as a system-level design choice, with limited analysis of its representation-level implications.

3 Method

3.1 Overview

Our goal is to build a video world model that can reliably follow complex camera motion commands. Formally, given an initial frame x_0, a text prompt τ, and an action sequence a_{seq}, we aim to generate a video whose dynamics align with both semantic contents and specified actions. To this end, we adopt a pretrained image-to-video diffusion model (Wan et al., 2025) as the backbone, which synthesizes videos conditioned on x_0 and τ. The overall pipeline consists of three stages: (i) an **Action Discretization** stage that maps continuous camera trajectories $\mathbf{T}$ to discrete action sequences a_{seq} , (ii) an **Action-Controllable Teacher Model** that generates videos conditioned on a_{seq}, and (iii) an **Action-Guided Distillation** procedure that distills the teacher model into an efficient autoregressive student for long-horizon generation.

3.2 Action Discretization

Given a camera trajectory $\mathbf{T} = \{\mathbf{T}_t\}_{t=0}^{T}$, where each $\mathbf{T}_t$ is a 4×4 camera-to-world pose for the t-th frame, we map it into a discrete sequence a_{seq} with the following steps.

Temporal partitioning. We first divide $\mathbf{T}$ into a sequence of short temporal segments to capture locally coherent camera motions. The initial pose $\mathbf{T}_0$ is treated as a reference pose, and the remaining poses are partitioned into consecutive segments of fixed length $k = 4$. As a result, the trajectory is divided into N segments, with each summarizing camera motion over a short temporal interval and serving as the basic unit for action extraction.

Relative pose estimation. Let $\mathbf{T}_{k-1}$ and $\mathbf{T}_k$ denote the terminal poses of two adjacent segments, we then calculate their relative transformation to represent the camera action over the segment:

$$\Delta\mathbf{T} = \mathbf{T}_k \mathbf{T}_{k-1}^{-1} = \begin{bmatrix} \Delta\mathbf{R} & \Delta\mathbf{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{1}$$

where $\Delta\mathbf{T} \in SE(3)$ is a 4×4 homogeneous transformation matrix, $\Delta\mathbf{R} \in SO(3)$ is the rotation matrix and $\Delta\mathbf{t} = [dx, dy, dz]^\top$ is the translation vector.

Quantization and primitive mapping. We follow DeepVerse (Chen et al., 2025) to decompose $\Delta\mathbf{T}$ into two decoupled motion primitives through a threshold-based quantization process. Specifically, the translation vector $\Delta\mathbf{t} = [dx, dy, dz]^\top$ and rotation matrix $\Delta\mathbf{R}$ are separately mapped to a set of symbolic labels in the following manner:

- **Translation ($N_t = 9$ types):** we compute the horizontal displacement $d = \sqrt{dx^2 + dz^2}$. If d is below a predefined threshold τ_t, the motion is assigned a *Stationary* label s_0. Otherwise, the translation is categorized into one of eight directional bins based on the quantized polar angle $\phi = \arctan2\,(dx, dz)$, representing: *Forward* (F), *Backward* (B), *Left* (L), *Right* (R), and their diagonal counterparts (*fL, fR, bL, bR*).
- **Rotation ($N_r = 3$ types):** we extract the yaw angle ψ from $\Delta\mathbf{R}$. The rotation is mapped to a discrete state $r \in \{\text{N, L, R}\}$, where N represents a *Neutral* state ($|\psi| \leq \tau_r$), L (Left) and R (Right) represent counter-clockwise and clockwise rotations exceeding the threshold τ_r, respectively.

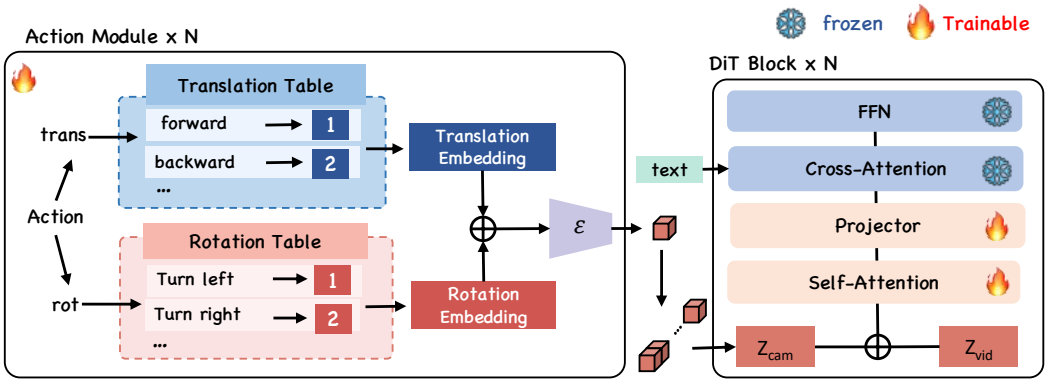

Figure 2: Architecture of the action-controllable teacher model. Each discretized action is decomposed into translation and rotation indices, which retrieve embeddings from separate lookup tables. The embeddings are projected and concatenated to form the action feature, and then fused with video latents via element-wise addition.

Combining these primitives, each camera action is represented as an action pair $a_k = (i_k, j_k)$, defined in a discrete action space $\mathcal{A}$ of size $9 \times 3 = 27$, where $i_k \in \{0, \ldots, 8\}$ and $j_k \in \{0, \ldots, 2\}$ index the corresponding translation category (e.g., S, F, fL) and rotation category (e.g., N, L, R) respectively. In this way, we convert the continuous camera trajectory $\mathbf{T}$ into a discrete action sequence $a_{seq} = \{a_1, a_2, \ldots, a_N\}$.

3.3 Action-Controllable Teacher Model

To supply our image-to-video backbone (Wan et al., 2025) with action controllability, we introduce an action module that maps a_{seq} into a high-dimensional latent space and then fuse it with video latents, as illustrated in Fig. 2.

Specifically, we construct two learnable embedding tables, $\mathbf{E}_{trans} \in \mathbb{R}^{9\times d}$ and $\mathbf{E}_{rot} \in \mathbb{R}^{3\times d}$ to obtain the action embedding $e_k \in \mathbb{R}^d$ based on $a_k = (i_k, j_k)$:

$$\mathbf{e}_k = \begin{cases} \mathbf{e}_{start}, & k = 0 \\ \text{Proj}\left(\mathbf{E}_{trans}\left[i_k\right] + \mathbf{E}_{rot}\left[j_k\right]\right), & k > 0 \end{cases} \tag{2}$$

where $\mathbf{e}_{start} \in \mathbb{R}^d$ is a learnable parameter representing the initial state. $\mathbf{E}[\cdot]$ denotes a lookup operation that retrieves the i-th or j-th row from the respective embedding tables, indexed by the action categories. $\text{Proj}(\cdot)$ is a linear projection layer, initialized as zeros to stabilize training by gradually introducing action-conditioned signals. The resulting action embedding can be represented as $e_{act} = \{e_0, e_1, \ldots, e_N\}$.

The action embedding e_{act} is integrated into the video features $\mathbf{f}_{vid} \in \mathbb{R}^{b\times(t\times h\times w)\times d}$ through a structured expansion and fusion process. Specifically, the sequence of action embeddings $\{\mathbf{e}_k\}_{k=0}^{t-1}$ is stacked along the temporal axis to form a temporal action representation $\mathbf{f}_{act}^{temp} \in \mathbb{R}^{b\times t\times d}$. To match the resolution of the video features, each temporal embedding is replicated $h \times w$ times and flattened to the token dimension, resulting in $\mathbf{f}_{act} \in \mathbb{R}^{b\times(t\times h\times w)\times d}$. This expanded action feature is combined with the video features via element-wise addition to get the fused feature:

$$\mathbf{f}_{fuse} = \mathbf{f}_{vid} + \mathbf{f}_{act}. \tag{3}$$

Following ReCamMaster (Bai et al., 2025a), $\mathbf{f}_{fuse}$ are further processed by a self-attention layer together with an additional linear projection layer, enabling the action-conditioned signals to modulate the video representation more effectively.

The model is then optimized via the following objective:

$$\mathcal{L} = \mathbb{E}_{z,\epsilon,t,\mathbf{f}_{act},\tau,x_0}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \mathbf{f}_{act}, \tau, x_0\right)\|^2\right] \tag{4}$$

where t denotes the diffusion timestep, τ is the text prompt, x_0 is the conditional frame, ϵ is the added noise, and ϵ_θ is the denoising network (Wan et al., 2025).

3.4 ACTION-GUIDED DISTILLATION

To enable efficient long video generation, we distill the teacher model $G_{teacher}$ into a four-step autoregressive student $G_{student}$ using self-forcing strategy (Huang et al., 2025b), which consists of two stages: ode initialization and DMD (Yin et al., 2024)-based training.

Attention pattern. During both training and inference, we adopt a truncated causal attention mechanism consisting of a fixed-size local window and a set of sink tokens(Yang et al., 2025a; Shin et al., 2025). Concretely, each token attends to (i) all preceding sink tokens and (ii) a sliding window of recent tokens. We denote the number of sink tokens as S and the local window size as W.

Ode initialization. We first construct a dataset $\mathcal{D}$ comprising ODE trajectories $\{\mathbf{z}_t^i\}_{i=1}^N$ generated by $G_{teacher}$, where the timesteps t are sampled from a 4-step subset of $[0, T]$.

Following the initialization protocol from CausVid (Yin et al., 2025), the input of N noisy latents is partitioned into L non-overlapping chunks $\{\mathbf{z}_{t^i}^i\}_{i=1}^L$ with independent timesteps, and we adopt a causal attention mask M that permits the attention pattern we use. Specifically, the mask $M \in \{0,1\}^{N\times N}$ is defined as:

$$M\left[m,n\right] = \begin{cases} 1, & n < S \ \wedge \ n \le m, \\ 1, & S \le n \le m \ \wedge \ n \ge m - W, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For each chunk i, the attention mechanism is constrained to M. The final optimizing objective can be formulated as:

$$\mathcal{L}_{\text{ode}} = \frac{1}{L}\sum_{i=1}^{L} \mathbb{E}_{\mathbf{z},t^i} \left\| G_{student}\left(\mathbf{z}_{t^i}^i, c^i, t^i, M\right) - \mathbf{z}_0^i \right\|^2 \tag{6}$$

where c^i includes the text τ and action conditions a_i.

DMD-based self-forcing training. Subsequently, we perform self-forcing style distillation leveraging distribution matching distillation (DMD) (Yin et al., 2024). Unlike traditional teacher-forcing, $G_{student}$ autoregressively produces video chunks $\{\mathbf{z}_0^i\}_{i=1}^L$ conditioned on its own previous outputs, mitigating the training-inference gap. To balance long-range stability and computational efficiency, we implement a rolling KV-cache mechanism with an anchored attention sink. Specifically, for each chunk i, the generative process is conditioned on the corresponding action signals a_i, text condition τ, and contexts $\mathcal{C}_i$:

$$\mathcal{C}_i = \{\mathbf{z_0^j} \mid 0 \le j < S \text{ or } \max\left(0, i - W\right) \le j < i\} \cup \{\mathbf{z_t^i}\} \tag{7}$$

The generator is optimized by minimizing the approximate KL-divergence between the synthesized distribution p_θ and the real data distribution p_{real}. An auxiliary distribution p_{fake} is utilized to estimate p_θ, and the gradient for the generator parameters θ can be formulated as:

$$\nabla_\theta \mathcal{L}_{\text{DMD}} = -\mathbb{E}_t\left(\nabla_\theta \text{KL}\left(p_{\text{fake},t} \| p_{\text{real},t}\right)\right) \tag{8}$$

4 EXPERIMENTS

Training data. To enable our model to handle scenarios with highly dynamic contents, we select around 10k high-quality clips from OmniWorld (Zhou et al., 2025), and re-annotate their camera poses with VIPE (Huang et al., 2025a). However, We observe a long-tail distribution in action types within this subset, where forward-dominant sequences account for the majority of samples. To relieve this imbalance and improve generalization, we further augment it with synthetic data from a UE simulator (Yu et al., 2025a). The final dataset comprises approximately 20k videos of 81 frames, featuring a rich variety of camera motions.

Implementation Details. We build DisCo based on the Wan2.2-5B image-to-video generation model (Wan et al., 2025). For action discretization, the translation threshold τ_t is 0.1 and the rotation threshold τ_r is 1. For the ode initialization, we use $G_{teacher}$ to generate 4k ode pairs with

a balanced action distribution as our training data. In the self-forcing distillation stage, $G_{student}$ generates videos chunk-wisely, it takes the first conditional image as the first chunk and splits the remaining video with the chunk size=4. In both ode initialization and self-forcing stage, the sink size is 1 and the local attention window size is 8. More details are in Appendix A

DisCoBench. To comprehensively evaluate our model, we introduce DisCoBench, which consists of 3 subsets to assess generation capabilities from different perspectives:

- **DisCo-ActionBench**: This subset contains 300 initial images from diverse scenes paired with predefined complex action sequences. Each sequence contains at least two kinds of actions from our unified action set $\mathcal{A}$. This subset evaluates the controllability and short video generation ability of the model. More details can be found in Appendix.C.
- **DisCo-DynamicBench**: This subset includes 200 dynamic third-person exploration videos to specifically assess the ability of models to generate videos with highly dynamic elements.
- **DisCo-LongBench**: This subset is an extension of the DisCo-ActionBench with prolonged action sequences to evaluate long-horizon stability. Specifically, we extend each action for 3 times, supporting evaluation for long videos with at least 241 frames generation

4.1 Evaluation on DisCo-Bench

Evaluation on DisCo-ActionBench. In this section, we mainly evaluate the action-following performance and the visual-temporal quality of the generated videos on DisCo-ActionBench. For the evaluation of video quaility, we adopt the widely used VBench++ protocol (Huang et al., 2025c), including Subject Consistency, Background Consistency, Motion Smoothness, Aesthetic Quality, and Imaging Quality. For the evaluation of action following ability, we use a carefully designed metric, *i.e.*, Action Score. Specifically, we utilize VIPE (Huang et al., 2025a) to obtain the c2w poses from the generated videos, and then transform the resulting camera trajectories into discrete action sequences with the discretization approach described in 3.2. The Action Score is defined as the proportion of actions that exactly match the ground-truth actions across all samples.

We compare our model against two categories of baselines: (1) Continuous-actions-based models: CameraCtrl (He et al., 2025a), MotionCtrl (Wang et al., 2024b), WanX-Cam (Wan et al., 2025), and Hunyuan-GameCraft Li et al. (2025b). The CameraCtrl and MotionCtrl adopts the image-to-video SVD version. WanX-Cam corresponds to the implementation of VideoX-Fun. (2) Discrete action based models: Matrix-Game2(He et al., 2025b). As shown in Tab.1, our method achieves the highest action score while maintaining superior visual and temporal quality.

Table 1: Comparison on DisCo-ActionBench. **Bold** and underlined indicate the best and second-best results respectively.

	Model	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Action Score
Continuous	CameraCtrl	0.8265	0.9202	0.9604	0.4536	0.5358	0.5731
	MotionCtrl	<u>0.8709</u>	0.9294	0.9797	0.4608	0.5618	0.3636
	WanX-Cam	0.8139	0.8977	0.9764	0.5070	0.6351	0.5143
	GameCraft	0.8559	0.9169	0.9757	0.5074	**0.6792**	0.2736
Discrete	Matrix-Game2	0.8538	<u>0.9333</u>	<u>0.9824</u>	<u>0.5148</u>	<u>0.6474</u>	<u>0.5880</u>
	Ours	**0.8869**	**0.9353**	**0.9842**	**0.5308**	0.6420	**0.6332**

Evaluation on DisCo-DynamicBench. The ability to generate exploration videos with dynamic elements is crucial for interactive experiences. To evaluate this capability, we construct DisCo-DynamicBench, a benchmark tailored for assessing dynamic exploration video generation. We adopt two metrics for evaluation: Fréchet Video Distance (FVD) (Unterthiner et al., 2018) and VQA-Score (Lin et al., 2024). FVD measures the distributional similarity between generated videos and reference exploration videos, where a lower score indicates closer alignment with the target video distribution. VQA-Score evaluates whether the generated videos correctly follow the textual prompts. For instance, if a prompt requires a character to perform a running action but the generated video only exhibits camera motion with a static character, the resulting VQA-Score will be penalized. We use Qwen2.5-VL-7B (Bai et al., 2025b) as the underlying vision-language model for

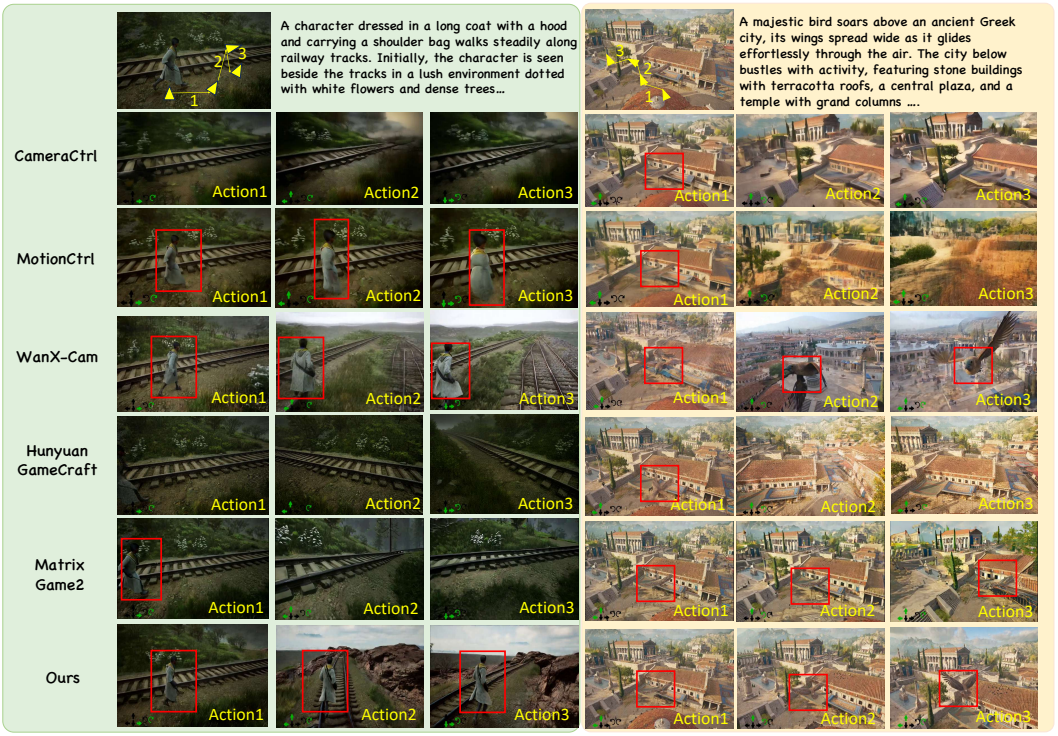

Figure 3: Comparison with baselines on complicated camera actions. The red box means the dynamic element in the generated video. DisCo is the only model with both great controllability and the ability to generate videos with dynamic elements.

VQA scoring. The results are illustrated in Tab. 2. DisCo achieves the best FVD score and competitive VQA-Score performance, demonstrating its effectiveness in generating exploration videos with dynamic elements. More details on the effectiveness of VQA-Score are provided in Appendix C.

Table 2: Comparison on the **DisCo-DynamicBench**

Model	FVD	VQAScore
CameraCtrl	470.22	0.3321
MotionCtrl	481.19	0.4126
WanX-Cam	<u>285.95</u>	**0.5637**
GameCraft	374.06	0.4033
Matrix-Game2	380.28	0.3884
Ours	**220.51**	<u>0.5448</u>

Evaluation on DisCo-LongBench. To evaluate the long video generation ability of DisCo, we compare it with two controllable long video generation models: Matrix-Game2 (He et al., 2025b) and Hunyuan-GameCraft (Li et al., 2025b). As shown in Tab.3, DisCo outperforms in consistency, motion smoothness and aesthetic quality, while achieving competitive imaging quality.

Table 3: Comparison with relative baselines on the **DisCo-LongBench**

Model	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality
GameCraft	0.7615	0.8842	0.9740	0.4758	0.6441
Matrix-Game2	0.7593	0.8969	0.9814	0.4883	**0.6474**
Ours	**0.8138**	**0.9103**	**0.9846**	**0.4938**	0.6372

Qualitative Comparison. As shown in Fig. 3, MotionCtrl and Hunyuan-GameCraft struggle with both action following and dynamic content generation. CameraCtrl and Matrix-Game2 follow complex action commands but fails to generate the required dynamic elements. WanX-cam exhibits noticeable visual inconsistencies. Comparatively, our method achieves accurate action following while maintaining stable generation of dynamic elements. More demos are shown in Appendix F.

4.2 Continuous versus Discrete Actions

As discussed in Sec.1, limited separability of action features leads to ambiguous guidance and degraded controllability in continuous-action-based models, whereas carefully designed discrete action features alleviate this issue. In this subsection, we empirically validate this by analyzing the similarity of action features and the resulting controllability.

For a direct analysis of action features, we focus on adapter-based models that explicitly extract action features before fusion with visual features. Based on this criterion, we select representative continuous-action-based baselines, including CameraCtrl (He et al., 2025a), WanX-Cam (Wan et al., 2025), and Hunyuan-Gamecraft (Li et al., 2025b) for further comparison.

Action feature similarity. To assess whether our method produces more discriminative conditioning features, we analyze the cosine similarity of action features across different motion patterns. Specifically, we define the similarity score as the average pairwise cosine similarity:

$$\mathcal{S} = \frac{1}{N(N-1)} \sum_{i \neq j} \frac{\mathbf{f}_i^\top \mathbf{f}_j}{\|\mathbf{f}_i\|_2 \, \|\mathbf{f}_j\|_2}. \tag{9}$$

where $\mathbf{f}_i$ denotes the action feature extracted from the action encoders. Each action pattern is sampled from the unified action set $\mathcal{A}$ mentioned in 3.2. For models conditioned on continuous camera controls, c2w matrices are used as conditioning inputs, of which different transformations correspond to different discretized actions in $\mathcal{A}$. As illustrated in Fig. 4, our method produces action features with more regular and interpretable similarity structures compared to continuous-action-based methods. Specifically, actions sharing the same rotation pattern form coherent clusters, and actions with the same translation pattern also exhibit certain similarity. In contrast, continuous-action-based models produce highly entangled and less interpretable feature representations, resulting in uniformly higher similarity scores, as shown in Tab. 4.

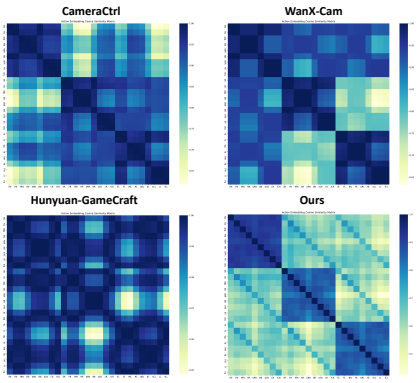

Figure 4: Heatmaps of action feature similarity for different adapter-based models. Darker colors indicate higher similarity.

Controllability evaluation. As illustrated in Tab.4, Our Method achieves the best action score, while continuous-action-based models with entangled representations exhibit weaker controllability. These findings provide empirical support for our claim that preserving sufficient action separability in the conditioning feature space is critical for reliable action control. We further substantiate this with ablation studies in Sec. 4.3.

Table 4: Comparison with adapter-based models using continuous action condition

Model	Similarity Score	Action Score
CameraCtrl	0.8826	0.5731
WanX-Cam	0.8953	0.5143
GameCraft	0.9913	0.2736
Ours	**0.7029**	**0.6332**

Table 5: Ablation on continuous and discrete representations

Action representation	Similarity Score	Action Score
continuous (poses)	0.9377	0.5245
continuous (Plücker)	0.9201	0.5563
discrete (ours)	**0.7029**	**0.6332**

4.3 Ablations

Discrete Action Representation. As analyzed in Sec. 4.2, continuous action representations are associated with higher action feature similarity and degraded controllability. To isolate the effect of action representation from other factors, we conduct an ablation by replacing the discrete action embeddings in DisCo with directly encoding camera-to-world matrices and continuous Plücker embeddings derived from poses, while keeping all other architectures, training settings unchanged.

As shown in Tab. 5, conditioning on continuous poses or Plücker embeddings results in significantly higher feature similarity and lower action-following accuracy compared to the discrete representation. This strictly controlled ablation provides direct evidence that preserving action separability in the conditioning feature space is critical for effective action control, and suggests that the gains stem from the intrinsic difference between continuous and discrete action representations rather than architectural or training differences.

Attention Sink. The attention sink is designed to enhance consistency in long video generation. In our setting, we adopt a sink size of 1(take the first frame as the global sink), which aligns with the image-conditioning paradigm used in image-to-video generation. To further investigate the impact of the attention sink, we conduct an ablation study by training models without the attention sink, and evaluate their consistency performance on DisCo-LongBench. The results are summarized in Tab. 6, with qualitative comparisons shown in Fig. 5. Both demonstrate that introducing an attention sink improves the long-horizon consistency.

Table 6: Ablation on attention sink, test on the DisCo-LongBench

w/wo sink	Subject Consistency	Background Consistency
w/o sink	0.8005	0.9037
w 1 sink	**0.8138**	**0.9103**



Figure 5: Ablation on long video generation with and without attention sink.

Data Ablation. We further conduct ablation experiments to analyze the contribution of different components of our training data. By default, we train DisCo on a 1:1 mixture of motion-intensive exploration videos (D) and camera-stabilized synthetic scenes (S); the former features complex actions and dynamic elements, while the latter prioritizes smooth motion and static environment. To isolate their respective effects, we consider two alternative settings: (i) training exclusively on D, and (ii) training solely on S.

As shown in Tab. 7, training with D enables the model to generate highly dynamic exploration videos. However, the long-tail distribution of its camera actions and the high complexity of such data make it challenging for the model to reliably learn the correct motion patterns. In contrast, S provides more structured and balanced motion supervision, and substantially improves action-following performance. Nevertheless, in setting (ii), the model exhibits degraded FVD and VQA-Score, indicating limited capability in generating videos with rich dynamic elements. These results indicate that combining D and S is crucial for achieving both action controllability and dynamic scene generation.

Table 7: Ablation on Dynamic and synthetic ratio

D:S	FVD	VQAScore	Action Score
1:0	**198.96**	**0.5779**	0.4828
0:1	699.83	0.5237	<u>0.6005</u>
1:1(ours)	<u>220.51</u>	<u>0.5448</u>	**0.6332**

5 Conclusion

This paper presents DisCo, a video world model with discrete camera motion control. By discretizing camera trajectories into a compact set of predefined action types, DisCo alleviates action representation entanglement, enabling more reliable action following under complex and compositional motion sequences. Combined with self-forcing distillation and attention sinks, DisCo supports efficient controllable long video generation. To facilitate comprehensive evaluation of model performance, we also introduce DisCoBench, which assesses complementary aspects of controllable video generation with different subsets. Extensive experiments demonstrate that DisCo consistently outperforms prior methods in action controllability, while preserving visual consistency and temporal coherence in dynamic and long-horizon video generation.

AI USE STATEMENT

We used generative AI tools for language editing, proofreading, and code debugging. All AI-assisted content was reviewed and verified by the authors.

ETHICS STATEMENT

This work involves a human evaluation of generated videos. Participants only provided subjective preferences, and no personally identifiable information was collected. The study poses no known ethical risks beyond those associated with standard visual evaluation.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of our method, experimental settings, and evaluation protocols in the main paper and appendix. We will release the source code, benchmark and relevant evaluation scripts to facilitate reproducibility, subject to the licensing requirements of the underlying models and datasets.

REFERENCES

Sherwin Bahmani, Ivan Skorokhodov, Guocheng Qian, Aliaksandr Siarohin, Willi Menapace, Andrea Tagliasacchi, David B Lindell, and Sergey Tulyakov. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers. In *CVPR*, 2025.

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. *arXiv preprint arXiv:2503.11647*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *CVPR*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NeurIPS*, 2024.

Junyi Chen, Haoyi Zhu, Xianglong He, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Yang Zhou, Zizun Li, Zhoujie Fu, Jiangmiao Pang, et al. Deepverse: 4d autoregressive video generation as a world model. *arXiv preprint arXiv:2506.01103*, 2025.

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Junhao Zhuang, Chengming Xu, Jianfeng Feng, Yu Qiao, Yanwei Fu, Chenyang Si, and Ziwei Liu. Longvie 2: Multimodal controllable ultra-long video world model. *arXiv preprint arXiv:2512.13604*, 2025.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, and Jun Xiao. Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models. *arXiv preprint arXiv:2406.10981*, 2024.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025.

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

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for video diffusion models. In *ICLR*, 2025a.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025b.

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

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025b.

Ziqi Huang, Fan Zhang, Xiaojie Xu, Yinan He, Jiashuo Yu, Ziyue Dong, Qianli Ma, Nattapol Chanpaisit, Chenyang Si, Yuming Jiang, et al. Vbench++: Comprehensive and versatile benchmark suite for video generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025c.

Zhennan Jiang, Kai Liu, Yuxin Qin, Shuai Tian, Yupeng Zheng, Mingcai Zhou, Chao Yu, Haoran Li, and Dongbin Zhao. World4rl: Diffusion world models for policy refinement with reinforcement learning for robotic manipulation. *arXiv preprint arXiv:2509.19080*, 2025.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*, 2024.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Pathdreamer: A world model for indoor navigation. In *CVPR*, 2021.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

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.

Guangyuan Li, Siming Zheng, Shuolin Xu, Jinwei Chen, Bo Li, Xiaobin Hu, Lei Zhao, and Peng-Tao Jiang. Magicworld: Interactive geometry-driven video world exploration. *arXiv preprint arXiv:2511.18886*, 2025a.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2025b.

Quanhao Li, Zhen Xing, Rui Wang, Hui Zhang, Qi Dai, and Zuxuan Wu. Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. In *ICCV*, 2025c.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025d.

Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. *arXiv preprint arXiv:2410.20502*, 2024.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *ECCV*, 2024.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *arXiv preprint arXiv:2509.25161*, 2025.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. *arXiv preprint arXiv:2512.04678*, 2025.

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*, 2025.

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 *CVPR*, 2025.

Joonghyuk Shin, Zhengqi Li, Richard Zhang, Jun-Yan Zhu, Jaesik Park, Eli Shechtman, and Xun Huang. Motionstream: Real-time video generation with interactive motion controls. *arXiv preprint arXiv:2511.01266*, 2025.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

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.

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:1812.01717*, 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.

Junke Wang, Yi Jiang, Zehuan Yuan, Binyue Peng, Zuxuan Wu, and Yu-Gang Jiang. Omnitokenizer: A joint image-video tokenizer for visual generation. In *NeurIPS*, 2024a.

Junke Wang, Zhi Tian, Xun Wang, Xinyu Zhang, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Simplear: Pushing the frontier of autoregressive visual generation through pretraining, sft, and rl. *arXiv preprint arXiv:2504.11455*, 2025a.

Junke Wang, Xun Wang, Qiushan Guo, Peize Sun, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Omnigen-ar: Autoregressive any-to-image generation. In *NeurIPS*, 2025b.

Zhouxia Wang, Ziyang Yuan, Xintao Wang, Yaowei Li, Tianshui Chen, Menghan Xia, Ping Luo, and Ying Shan. Motionctrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024b.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021.

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*, 2025a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025b.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *NeurIPS*, 2024.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, 2025.

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*, pp. 1–11, 2025a.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. *arXiv preprint arXiv:2501.08325*, 2025b.

David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. In *CVPR*, 2025.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

A More Experimental Details

All training stages are conducted on 16 NVIDIA H20 GPUs using the AdamW (Kingma & Ba, 2014) optimizer.

Teacher Model. The teacher model is trained for approximately 10k steps with a batch size of 64 and a learning rate of 1×10^{-5}.

ODE Initialization Stage. The ODE initialization stage is trained for approximately 8k steps with a batch size of 64 and a learning rate of 2×10^{-6}.

Self-Forcing DMD Distillation. In this stage, we use a batch size of 32 and train for approximately 2k steps. The learning rates are set to 2×10^{-6} for the generator and 4×10^{-7} for the critic (i.e., the fake score function), with an update ratio of 1:5 between the generator and the critic.

B Clarifications

B.1 Clarification on contributions.

Our contribution lies not only in discretization-based formulation, and proposed benchmark, but also in a deeper representation level insight that may influence and guide the future work. Specifically, continuous action representations can lead to feature entanglement where different motion patterns may exhibit highly similar representations and thus degrade controllability, while the proposed discretization-based embedding and fusion strategy effectively mitigates this issue. To the best of our knowledge, this hasn't been systematically studied in prior controllable video generation works.

B.2 Clarification on action categories

Currently, our action space does not include some camera motions such as pitch and roll. We acknowledge this limitation. Supporting these additional motion categories requires collecting sufficiently diverse training trajectories, which is more challenging in practice, making it difficult to adequately train reliable embeddings for such actions with our current data. Nevertheless, our framework is readily extensible: additional action categories can be incorporated by introducing

corresponding action embeddings. We therefore leave extending the action space to cover a broader range of camera motions, including pitch and roll, as future work.

Importantly, this limitation does not affect the core finding of this work: discrete camera-motion representations can reduce feature similarity between different actions and thereby improve controllability compared with continuous representations. To isolate the effect of the action representation, we conduct strictly controlled experiments using the same backbone, training data, training settings, and computational resources, while varying only the camera-motion representation. The results in Tab. 5 provide direct evidence supporting this finding.

B.3 Clarification on the motion-magnitude information.

Our current action space does not explicitly distinguish different motion magnitudes. This is a deliberate design choice rather than an inherent limitation of the discrete action space, which can be naturally extended by introducing additional discrete action primitives with motion-magnitude information, for example, adding separate action embeddings for different movement speeds or rotation angles. In this work, we intentionally adopt a compact action space to ensure sufficient training samples for each action category.

B.4 Clarification on the main comparison

Table 1 presents a system-level comparison. The corresponding backbone, frame count, and spatial resolution of each method are summarized in Table 8. Specifically, compared with WanX-Cam, our method achieves higher video quality and stronger controllability while using the same backbone. Moreover, our method also outperforms GameCraft despite using a smaller backbone. We acknowledge that there are differences in training data, training resources, and other implementation details across methods. However, such differences are common in controllable video generation works, where obtaining perfectly matched training and computational setups is often impractical. We therefore focus on the system-level performance under the respective official settings of each method.

To address this issue, we further conduct strictly controlled ablations using the same Wan2.2 backbone, training data, training settings, and computational resources, while varying only the representation of camera motion. The results, reported in Tab. 5 of the main paper, further support the effectiveness of our method and the benefits of discrete camera-motion representations.

Table 8: Baseline settings.

Model	Backbone	Resolution	Frame count
CameraCtrl	SVD	576x320	14
MotionCtrl	SVD	1024x576	14
WanX-Cam	Wan(5B)	1280x704	81
GameCraft	Hunyuan-Video(13B)	1216x704	99
Matrix-Game-v2	SkyReel(1.3B)	640x352	81
DisCo	Wan(5B)	1280x704	81

C Details on DisCoBench and Evaluation

Source of the images. Images in DisCo-ActionBench and DisCo-LongBench come from game scenes, AI-Generated images and real-world images. Videos in DisCo-DynamicBench source from game videos, movie clips and real-world exploration videos.

Design of the DisCo-ActionBench. DisCo-ActionBench is constructed to evaluate models under a wide range of complicated camera action sequences rather than repeated instances of a small number of simple trajectories. Each sample is associated with a structured camera motion sequence composed of discrete motion primitives from our predefined action set $\mathcal{A}$, and contains at least two types of actions. Each camera action sequence can be formulated as $seq = (a_1 \times N_1)(a_2 \times N_2)...$ where $a_i \in \mathcal{A}$ and N_i represents the lasting number of action a_i. To avoid changing actions too quickly, we set $Ni \geq 4$.

We also provide visualizations to present the per-action distribution and the trajectories of the DisCo-ActionBench. As illustrated in Fig.6 and Fig.7, DisCo-ActionBench covers a wide range of action sequences structures, offering a strong evaluation suite for the assessment of the controllability of models.

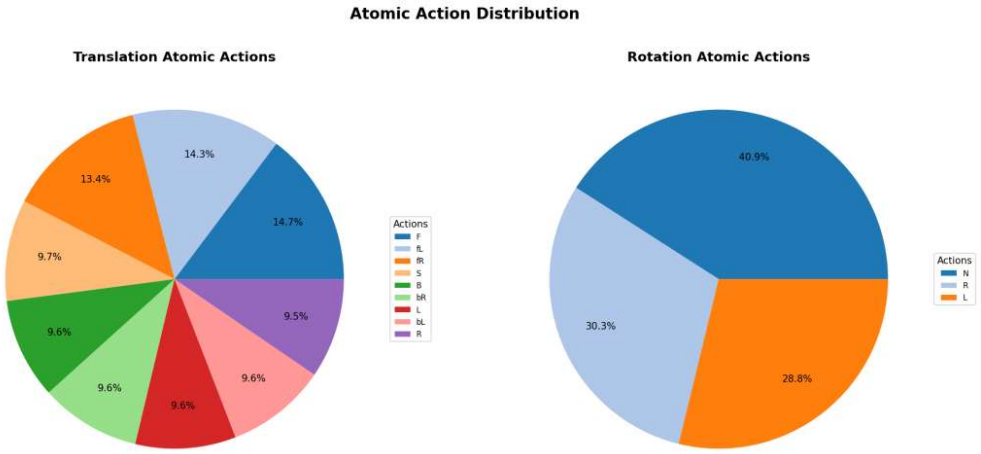

Figure 6: Per-action distribution of DisCo-ActionBench.

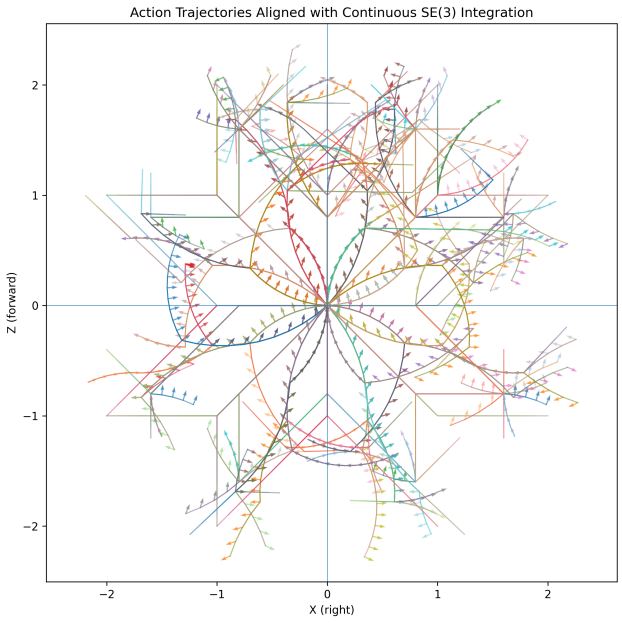

Figure 7: Each polyline corresponds to a unique camera motion sequence in the DisCo-ActionBench. All trajectories start from the same origin, and are constructed by sequentially applying discrete camera motion primitives. The visualization highlights the diversity of trajectory structures and directional coverage.

Moreover, the benchmark data underwent manual human quality inspection, requiring over 30 human$*$hours of verification. We manually inspected every sample to ensure that (1) the initial

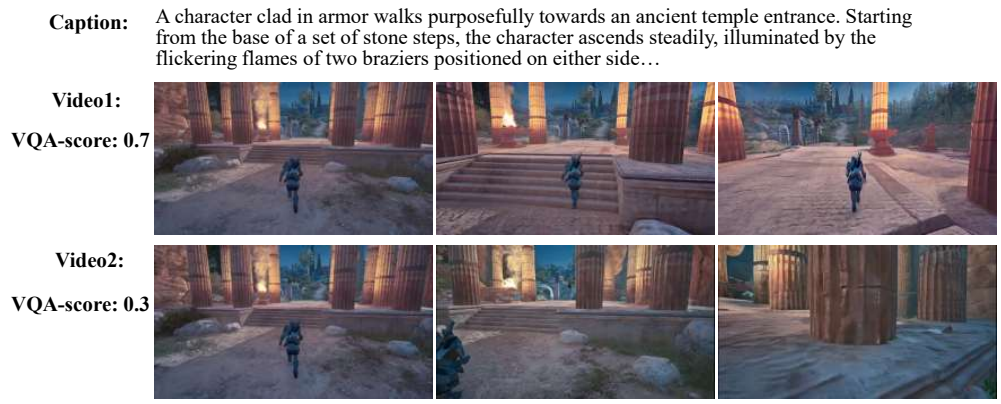

Figure 8: Comparison of two generated videos under the same caption with their corresponding VQA-Scores. Video1 aligns well with the caption by exhibiting the required dynamic content and achieves a higher VQA-Score, whereas video2 mainly contains camera motion in a static scene and receives a significantly lower score.

frame contains a clear and visually recognizable scene/foreground object that suitable for controllable generation, (2) the predefined action sequences jointly cover all action types with a balanced distribution, and each sequence contains at least two action types. During this process, we use the per-action distribution as a quantitative indicator to analyze benchmark balance and adjust the sampled sequences accordingly to ensure that all action types were represented in a balanced manner.

Effectiveness of VQA-Score in DisCo-DynamicBench. In this section, we provide additional details to demonstrate the effectiveness of the VQA-Score used in DisCo-DynamicBench. Since most samples in DisCo-DynamicBench are sourced from OmniWorld-Game (Zhou et al., 2025), we directly adopt the captions provided by the dataset for VQA-Score calculation. For the remaining samples, we follow the same caption generation protocol as OmniWorld-Game to produce corresponding textual descriptions.

As shown in Fig.8, since the captions explicitly describe the dynamic content expected in the generated videos, VQA-Score is able to effectively assess whether the generated videos contain the required dynamic elements, rather than merely exhibiting camera motion in static scenes.

D More Experiments and Discussions

D.1 Sensitivity Analysis of Action Score

To examine the sensitivity of the action score, we vary the translation threshold τ_t and rotation threshold τ_r and report the results in Figure 9.

We find that the action score remains stable across a reasonably broad range of thresholds: it stays above 0.61 for $\tau_r \in [0.6, 1.3]$ and above 0.62 for $\tau_t \in [0.05, 0.12]$. This demonstrates the robustness of the metric and the validity of our chosen settings.

D.2 Trade-off between Discretization and Fine-grained Control

Discretizing continuous trajectories into 27 bins inevitably introduces some fine-grained information loss. However, continuous camera trajectories, though fully fine-grained, often lead to feature entanglement and degraded controllability. Our discretized actions instead trade off sub-bin precision for a more structured and distinguishable action space, which we validate is fundamental to fine-grained control as complex motions can only be modeled reliably when different actions can be represented in a sufficiently separable manner.

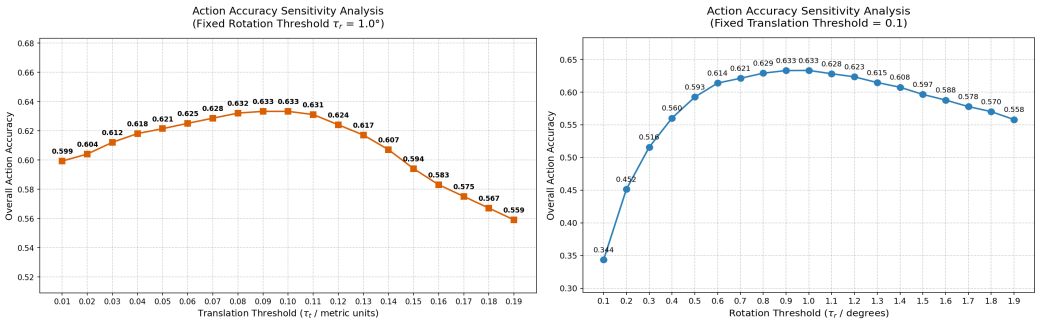

Figure 9: Sensitivity Analysis of Action Score

We further support this claim with an additional ablation on the number of bins. As shown in Table 9, increasing the number of bins indeed provides a finer-grained action space, but it also leads to worse Action Score. This result further validates our argument that stronger expressiveness alone does not necessarily improve controllability; when the action space becomes overly fine-grained and less distinguishable, control fidelity can instead degrade.

Table 9: Effect of number of bins.

Number of Bins	Action Score
27 (9×3)	**0.6332**
39 (13×3)	0.5638
48 (16×3)	0.5073

D.3 Human Evaluation

We further conduct an extensive human evaluation to assess the ability of our method to generate videos with the intended dynamic content, as well as its action-following and motion smoothness. We compare our method with three representative baselines, Matrix-Game2, WanX-Cam, and Hunyuan-GameCraft, using 30 video pairs per baseline. In total, 30 human evaluators participate in the evaluation and are asked to select the video that better matches the intended dynamic content while maintaining visual quality. We report the preference rate together with Fleiss' Kappa to measure inter-rater agreement.

Importantly, for the evaluation on controllability, our human evaluation is designed to assess on the sequence-level, rather than the correctness of isolated individual actions. In our setting, each evaluation sequence contains multiple action types, and action-following errors predominantly occur around transitions between different actions. Judging each individual action independently would therefore not faithfully reflect the actual controllability of the generated sequence. Moreover, isolated actions such as a single forward or left/right motion can be visually ambiguous, whereas deviations become much more apparent when the model needs to transition between different motion patterns. Therefore, we intentionally ask evaluators to judge the overall correspondence between the generated video and the complete target action sequence.

The results are summarized in Table 10. For dynamic content generation, our method is consistently preferred over all three baselines, with preference rates of 74.7%, 63.6%, and 70.2% over Matrix-Game2, WanX-Cam, and Hunyuan-GameCraft, respectively. Our method also achieves higher preference rates in both action following and motion smoothness. The corresponding Fleiss' Kappa values range from 0.41 to 0.64, indicating moderate to substantial agreement among evaluators. These human evaluation results are consistent with our VQA-Score evaluation and provide additional evidence that our method can generate more faithful dynamic content while preserving visual quality.

Overall, the human evaluation consistently favors our method across dynamic content generation, action following, and motion smoothness, further supporting the effectiveness of our approach for controllable video generation.

Table 10: Human preference results across different evaluation criteria. The preference rates indicate the percentage of evaluators preferring our method or the corresponding baseline.

Criterion	Compared Baseline	Preference for Ours	Preference for Baseline	Fleiss' Kappa
Dynamic Content Generation	Matrix-Game2	**74.7%**	25.3%	0.64
	WanX-Cam	**63.6%**	36.4%	0.43
	Hunyuan-GameCraft	**70.2%**	29.8%	0.59
Action Following	Matrix-Game2	**58.8%**	41.2%	0.45
	WanX-Cam	**61.7%**	38.3%	0.41
	Hunyuan-GameCraft	**76.1%**	23.9%	0.62
Motion Smoothness	Matrix-Game2	**73.4%**	26.6%	0.46
	WanX-Cam	**69.0%**	31.0%	0.48
	Hunyuan-GameCraft	**78.7%**	22.3%	0.53

D.4 Generalization Ability

To evaluate the generalization capability of DisCo, we randomly select 100 videos from the external Sekai (Li et al., 2025d) dataset, and compare our method with Matrix-Game2. As shown in Table 11, our model also exhibits competitive performance on unseen real-world data.

Table 11: Performance on unseen real-world data.

Method	FVD ↓	Action Score ↑
Ours	**333.93**	**0.602**
Matrix-Game2	369.10	0.567

D.5 Multi-seed Results

To further validate that our gain in controllability is reproducible rather than due to randomness, we evaluate our model over four independent runs with different random seeds on DisCo-ActionBench. As shown in Table 12, the Action Score remains stable with low standard deviations, which reflects genuinely better controllability.

Table 12: Multi-seed results.

Method	score-r1	score-r2	score-r3	score-r4	Std.
Ours	**0.6332**	**0.6315**	**0.6253**	**0.6357**	**0.0044**
Matrix-Game2	0.5880	0.5863	0.5912	0.5800	0.0047

D.6 Evaluate on independent measures

To evaluate the controllability by measures independent from the discretization pipeline, we further report the Relative Pose Error -Translation (RPE-T) and Relative Pose Error- Rotation (RPE-R) adpted from CameraCtrl (He et al., 2025a) on 100 real-world videos used in Appendix.D.4. The result is illustrated in the table. 13, which demonstrates that our model still outperforms baselines in metrics that are independent from the discretization rule.

D.7 Ablation on the attention window size

We further conduct an ablation on the attention window size using DisCo-LongBench. The results shown in Tab.14 indicate that enlarging the attention window provides no measurable improvement in long-horizon consistency.

Table 13: Performance on independent measures

Method	RPE-T	RPE-R
CameraCtrl	0.252	0.748
MotionCtrl	0.315	3.397
GameCraft	0.285	2.682
WanX-Cam	0.065	0.394
Matrix-Game2	0.092	0.544
ours	**0.054**	**0.326**

Table 14: Ablation on the attention window size

window size	Subject Consistency	Background Consistency
size = 8(ours)	0.8138	0.9103
size = 16	0.8132	0.9093

We acknowledge that such a local attention design is not optimal for extremely challenging cases requiring explicit global memory, such as 360-degree rotations or trajectories with loop closure. In these scenarios, consistency may gradually degrade due to the limited accessible history.

D.8 Compare with hybrid discrete-plus-residual baseline

We further compare our method with a hybrid discrete-plus-residual baseline: Hunyuan-WorldPlay (Sun et al., 2025), on the DisCo-ActionBench. The results are shown in Table 15. Our method still exhibits superior controllability

Table 15: Compare with hybrid discrete-plus-residual baseline.

Method	Action Score
WorldPlay	0.5427
Ours	0.6332

D.9 Runtime Characteristics

The runtime characteristics of our method and the comparison with self-forcing are reported in Table 16.Our method is more resource expensive mainly for its larger backbone and higher supporting resolution. This evaluation is conducted on a single 48G 4090.

Table 16: Efficiency comparison.

Method	Frame-rate	Latency	Resolution	Memory
Ours (5B)	16	23.4s	1280×704	33G
Self-forcing (1.3B)	16	13.8s	832×480	16G

E Limitation

Although DisCo covers most common actions for user interactions, including basic translational movements, diagonal movements, yaw rotations, and their combinations, some actions such as pitch, roll, and variable velocities are currently not supported. This limitation stems mainly from the fact that constructing large-scale, action-balanced datasets that cover all types of action patterns is highly resource-intensive. We leave this to future work. Another limitation arises in applications that

require very precise camera placement, as the discretization process inevitably sacrifices some fine-grained information and expressiveness. However, we observe that overly fine-grained action spaces (e.g., continuous actions) lead to feature entanglement and degraded controllability. Therefore, we trade off sub-bin precision for improved controllability.

F DEMOS

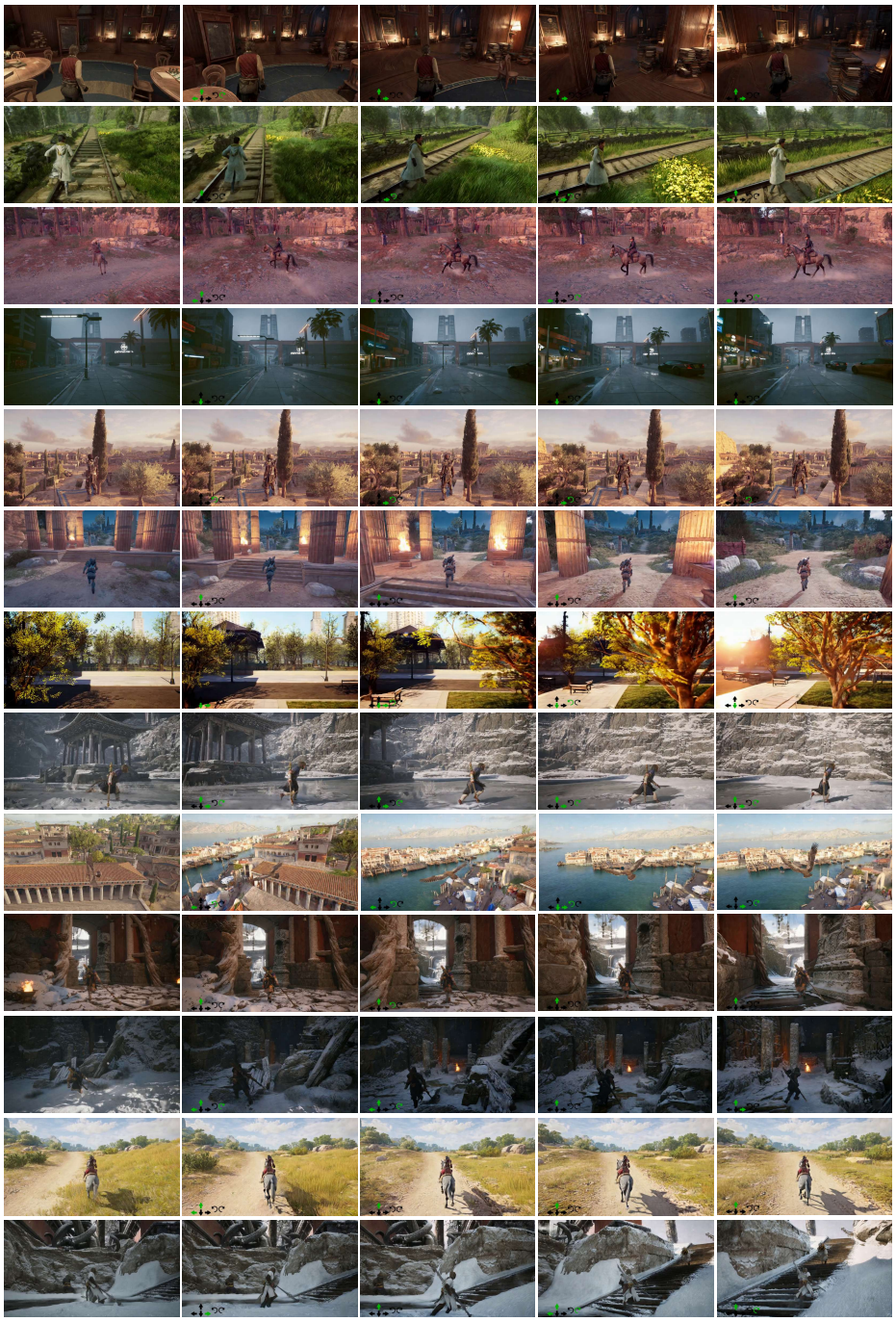

Figure 10: Demos