A DISCRETE WORLD EXPLORATION MODEL WITH SUPERIOR ACTION ALIGNMENT

Anonymous authors
Paper under double-blind review

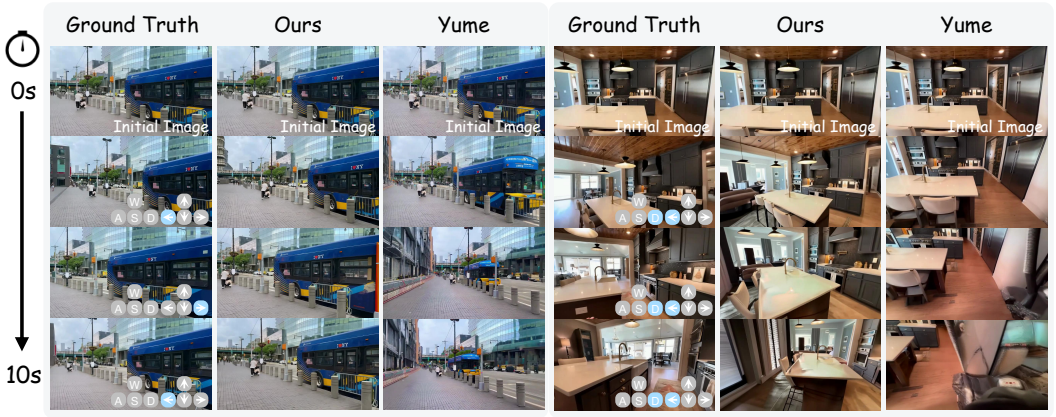


Figure 1: Compared with Yume Mao et al. (2025), a state-of-the-art diffusion-based world model, our method *DiscreteWorld* follows the actions more accurately while perserving sharper visual quality over 10-second generation. In contrast, Yume exhibits action misalignment (like directional mismatch) and long-horizon degradation. 'WASD' denotes 'forward/left/backward/right' translation controls respectively, and 'arrows' denote camera rotations.

ABSTRACT

As a fundamental world-modeling task, world exploration requires accurate action alignment and stable long-horizon generation. Existing approaches are predominantly diffusion-based but still exhibit limitations, particularly in accurate action alignment. Moving beyond diffusion, we introduce *DiscreteWorld*, a discrete generative model that produces stable 10-second, 480p, camera-controllable open-world videos with competitive action controllability. A key design of *DiscreteWorld* is to quantize both camera actions and visual observations into a unified discrete token space for joint reasoning and generation, yielding two essential advantages. First, unified action-vision modeling in the same token space enables precise semantic alignment between the two modalities, leading to accurate and reliable action following. Second, viewing discretization as a form of refinement and compression, we conduct the entire world modeling—including semantic understanding and video generation—in a discrete token space to facilitate the learning of higher-quality latent compressed representations, which is crucial for world models. Additionally, we build our *DiscreteWorld* upon GRNs—a text-to-video discrete generative model—and thereby naturally inherit its iterative visual refinement capability for high-quality video synthesis. Extensive experiments on real-world and game scenarios show that *DiscreteWorld* achieves more accurate action following and more stable long-horizon generation than prior diffusion-based world models. All models and code will be open-sourced to foster further research.

1 INTRODUCTION

World modeling constitutes a foundational capability for embodied intelligence, providing a predictive substrate for anticipating how environments evolve under actions in planning, control, and

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.

interaction (Ha & Schmidhuber, 2018; Hafner et al., 2020). As a concrete instantiation of this formulation, world exploration studies a navigation-centric setting in which future observations are generated under directional controls, requiring the model to simulate scene evolution along intended trajectories.

Recent world exploration models (Huang et al., 2025; Zhang et al., 2025; He et al., 2025; Mao et al., 2025; Zhu et al., 2025) have achieved substantial progress by building upon diffusion-based video generators (Labs, 2024; Wang et al., 2025a). Despite their strong perceptual fidelity, maintaining precise action alignment and long-horizon trajectory stability remains challenging. Similar observations have been reported in GLM-Image (Z.ai, 2026), which suggests that diffusion-centric formulations can be less natural for complex instruction following and dense semantic expression. Meanwhile, the remarkable success of large-scale token-based modeling provides a complementary perspective. LLMs demonstrate the effectiveness of token representations for complex reasoning, while recent multimodal models show that heterogeneous signals can be unified in a shared token space for understanding and generation (OpenAI, 2025a;b; Team et al., 2026). These observations motivate a discrete generative formulation for world modeling in which actions and visual states are represented within a unified token space.

Among discrete generative models, GRNs (**G**enerative **R**efinement **N**etworks) Han et al. (2026) provide a state-of-the-art backbone, but it is originally designed for standalone text- or class-conditioned generation. Extending GRNs to world exploration requires a formulation that can jointly represent actions, visual context, and future observations within a single generative process. To this end, we present *DiscreteWorld*, which quantizes camera actions and visual observations into a unified discrete action-vision token space for joint semantic understanding and video generation. This unified token space promotes semantic alignment between actions and visual dynamics, leading to accurate and reliable action following. Moreover, by treating discretization as refinement and compression, DiscreteWorld performs the full world modeling process in discrete tokens, supporting compact latent representations that are essential for long-horizon world modeling. Built upon GRNs, DiscreteWorld further inherits iterative discrete refinement for high-quality video synthesis and adapts it to long-horizon rollout through anchor-based global grounding, historical-context conditioning, and future-token refinement under aligned actions.

We evaluate DiscreteWorld in world exploration settings that require persistent action following across both real-world and game scenarios. Experiments show that DiscreteWorld achieves superior action following and more stable long-horizon prediction than prior world exploration models. In particular, the results demonstrate that discrete-based modeling provides an effective formulation for world exploration, suggesting that discrete world modeling is not merely an alternative implementation of existing video generators, but a promising paradigm for scalable and controllable world exploration.

To summarize, our contributions are as follows:

- We establish a viable discrete generative paradigm for world exploration through *DiscreteWorld*, enabling stable, high-quality generation of 10-second, 480p open-world videos under camera-action control.
- We unify semantic understanding and video generation within a discrete action-vision token space, enabling stronger action-vision alignment and more reliable action following.
- We extend GRNs from generic discrete video generation to action-conditioned world exploration, achieving stronger action following and more stable long-horizon generation on real-world and game scenarios.

2 Related Work

2.1 Discrete Generative Models

Encouraged by the immense success of large language models, recent approaches Esser et al. (2021); Sun et al. (2024); Kondratyuk et al. (2024); Wang et al. (2025c) explore generating images or videos by predicting the next visual token. MaskGIT Chang et al. (2022) accelerates generation speed and improves quality by using parallel decoding and a masking strategy that predicts high-confidence tokens first and then iteratively fills in the rest. VAR Tian et al. (2024) shifts the autoregressive paradigm form next-token to next-scale prediciton, delivering better generation results as well as over

$10\times$ inference speedup. Nevertheless, autoregressive models are still constrained by lossy discrete tokenization and error accumulation. Their overall performance, especially for video generation, is still lag behind diffusion/flow methods. Although Infinity Han et al. (2024) proposes self-correciton by randomly flipping bitwise labels, the assumption of less than 30% diffuse errors covers limited error patterns. Recently, GRNs Han et al. (2026) perform generation via iterative refinement over hierarchical binary tokens, mitigating error accumulation and achieving state-of-the-art performance among discrete generative models. In this work, we extend GRNs beyond standalone visual generation by reformulating their refinement process for action-conditioned world exploration.

2.2 World Exploration Model

Most existing methods Huang et al. (2025); He et al. (2024); Bai et al. (2025); Zhang et al. (2025); He et al. (2025) built upon diffusion-based video generators Labs (2024); Wang et al. (2025a). Recently, Matrix-Game 2.0 He et al. (2025) focuses on real-time streaming generation via few-step autoregressive diffusion, combining a game-centric data pipeline, frame-level keyboard and mouse action injection, and causal distillation. YUME Mao et al. (2025) develops an interactive world generation framework based on a Masked Video Diffusion Transformer, incorporating quantized camera actions, memory mechanisms, and accelerated sampling for long-horizon controllable exploration from images, text, or videos. Astra Zhu et al. (2025) further extends diffusion-based world modeling with an autoregressive denoising architecture, temporal causal attention, noise-augmented history memory, an action-aware adapter, and a mixture of action experts for heterogeneous control modalities.

Beyond diffusion-based formulations, early efforts have also explored token-based world modeling. iVideoGPT (Wu et al., 2024) formulates interactive world modeling as autoregressive next-token prediction, discretizing visual observations through a compressive tokenizer while incorporating actions as continuous conditioning signals. However, it does not fully realize unified discrete modeling of both actions and vision, and mainly targets robotic manipulation at relatively low resolutions of 64×64 resolution with about 138M parameters. In contrast, our DiscreteWorld models camera actions and visual observations within a unified discrete token space, enabling high-quality 480p long-horizon world exploration with scalable 2B parameters.

3 Preliminary: Generative Refinement Networks

GRNs Han et al. (2026) operate in a discrete latent space obtained via HBQ (**H**ierarchical **B**inary **Q**uantization). Given a continuous feature $F \in (-1, 1)$, HBQ performs multi-round binary quantization in a coarse-to-fine manner. At round i, the quantization decision is defined as:

$$c_i = \sum_{j=1}^{i-1} \frac{\delta[q_j]}{2^j}, q_i = \begin{cases} 0 & \text{if } F \leq c_i, \\ 1 & \text{if } F > c_i, \end{cases} \tag{1}$$

where $\delta[q] = -1$ if $q = 0$ and $+1$ otherwise. The final quantized representation is reconstructed as:

$$\hat{F} = \sum_{i=1}^{n} \delta[q_i] \cdot 2^{-i}, \tag{2}$$

This formulation yields a binary-tree-like refinement process, where each additional bit reduces the quantization error exponentially (i.e., $e_i < 2^{-i}$), enabling near-lossless discrete representations.

Based on this discrete representation, GRNs supports generation via iterative refinement over token maps $Y_{bit} = Concat(q_1, q_2, ..., q_n)$. At each step m, a partially observed input state F_m is constructed using a binary selection map S_m:

$$F_m = S_m \cdot Y_m \oplus \overline{S_m} \cdot Y_{rand}, \tag{3}$$

where Y_m denote the predicted token map at step m, and Y_{rand} denote a randomly initialized token map. For the next step, given F_m, the model $\Phi(\cdot)$ predicts the next token distribution via $p(Y_{m+1}) = \Phi(F_m)$. The binary map S is sampled as $S_{m+1} = RandLike(Y_t) < p(Y_{m+1}, m+1)$, where $RandLike(\cdot)$ generates i.i.d. uniform random variables in [0,1]. Here, $p(Y_{m+1}, m+1)$ is an entropy-based schedule function for dynamic-step generation.

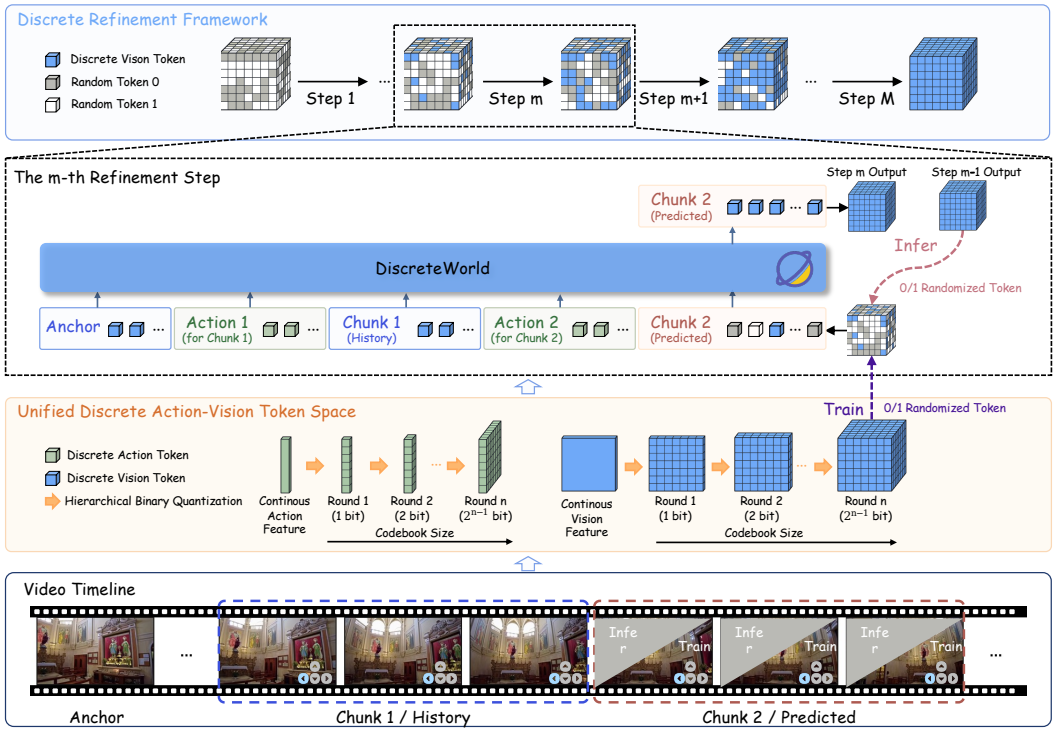


Figure 2: **Overview of DiscreteWorld.** Video frames are factorized into an anchor, history chunk, and a predicted chunk, each chunk aligned with its actions. Hierarchical Binary Quantization maps both action and vision into a unified discrete action-vision token space, enabling joint modeling. The predicted chunk is progressively refined over multiple steps: training uses partially corrupted ground-truth tokens, while inference starts from random tokens and progressively refine them.

4 METHOD

Figure 2 illustrates the framework of *DiscreteWorld*. We formulate world exploration in a unified discrete action-vision token space, enabling joint semantic understanding and video generation over compact discrete representations. This promotes action-vision alignment and uses discretization as compression to support compact latent representations for world modeling. Built upon GRNs, DiscreteWorld adapts iterative discrete refinement to long-horizon action-conditioned world exploration.

4.1 UNIFIED DISCRETE ACTION-VISION TOKEN SPACE

Existing diffusion-based methods typically inject actions as external conditions into visual denoising networks. In contrast, DiscreteWorld represents actions as native tokens in the same discrete space as vision. This allows action and visual tokens to interact through in-context self-attention, promoting semantic alignment between control signals and visual dynamics for reliable action following.

For each frame t, we are given camera intrinsics $\mathbf{K}_t$ and camera extrinsics $\mathbf{T}_t$:

$$\mathbf{K}_t = \begin{bmatrix} f_{x,t} & 0 & c_{x,t} \\ 0 & f_{y,t} & c_{y,t} \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{T}_t = \begin{bmatrix} \mathbf{R}_t & \mathbf{t}_t \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{4}$$

Here $(f_{x,t}, f_{y,t}, c_{x,t}, c_{y,t})$ denote focal lengths and principal-point coordinates, while $\mathbf{R}_t$ and $\mathbf{t}_t$ denote camera rotation and translation under the world-to-camera convention. Since world exploration is governed by local state transitions, we compute relative extrinsic along the temporal axis:

$$\Delta\mathbf{T}_0 = \mathbf{I}, \qquad \Delta\mathbf{T}_t = \mathbf{T}_{t-1}^{-1}\mathbf{T}_t, \quad t \geq 1. \tag{5}$$

We then concatenate intrinsics and relative extrinsic into an 11-dimensional action descriptor

$$\mathbf{a}_t = [f_{x,t}, f_{y,t}, c_{x,t}, c_{y,t}, q_{x,t}, q_{y,t}, q_{z,t}, q_{w,t}, t_{x,t}, t_{y,t}, t_{z,t}], \tag{6}$$

where $(q_{x,t}, q_{y,t}, q_{z,t}, q_{w,t})$ is the unit quaternion corresponding to $\Delta\mathbf{T}_t$ and $(t_{x,t}, t_{y,t}, t_{z,t})$ is the corresponding translation. Collecting descriptors over time yields the action sequence $\mathbf{A} = [\mathbf{a}_t]_{t=1}^{T}$.

To unify actions and visual observations in the unified token space, we quantize both modalities using HBQ. Let Q_{HBQ} denote the quantizer, and let $\mathbf{V}$ denote the visual latent sequence. We obtain the corresponding binary codes

$$\mathbf{Y}^a, \mathbf{Y}^v = Q_{\mathrm{HBQ}}(\mathbf{A}; \mathbf{V}). \tag{7}$$

We concatenate the action and visual token sequences to form the input $[\mathbf{Y}^a; \mathbf{Y}^v]$ of the generative Transformer. This places actions and visual observations on equal footing within the same discrete sequence, allowing self-attention to directly capture the dependencies between control signals and visual dynamics.

4.2 Discrete Refinement Framework for Long-Horizon World Exploration

Extending GRNs from class- and text-conditioned visual generation to world exploration requires more than changing the conditioning signal. It calls for a reformulation of discrete refinement under temporally evolving actions and historical context. DiscreteWorld introduces a discrete refinement framework for long-horizon world exploration.

Horizon Factorization. For each prediction horizon, the visual token map is decomposed as

$$\mathbf{Y}^v = [\mathbf{Y}_{\mathrm{anchor}}, \mathbf{Y}_{\mathrm{hist}}, \mathbf{Y}_{\mathrm{pred}}], \tag{8}$$

where $\mathbf{Y}_{\mathrm{anchor}}$ is a persistent global anchor, $\mathbf{Y}_{\mathrm{hist}}$ is a history chunk, and $\mathbf{Y}_{\mathrm{pred}}$ is a predicted chunk. The anchor corresponds to the first frame and provides a stable global reference for scene identity, the history chunk provides a frame or one-second context for local transition dynamics, and the predicted chunk contains the latent frames of the next two seconds. The asymmetric design also matches inference: after predicting a two-second future from the initial image, each subsequent horizon reuses only the last one second as historical context for efficient generation.

Unified Action-Vision Learning. Given the horizon factorization, training treats future prediction as discrete token recovery in the compressed action-vision space. For each training instance, we sample a random probability $\alpha_m \sim \mathcal{U}(0, 1)$ for refinement step m and construct a partially observed predicted chunk:

$$\tilde{\mathbf{Y}}_{\mathrm{pred}}^{(m)} = \mathbf{S}^{(m)} \cdot \mathbf{Y}_{\mathrm{pred}} \oplus \overline{\mathbf{S}^{(m)}} \cdot \mathbf{Y}_{rand}, \tag{9}$$

where $\mathbf{S}^{(m)} \sim \mathrm{Bernoulli}(1 - \alpha_m)$ is a binary preservation map, and $\mathbf{Y}_{rand}$ is a random binary token map. The full visual input is then formed as $\tilde{\mathbf{Y}}^{v,(m)} = [\mathbf{Y}_{\mathrm{anchor}}, \mathbf{Y}_{\mathrm{hist}}, \tilde{\mathbf{Y}}_{\mathrm{pred}}^{(m)}]$. The model is trained with cross-entropy objective:

$$\mathcal{L} = -\mathbb{E}\left[\sum_{m=1}^{M} \sum_{y \in \mathbf{Y}_{\mathrm{pred}}} \log p\Big(y \mid [\mathbf{Y}^a; \tilde{\mathbf{Y}}^{v,(m)}]\Big)\right] \tag{10}$$

Here, $\mathcal{L}$ supervises recovery of clean predicted visual tokens from corrupted refinement states, so that action and video are learned jointly in one discrete representation, supporting both semantic understanding and video generation. The training procedure is summarized in Algorithm 1 in the Appendix.

Chunk-wise Autoregressive Inference. At inference time, DiscreteWorld performs long-horizon generation with a fixed anchor and overlapping prediction windows. For each horizon, $\mathbf{Y}^a$ is the action tokens of the predicted chunk, which is initialized from random binary tokens, i.e., $\mathbf{Y}_{\mathrm{pred}}^{(0)} = \mathbf{Y}_{rand}$. Given a 50-step refinement, DiscreteWorld iteratively predicts and updates the predicted chunk:

$$\hat{\mathbf{Y}}_{\mathrm{pred}}^{(m)} = f_\theta\Big([\mathbf{Y}^a; \mathbf{Y}_{\mathrm{anchor}}, \mathbf{Y}_{\mathrm{hist}}, \mathbf{Y}_{\mathrm{pred}}^{(m)}]\Big), \tag{11}$$

$$\mathbf{Y}_{\mathrm{pred}}^{(m+1)} = \mathbf{S}^{(m+1)} \cdot \hat{\mathbf{Y}}_{\mathrm{pred}}^{(m)} \oplus \overline{\mathbf{S}^{(m+1)}} \cdot \mathbf{Y}_{rand}, \tag{12}$$

where $\mathbf{S}^{(m+1)} \sim \mathrm{Bernoulli}(1 - \rho_{m+1})$, with ρ_m decreasing from 1 to 0. As ρ_m decreases, the predicted chunk gradually transitions from random initialization to model-predicted tokens. After the final refinement step, the predicted chunk is decoded into video, and its last one-second segment is used as the historical context for the next horizon. This forms an autoregressive inference procedure for long-horizon world exploration, as summarized in Appendix Algorithm 2.

Table 1: **Controlled comparison under matched settings.** Cont. and Disc. denote continuous and discrete representations. Rot. and Trans. are rotation and translation errors; Subj. and Back. are subject and background consistency; Motion, Aesth., and Imaging denote motion smoothness, aesthetic quality, and imaging quality.

Action	Video	Paradigm	Instruction Following		Visual Quality				
			Rot. ↓	Trans. ↓	Subj. ↑	Back. ↑	Motion ↑	Aesth. ↑	Imaging ↑
Cont.	Cont.	Diffusion	4.43	9.44	0.870	**0.930**	0.987	**0.382**	0.584
Disc.	Cont.	Diffusion	4.88	9.90	**0.872**	0.924	**0.989**	0.379	**0.588**
Cont.	Disc.	GRN	3.35	7.87	0.868	0.919	0.979	0.371	0.563
Disc.	Disc.	GRN	**2.81**	**6.98**	0.866	0.921	0.983	0.373	0.565

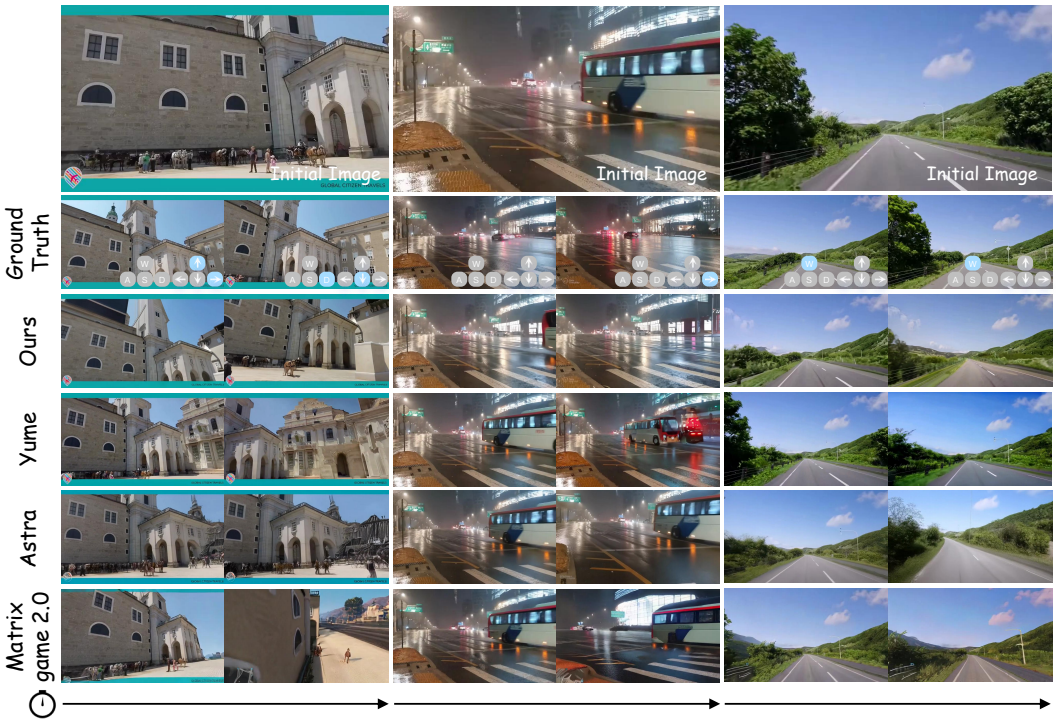


Figure 3: **Qualitative comparison on real-world scenes with state-of-the-art methods.**

5 Experiments

Dataset. For the real-world experiments, we train on SpatialVID-HQ, a high-quality subset of SpatialVID Wang et al. (2025b) containing diverse in-the-wild videos with per-frame camera intrinsics and extrinsics. For the game experiments, we train on a private 41-hour dataset collected from five game environments, also with per-frame camera intrinsics and extrinsics. Following the evaluation scale of Yume Mao et al. (2025) and Astra Zhu et al. (2025), we use 50 held-out videos from SpatialVID-HQ to evaluate real-world scenes and 50 held-out videos from the private game dataset to evaluate game scenes. Appendix Figure 9 shows that the test set covers diverse actions.

Metrics. Instruction following and visual quality are two key aspects of world exploration. Following the evaluation protocol of Astra Zhu et al. (2025), we measure instruction following using MegaSaM-based Li et al. (2025) rotation and translation errors, and visual quality using five VBench Huang et al. (2024) metrics: subject consistency, background consistency, motion smoothness, aesthetic quality, and imaging quality. For comparison with state-of-the-art methods, the evaluated methods use action conditions with different representations. We therefore additionally report a human score, defined as the proportion of generated videos judged by 100 participants to follow the intended camera action.

Table 2: **Quantitative comparison with state-of-the-art methods on real-world and game scenarios.** Human is the human-rated instruction-following score; Rot. and Trans. are rotation and translation errors; Subj. and Back. are subject and background consistency; Motion, Aesth., and Imaging denote motion smoothness, aesthetic quality, and imaging quality. Best and second-best results within each scene type are shown in **bold** and *italic*.

Method	Scene	Instruction Following			Visual Quality				
		Human ↑	Rot. ↓	Trans. ↓	Subj. ↑	Back. ↑	Motion ↑	Aesth. ↑	Imaging ↑
Matrix-Game 2.0 He et al. (2025)	Real	0.564	4.21	9.14	0.816	0.885	0.988	0.482	0.655
Astra Zhu et al. (2025)		0.543	4.39	9.36	0.874	0.918	0.984	0.486	0.642
Yume Mao et al. (2025)		*0.678*	*3.47*	*8.26*	*0.906*	**0.934**	**0.991**	*0.513*	**0.708**
DiscreteWorld (Ours)		**0.723**	**2.32**	**6.03**	**0.928**	*0.930*	*0.990*	**0.527**	*0.701*
Matrix-Game 2.0 He et al. (2025)	Game	*0.644*	*2.97*	*7.29*	0.836	0.924	0.984	0.556	0.638
Astra Zhu et al. (2025)		0.620	3.43	8.05	0.797	0.869	0.978	0.511	0.603
Yume Mao et al. (2025)		0.367	9.42	12.93	**0.941**	**0.943**	**0.991**	**0.622**	**0.705**
DiscreteWorld (Ours)		**0.736**	**2.16**	**5.87**	*0.866*	*0.936*	*0.975*	*0.590*	*0.654*



Figure 4: **Qualitative comparison on game scenes with state-of-the-art methods.**

5.1 Unified Discrete Action–Vision Space Improves Instruction Following

Direct comparison with state-of-the-art methods involves different backbones, training data, and action interfaces, as shown in Appendix Table 7, making it difficult to isolate the sources of DiscreteWorld's performance gains. We therefore conduct a controlled study under strictly matched settings to examine how continuous and discrete modeling of different modalities affects world exploration. ***We observe that a unified discrete action–vision space outperforms mixed setups on instruction following.***

Experimental Setup. We use two pretrained tokenizers released with GRN Han et al. (2026): the 16-channel continuous HBQ tokenizer without quantization and its eight-round discrete counterpart. As reported in GRN, they achieve comparable reconstruction PSNRs of 31.14 and 31.10, respectively, minimizing reconstruction quality as a confounding factor.

All variants adopt the same Transformer architecture based on the

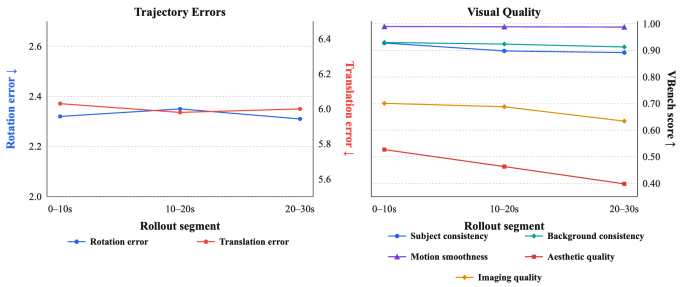


Figure 5: **Long-horizon rollout.**

Table 3: **Action quantization under the matched setting.** The one-level baseline uses an LFQ-style codebook-free autoencoder with the same eight-bit budget.

Action quantization	Bits	Rot. ↓	Trans. ↓	Subj. ↑	Back. ↑	Motion ↑	Aesth. ↑	Imaging ↑
Hierarchical	8	**2.81**	**6.98**	**0.866**	0.921	**0.983**	**0.373**	**0.565**
One-level	8	2.95	7.10	0.864	**0.925**	0.982	0.366	0.564

official GRN implementation[1] and are trained from scratch under the same optimization budget. They are trained on SpatialVID-HQ, with camera intrinsics and extrinsics injected. The variants differ only in whether the action and visual modalities are modeled continuously or discretely and in the corresponding diffusion or GRN training objective.

Experimental Analysis. As shown in Table 1, we analyze the results from three perspectives. (1) *Generative objective.* GRN outperforms diffusion in instruction following even with continuous actions. One possible explanation is that actions in diffusion interact with noisy visual features throughout denoising, whereas GRN retains partially uncorrupted visual tokens during refinement, potentially providing more stable context for action–vision alignment. (2) *Action representation.* Discrete actions further improve instruction following within GRN, suggesting that a unified discrete action–vision space facilitates cross-modal interaction. (3) *Visual quality.* Diffusion performs slightly better on several visual-quality metrics, indicating that improved action alignment does not necessarily yield higher perceptual quality. The visual quality of DiscreteWorld is also constrained by the underlying video-generation capability of GRN.

5.2 Comparison with State-of-the-Art Methods

We compare DiscreteWorld with state-of-the-art world models to evaluate its end-to-end performance in real and game scenarios. Each method is evaluated under its native configuration. Therefore, this experiment compares complete systems rather than individual design choice.

Experimental Setup. We initialize DiscreteWorld from GRN-2B and adopt a progressive-resolution training schedule on SpatialVID-HQ, consisting of 100k steps at 192p followed by 120k steps at 480p. We then fine-tune the model on the game dataset at 480p. We use AdamW with a learning rate of 10^{-5} and train on 64 H100 GPUs with a per-GPU batch size of three.

Experimental Analysis. As shown in Table 2, discreteWorld achieves the highest human scores and the lowest trajectory errors in both real-world and game scenarios. Appendix Table 9 further confirms its advantage across most action categories. DiscreteWorld also maintains competitive visual quality despite using GRN-2B, whose VBench score (82.99) is lower than those of Wan-1.3B (83.96) used by Matrix-Game 2.0 and Astra, and Wan-14B (86.22) used by Yume. This suggests that the remaining visual-quality gap is largely influenced by the underlying generation capability. Yume obtains higher consistency scores in game scenarios, but its weaker instruction following suggests that its smaller motion magnitude may favor consistency metrics. Qualitatively, Figures 3 and 4 show that DiscreteWorld produces camera motion that better matches the specified actions while preserving scene content and visual coherence over time.

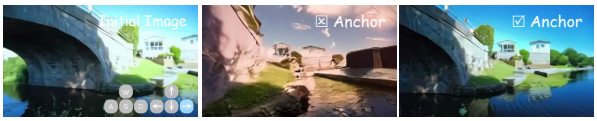


Figure 6: **Ablation on global anchor.**

To ensure comparability with prior camera-controllable world models such as Astra, which are typically evaluated over 8–10-second horizons (Table 8), we use 10-second videos for the main comparison . We additionally evaluate DiscreteWorld over 30-second rollouts. As shown in Figure 5, trajectory errors remain stable, indicating sustained action following, whereas aesthetic and imaging quality gradually decline due to accumulated visual drift.

[1] https://github.com/bytedance/GRN

5.3 Ablation Studies

Unless otherwise specified, all ablations use the real-world DiscreteWorld model trained on SpatialVID-HQ for 100k steps at 192p, following the first stage of the training schedule used in our comparison with state-of-the-art methods.

Ablation on Action Quantization Strategy. Building on the discrete GRN configuration in Table 1, we compare hierarchical and learned one-level action quantization under the same eight-bit budget. Following lookup-free quantization (LFQ) Yu et al. (2024), the one-level variant uses a lightweight codebook-free autoencoder that maps each normalized action value directly to an eight-bit binary vector. The autoencoder is first trained with an action-reconstruction objective and then frozen during world-model training. Except for the quantization strategy, all settings are identical to those in the controlled study in Section 5.1. As shown in Table 3, both strategies achieve strong instruction following, while hierarchical quantization yields slightly lower rotation and translation errors with comparable visual quality. These results indicate that the main benefit arises from representing actions and videos within a unified discrete space, with hierarchy providing a modest additional gain.

Ablation on Global Anchor. Figure 6 shows that removing the persistent global anchor causes the generated scene to drift from the initial observation, whereas the complete model better preserves scene identity. The corresponding quantitative results are reported in Appendix Table 5.



Figure 7: **Ablation on refinement steps.**

Ablation on Refinement Steps. Figure 7 shows that reducing refinement steps from 50 to 30 generally preserves scene structure, with occasional degradation in fine details. The small quantitative differences in Appendix Table 4 indicate that our performance gains do not come from sampling more refinement steps at inference time. We therefore settle on 50 refinement steps for all reported experiments.

Ablation on Sampling Temperature. Sampling temperature controls the stochasticity of token prediction. Figure 8 shows that a higher temperature increases diversity in newly exposed regions. Appendix Table 6 shows that increasing τ from 0.1 to 1.0 causes only minor changes in the VBench metrics while producing richer visual diversity. We therefore use $\tau = 1.0$ by default.



Figure 8: **Ablation on sampling temperatures.**

6 Conclusion

We presented *DiscreteWorld*, a discrete models for long-horizon, action-conditioned world exploration. DiscreteWorld unifies camera actions and visual observations in a discrete token space to improve action-vision alignment, and treats discretization as refinement and compression to support compact representations for high-quality world modeling. Experiments on real and game scenarios show more accurate action following and stable long-horizon generation than prior models.

7 Limitations

DiscreteWorld is not yet optimized for real-time inference. In Table 8, its runtime is comparable to that of diffusion-based long-horizon models at a similar scale: DiscreteWorld uses 50 refinement steps, while Astra and Yume use 50 denoising steps. Matrix-Game 2.0 is substantially faster because it specifically targets real-time streaming through diffusion-step distillation and other acceleration techniques. Since efficient distillation for discrete refinement remains relatively underexplored and is orthogonal to our formulation, we consider it is an important direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to assist with understanding existing code, diagnosing and fixing implementation bugs, and improving the clarity and grammar of the manuscript. All AI-assisted code changes were manually reviewed and validated through experiments, and all AI-assisted writing was checked and revised by the authors. Generative AI was not used to fabricate data, results, citations, or scientific claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

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*, pp. 14834–14844, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11315–11325, 2022.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

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

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. *arXiv preprint arXiv:2412.04431*, 2024.

Jian Han, Jinlai Liu, Jiahuan Wang, Bingyue Peng, and Zehuan Yuan. Generative refinement networks for visual synthesis. *arXiv preprint arXiv:2604.13030*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

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*, 2025.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. *arXiv preprint arXiv:2505.14357*, 2025.

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*, pp. 21807–21818, 2024.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024.

Black Forest Labs. Flux. `https://blackforestlabs.ai/announcing-black-forest-labs/`, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10486–10496. IEEE, 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.

OpenAI. Introducing 4o image generation. https://openai.com/index/introducing-4o-image-generation/, March 2025a. OpenAI blog post.

OpenAI. Addendum to gpt-4o system card: 4o image generation. https://openai.com/index/gpt-4o-image-generation-system-card-addendum/, March 2025b.

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.

Meituan LongCat Team, Bin Xiao, Chao Wang, Chengjiang Li, Chi Zhang, Chong Peng, Hang Yu, Hao Yang, Haonan Yan, Haoze Sun, et al. Longcat-next: Lexicalizing modalities as discrete tokens. *arXiv preprint arXiv:2603.27538*, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *arXiv preprint arXiv:2404.02905*, 2024.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025a.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025b.

Jiahuan Wang, Yuxin Chen, Jun Yu, Guangming Lu, and Wenjie Pei. Editinfinity: Image editing with binary-quantized generative models. *arXiv preprint arXiv:2510.20217*, 2025c.

Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: Interactive videogpts are scalable world models. *Advances in Neural Information Processing Systems*, 37:68082–68119, 2024.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion-tokenizer is key to visual generation. In *International Conference on Learning Representations*, volume 2024, pp. 765–783, 2024.

Z.ai. GLM-Image: Auto-regressive for dense-knowledge and high-fidelity image generation. https://z.ai/blog/glm-image, January 2026. Accessed: 2026-05-07.

Yifan Zhang, Chunli Peng, Boyang Wang, Puyi Wang, Qingcheng Zhu, Fei Kang, Biao Jiang, Zedong Gao, Eric Li, Yang Liu, et al. Matrix-game: Interactive world foundation model. *arXiv preprint arXiv:2506.18701*, 2025.

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.

A APPENDIX

The appendix is organized as follows:

- Subsection A.1 details the unified action-vision learning procedure.
- Subsection A.2 describes the chunk-wise autoregressive inference procedure.
- Subsection A.3 reports the quantitative results of ablation studies.
- Subsection A.4 summarizes system comparsion details.
- Subsection A.5 reports the full test-set action-combination distribution.
- Subsection A.6 reports the complete per-action trajectory-error breakdown.
- Subsection A.7 presents additional qualitative results on real-world and game scenarios.
- Subsection A.8 discusses potential broader impacts of discrete world exploration models.
- Subsection A.9 outlines safeguards for responsible use and deployment.
- Subsection A.10 discusses data, bias, and misuse-related ethical considerations.

A.1 TRAINING PROCEDURE

Algorithm 1 summarizes the unified action-vision learning procedure of DiscreteWorld. For each action-visual window $(\mathbf{Y}^a, \mathbf{Y}_{\text{anchor}}, \mathbf{Y}_{\text{hist}}, \mathbf{Y}_{\text{pred}})$, the anchor and history chunk are kept clean to provide stable visual grounding, while only the predicted chunk is partially corrupted. Specifically, for each refinement state m, we independently sample a corruption probability α_m and construct $\tilde{\mathbf{Y}}^{(m)}_{\text{pred}}$ by mixing ground-truth predicted tokens with random binary tokens. The objective trains the model to recover $\mathbf{Y}_{\text{pred}}$ across different refinement states.

Algorithm 1: Training Procedure

Input : Action-visual window $(\mathbf{Y}^a, \mathbf{Y}_{\text{anchor}}, \mathbf{Y}_{\text{hist}}, \mathbf{Y}_{\text{pred}})$
Hyperparameter : Number of refinement states M
Output : Training loss $\mathcal{L}$

$\mathcal{L} \leftarrow 0$;
for $m = 1, \ldots, M$ **do**
  $\alpha_m \sim \mathcal{U}(0, 1)$;
  $\mathbf{Y}^{(m)}_{rand} \sim \text{Unif}\{0, 1\}^{|\mathbf{Y}_{\text{pred}}|}$;
  $\mathbf{S}^{(m)} \sim \text{Bernoulli}(1 - \alpha_m)$;
  $\tilde{\mathbf{Y}}^{(m)}_{\text{pred}} \leftarrow \mathbf{S}^{(m)} \cdot \mathbf{Y}_{\text{pred}} \oplus \overline{\mathbf{S}^{(m)}} \cdot \mathbf{Y}^{(m)}_{rand}$;
  $\tilde{\mathbf{Y}}^{v,(m)} \leftarrow [\mathbf{Y}_{\text{anchor}}, \mathbf{Y}_{\text{hist}}, \tilde{\mathbf{Y}}^{(m)}_{\text{pred}}]$;
  $\mathcal{L} \leftarrow \mathcal{L} - \sum_{y \in \mathbf{Y}_{\text{pred}}} \log p\Big(y \mid [\mathbf{Y}^a; \tilde{\mathbf{Y}}^{v,(m)}]\Big)$;
return $\mathcal{L}$;

A.2 INFERENCE PROCEDURE

Algorithm 2 summarizes the chunk-wise autoregressive inference procedure. DiscreteWorld keeps the initial anchor $\mathbf{Y}_{\text{anchor}}$ fixed across all horizons and generates each predicted chunk from random binary initialization. For horizon h, the history chunk is initialized by the anchor for the first horizon and by the last one-second segment of the previous predicted chunk for subsequent horizons. Within each horizon, the model iteratively predicts $\hat{\mathbf{Y}}^{(h,m)}_{\text{pred}}$ and updates $\mathbf{Y}^{(h,m)}_{\text{pred}}$ according to the decreasing schedule $\{\rho_m\}_{m=0}^{M}$. As ρ_m decreases from 1 to 0, fewer positions are replaced by random binary tokens, so the predicted chunk progressively transitions from random initialization to model-predicted discrete tokens. The decoded predicted chunks are finally concatenated to form the long-horizon video.

Algorithm 2: Inference Procedure

Input : Initial visual token $\mathbf{Y}_{\text{anchor}}$, action token windows $\{\mathbf{Y}^{a,(h)}\}_{h=1}^{H}$
Hyperparameter : Refinement schedule $\{\rho_m\}_{m=0}^{M}$ with $M = 50$ and $1 = \rho_0 > \rho_1 > \cdots > \rho_M = 0$
Output : Predicted video $\mathbf{V}$

for $h = 1, \ldots, H$ **do**
  if $h = 1$ **then**
    $\mathbf{Y}_{\text{hist}}^{(1)} \leftarrow \mathbf{Y}_{\text{anchor}}$;
  else
    $\mathbf{Y}_{\text{hist}}^{(h)} \leftarrow \text{Tail}_{1\text{s}}(\mathbf{Y}_{\text{pred}}^{(h-1)})$;
  $\mathbf{Y}_{rand}^{(h)} \sim \text{Unif}\{0,1\}^{|\mathbf{Y}_{\text{pred}}|}$;
  $\mathbf{Y}_{\text{pred}}^{(h,0)} \leftarrow \mathbf{Y}_{rand}^{(h)}$;
  for $m = 0, \ldots, M-1$ **do**
    $\hat{\mathbf{Y}}_{\text{pred}}^{(h,m)} \leftarrow f_\theta\Big([\mathbf{Y}^{a,(h)}; \mathbf{Y}_{\text{anchor}}, \mathbf{Y}_{\text{hist}}^{(h)}, \mathbf{Y}_{\text{pred}}^{(h,m)}]\Big)$;
    $\mathbf{S}^{(m+1)} \sim \text{Bernoulli}(1 - \rho_{m+1})$;
    $\mathbf{Y}_{\text{pred}}^{(h,m+1)} \leftarrow \mathbf{S}^{(m+1)} \cdot \hat{\mathbf{Y}}_{\text{pred}}^{(h,m)} \oplus \overline{\mathbf{S}^{(m+1)}} \cdot \mathbf{Y}_{rand}^{(h)}$;
  $\mathbf{Y}_{\text{pred}}^{(h)} \leftarrow \mathbf{Y}_{\text{pred}}^{(h,M)}$;
$\mathbf{V} \leftarrow \text{Decode}(\{\mathbf{Y}_{\text{pred}}^{(h)}\}_{h=1}^{H})$;
return $\mathbf{V}$;

Table 4: **Ablation on refinement steps.**

Refinement steps	Subj. ↑	Back. ↑	Motion ↑	Aesth. ↑	Imaging ↑
30	**0.896**	**0.938**	**0.991**	**0.422**	0.593
40	0.892	0.933	0.990	0.415	0.604
50	0.894	0.935	0.990	0.419	**0.605**

Table 5: **Ablation on the global anchor.**

Setting	Subj. ↑	Back. ↑	Motion ↑	Aesth. ↑	Imaging ↑
w/o global anchor	0.877	0.924	**0.990**	0.411	0.600
DiscreteWorld	**0.894**	**0.935**	**0.990**	**0.419**	**0.605**

Table 6: **Ablation on sampling temperature τ.**

τ	Subj. ↑	Back. ↑	Motion ↑	Aesth. ↑	Imaging ↑
0.1	0.889	**0.937**	**0.991**	**0.422**	**0.619**
0.5	**0.897**	0.935	**0.991**	0.419	0.609
1.0	0.894	0.935	0.990	0.419	0.605

A.3 Additional Ablation Results

Tables 4–6 report supplementary quantitative ablations on refinement steps, the global anchor, and sampling temperature. Since these factors have negligible effect on instruction following, we report only the visual-quality metrics.

Table 7: **Comparison settings of state-of-art methods.**

Method	Domain	Training data	Control
Matrix-Game 2.0	Game-specific	Sekai, Unreal Engine, Minecraft, and private GTA5/Temple Run data	Keyboard/mouse
Astra	General	nuScenes, Sekai, SpatialVID, RT-1, and Multi-Cam Video	Camera poses
Yume	Walking-specific	Sekai-Real-HQ	Text instructions
DiscreteWorld (Ours)	General	SpatialVID-HQ and a private game dataset	Camera poses

A.4 System Comparison Details

Table 7 and Table 8 summarize the settings of state-of-the-art methods, highlighting differences in application domains, training data, control mechanisms, and backbone architectures.

Table 8: **Inference horizon and efficiency comparison.** Time is measured for generating a 10-second 480p video. All methods use one H100 GPU except Yume, which uses two due to memory limits.

Method	Backbone	Horizon	Time
Matrix-Game 2.0	Wan-1.3B	A few seconds	4.8 seconds
Astra	Wan-1.3B	8–10 seconds	392 seconds
Yume	Wan-14B	8–10 seconds	3420 seconds
DiscreteWorld (Ours)	GRN-2B	10 seconds	592 seconds

A.5 Full Test Set Action Combination Distribution

Figure 9 presents the complete action-combination distribution of the test set without aggregation. The distribution shows that the evaluation set contains diverse directional controls and frequent simultaneous action combinations, providing a challenging benchmark for assessing action following under complex control conditions.

A.6 Per-Action Trajectory Errors

Table 9 breaks down frame-to-frame trajectory errors by action type. Our method achieves the lowest errors across nearly all movement directions, demonstrating consistently accurate trajectory prediction rather than improvements concentrated in a particular direction.

A.7 Additional Qualitative Results

Figure 10 and Figure 11 provide additional qualitative comparisons on real-world and game scenarios. The real-world examples cover diverse environments, including outdoor streets, natural scenes, indoor home scenes, architectural spaces, different seasons, and both daytime and nighttime conditions. The game examples further span indoor and outdoor environments, architectural and natural scenes, and both daytime and nighttime settings. Across these settings, DiscreteWorld maintains stable long-horizon visual quality and follows diverse action commands, demonstrating the robustness of the proposed discrete world exploration framework.

A.8 Broader Impacts

By improving action alignment and temporal stability, discrete-token world models such as DiscreteWorld can lower the cost of data collection for embodied AI, robotics, navigation, game simulation, virtual training, and interactive content creation by letting agents explore plausible futures in simulation. The same capability carries risks: controllable video generation could be misused to fabricate misleading or deceptive content, and overconfident predictions could lead to unsafe decisions if used

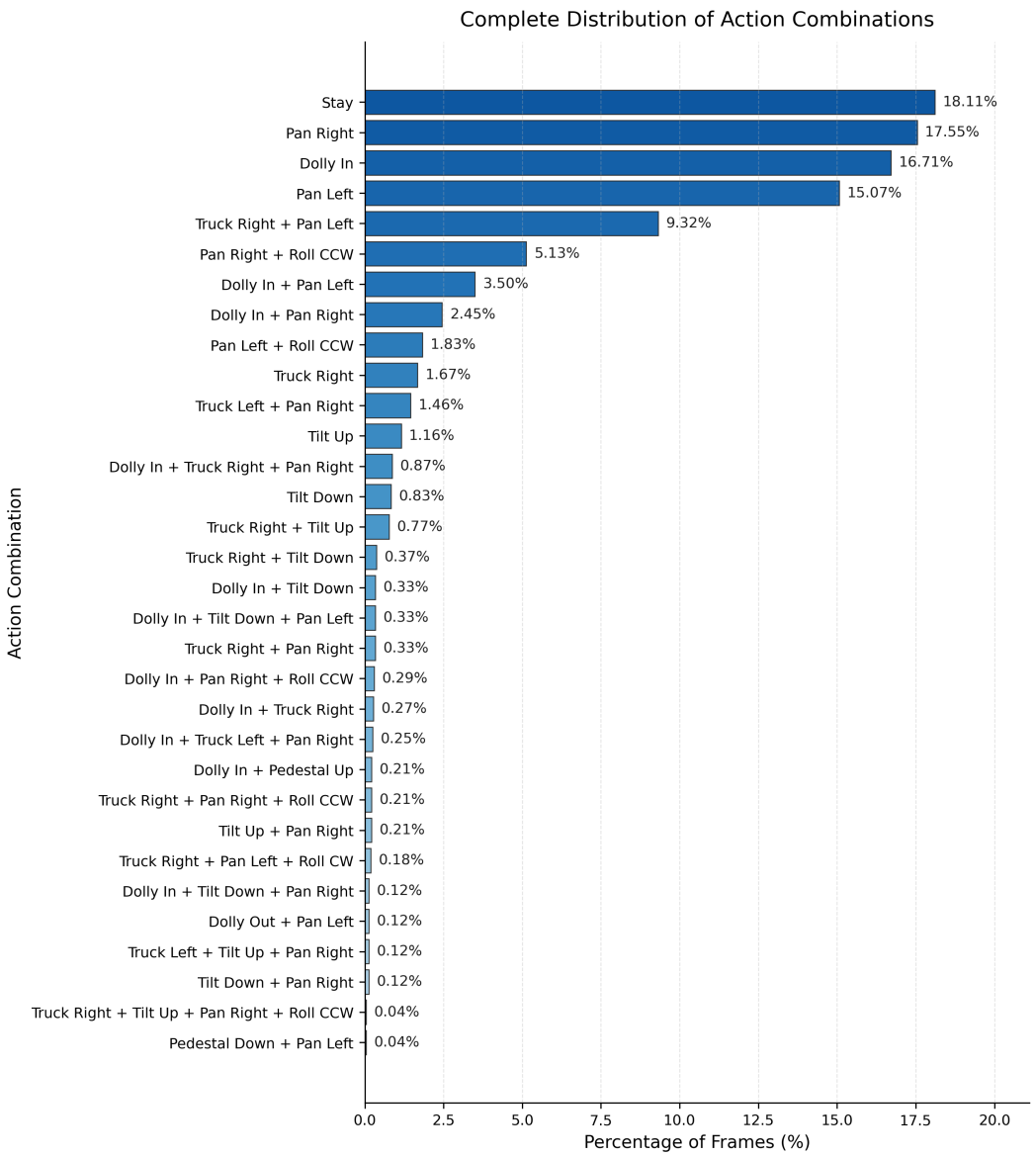


Figure 9: **Complete action-combination distribution.**

for planning without validation. We therefore view this work as a research step toward controllable world modeling, and stress that safety-critical deployment requires additional verification, uncertainty estimation, and domain-specific safeguards.

A.9 SAFEGUARDS

DiscreteWorld is an action-conditioned generative model, not an autonomous decision-making system; in our experiments it is evaluated only offline and is never connected to real-world actuation, which limits immediate physical risk. For future deployment, generated rollouts should be clearly labeled as synthetic, the model should be stress-tested under out-of-distribution actions, long horizons, and rare scenes before driving planning, and any control loop built on its outputs should add uncertainty estimation, human oversight, and fail-safe mechanisms—especially in robotics and navigation. Access to such systems should additionally be governed by usage policies that prohibit deceptive, harmful, or privacy-invasive use.

Table 9: **Per-action frame-to-frame trajectory errors.**

	Matrix-Game 2.0		Astra		Yume		DiscreteWorld	
Action type	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓	Rot.↓	Trans.↓
Stay	0.34	0.78	0.391	0.88	0.28	0.63	**0.192**	**0.45**
Pan right	0.46	1.04	0.50	1.12	0.451	1.01	**0.264**	**0.67**
Pan left	0.82	1.82	0.879	1.95	0.445	1.10	**0.40**	**0.87**
Truck right + pan left	**0.38**	**0.85**	0.429	0.92	0.412	0.93	0.409	0.92
Dolly in	0.35	0.76	0.393	0.86	**0.309**	**0.65**	0.36	0.79
Pan right + roll CCW	1.18	2.66	1.241	2.76	1.185	2.66	**0.724**	**1.61**
Dolly in + pan left	**0.29**	**0.66**	0.327	0.73	0.40	0.87	0.424	0.97
Truck left + pan right	0.40	0.88	0.454	1.03	0.419	0.94	**0.393**	**0.84**
Dolly in + pan right	0.54	1.22	0.596	1.31	0.545	1.23	**0.337**	**0.71**
Others	0.91	2.05	0.96	2.16	0.653	1.47	**0.525**	**1.18**

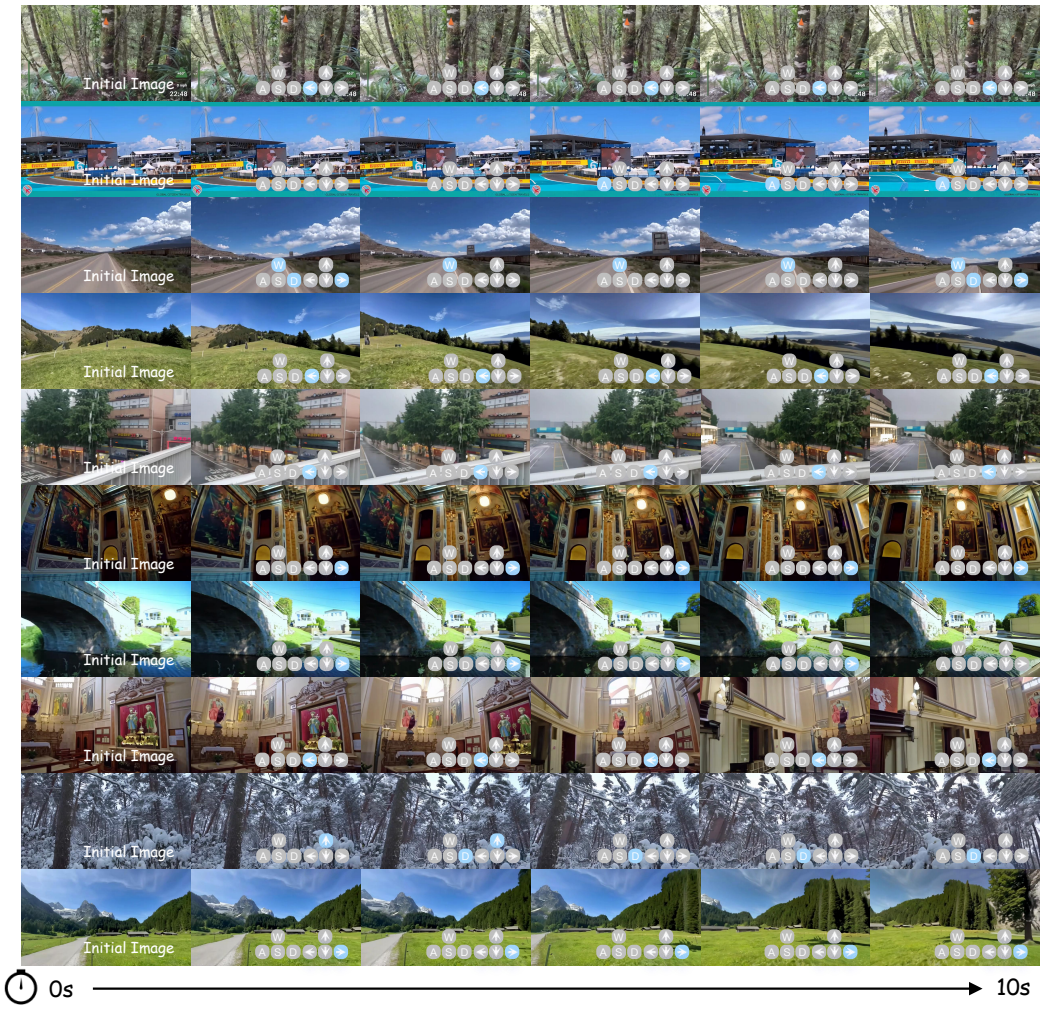


Figure 10: **More results on world exploration of real-world scenarios.**

A.10 Ethical Considerations

We use only existing curated real-world datasets (for research only) and simulated game-scene data, mindful of privacy, consent, and licensing.

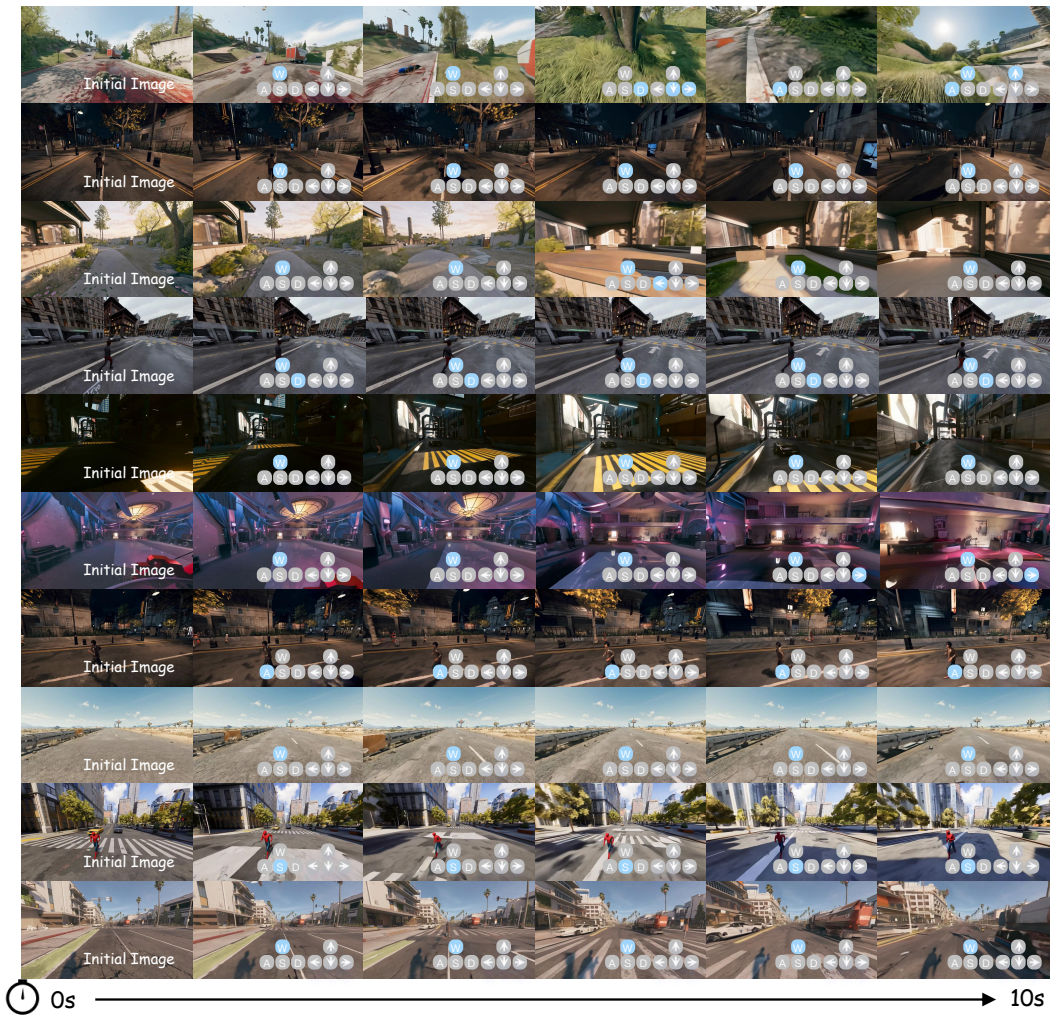


Figure 11: **More results on world exploration of game scenarios.**