WorldWeaver: Generating Long-Horizon Video Worlds via Rich Perception

Zhiheng Liu[1,2], Xueqing Deng[2], Shoufa Chen[1], Angtian Wang[2], Qiushan Guo[2],
Mingfei Han[2], Zeyue Xue[1], Mengzhao Chen[1], Ping Luo[1†], Linjie Yang[2†]

[1]The University of Hong Kong, [2]ByteDance Seed

Abstract

Generative video modeling has made significant strides, yet ensuring structural and temporal consistency over long sequences remains a challenge. Current methods predominantly rely on RGB signals, leading to accumulated errors in object structure and motion over extended durations. To address these issues, we introduce WorldWeaver, a robust framework for long video generation that jointly models RGB frames and perceptual conditions within a unified long-horizon modeling scheme. Our training framework offers three key advantages. First, by jointly predicting perceptual conditions and color information from a unified representation, it significantly enhances temporal consistency and motion dynamics. Second, by leveraging depth cues, which we observe to be more resistant to drift than RGB, we construct a memory bank that preserves clearer contextual information, improving quality in long-horizon video generation. Third, we employ segmented noise scheduling for training prediction groups, which further mitigates drift and reduces computational cost. Extensive experiments on both diffusion- and rectified flow-based models demonstrate the effectiveness of WorldWeaver in reducing temporal drift and improving the fidelity of generated videos. Page could be found here.

1   Introduction

Long-horizon video prediction [10, 76, 11, 77, 17, 45, 46] is a fundamental challenge for world modeling, as it requires models to internalize and reproduce the causal laws that govern scene dynamics over time. Although generative video modeling [65, 34, 34, 1, 48, 47, 49], has made rapid progress and shows impressive performance on short sequences less than 10 seconds, preserving realism and coherence over longer durations remains a significant challenge. Current approaches often struggle to depict correct motion and shape in long-horizon predictions, resulting in severe degradation including objects warp, motions drift, and violations of physical constraints as time progresses. These challenges highlight the difficulty of preserving both dynamics and temporal consistency in long video generation, which is critical for structured and predictive understanding of the world.

Most generative video models nowadays rely exclusively on RGB information as training signals, optimizing pixel-level reconstruction objectives. This causes them to favor color and texture over structural and dynamical attributes, such as motion trajectories, object geometry, and scale relationships [9, 40, 18]. In other words, training solely on RGB information can lead to the aforementioned instabilities. Such instabilities can be magnified over long sequences as small frame-wise errors accumulate and compound, resulting in pronounced temporal drift and structural degradation.

(a)

(b)

Prompt: A woman walks down the street, smiling as she puts on her sunglasses. She stops to wave hello, then turns around and walks away.
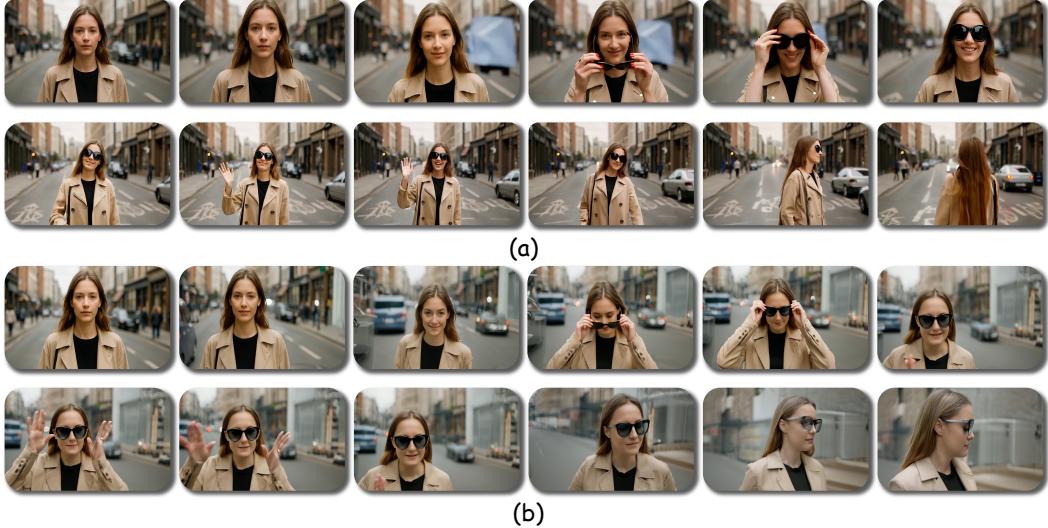
Figure 1: **WorldWeaver vs. existing approach on Long-Horizon Video Generation.** Compared to other methods (b), WorldWeaver (a) achieves superior temporal consistency and motion quality in long-horizon video generation.

To address these issues, as shown in Fig. 1, we introduce WorldWeaver, a robust framework for accurate long-horizon video prediction that seamlessly integrates joint modeling of RGB frames and perceptual signals. Our approach employs a perceptual-conditioned next prediction scheme to achieve precise and stable long-video generation. Our method encourages the model to capture not just color, but also structural and motion-related cues, thereby improving the robustness of the generated outputs. To achieve this, we incorporate multiple perceptual signals-such as video depth and optical flow-into a joint training framework. Extensive experiments confirm the effectiveness of using these auxiliary signals. Inspired by the observation that depth is more stable than RGB under temporal drift, we introduce a memory bank that stores reliable historical information from previous frames to enhance temporal consistency. This allows our model to maintain clearer long-term context and effectively mitigate the accumulation of drift in RGB predictions. In contrast, prior methods [10, 60, 77, 24] often inject substantial noise into historical information to prevent drift, which diminishes the utility of long-range context. Additionally, we adopt a segmented noise scheduling strategyassigning different noise levels to separate segments during training and constructing delayed denoising prediction groups during inferenceto further mitigate drift and reduce computational cost.

To evaluate our framework, we verify its effectiveness on both diffusion-based (CogVideoX-1.5B) and rectified flow-based (Wan2.1-1.3B) models on a robotic-arm manipulation dataset and a general-purpose video dataset. Extensive experiments demonstrate that our method substantially improves the stability and consistency of long-horizon video generation, effectively reducing temporal drift and structural distortions across diverse scenarios.

Our contributions can be summarized as follow:

- Systematically exploring the role of image-based perceptual condition, such as depth and optical flow, in enhancing long-horizon video generation as auxiliary signals.

- Proposing a unified framework that integrates perceptual conditioning and memory mechanisms for robust long-horizon video prediction.

- Extensive validation across different generative models and datasets, including both general-purpose and robotic manipulation domains, highlighting the potential of our approach as a foundation for scalable world models.

2 Related Works

Long video generation. With recent advances in video diffusion models [5, 34, 63, 38, 12], long video synthesis has garnered increasing attention for applications such as storytelling and simulations. Existing approaches span GAN-based methods [22, 41], training-free techniques [52, 64], distributed inference [61], autoregressive frameworks [74], and hierarchical strategies [31, 73]. Among these, FreeNoise [53] extends diffusion models by rescheduling noise without fine-tuning; StreamingT2V [32] integrates autoregressive modeling with memory modules for improved coherence; and TTT [13] adapts models at test time using specialized layers. LCT [24] enhances pre-trained models with 3D RoPE and asynchronous noise scheduling to produce scene-consistent multi-shot videos, while FAR [23] employs frame-level autoregressive modeling with FlexRoPE and context strategies for efficient long video generation. However, most existing methods rely solely on RGB information from historical frames, leading to cumulative structural and motion errors. In contrast, our approach incorporates perceptual signals—such as depth and optical flow—from historical frames to more effectively guide future frame generation.

Generative model for world simulation. World models, originating from foundational work [26], have evolved from RNN-based latent dynamics [27–29] to generative approaches such as diffusion-based [2, 15, 62, 19, 1] and autoregressive models [6, 44, 56, 39, 1]. Unlike agent-centric prediction, physical-world simulation emphasizes physics-informed realism for real-world applications. Early efforts aim to enhance simulator outputs: RL-CycleGAN [54] refines rendered images while aligning utility through Q-values, and RetinaGAN [33] preserves object features in simulated scenes for reinforcement learning. Recent diffusion [65, 57, 4] and autoregressive diffusion [10, 60, 74] frameworks surpass GANs by producing higher-fidelity, more temporally consistent, and realistic dynamic outputs. These advances enable video generation models to serve as learnable simulators, powering physics-grounded simulation in applications such as game environments [6, 62, 39, 7, 75, 14, 71], autonomous driving [35, 37, 51, 69, 20, 21], and robotic control [3, 16, 8, 30, 25, 78]. However, current video generation frameworks often prioritize color and texture over robust modeling of dynamics and spatial relationships [40, 9]. During inference, they frequently struggle to incorporate real-world knowledge beyond historical RGB data, limiting their ability to simulate complex interactions accurately.

3 Methods

In this section, we first introduce the preliminary and background on video diffusion models (Sec. 3.1). To enhance structure and motion modeling capabilities in video generation, we propose to jointly learn perceptual cues, such as depth and optical flow from video data that captures rich spatial structures and temporal dynamics (Sec. 3.2). We further propose a perception-conditioned long-horizon generation framework, which leverages a perceptual memory bank and group-wise noise scheduling to maintain robust context and mitigate drift in long video prediction (Sec. 3.3).

3.1 Preliminary

Video diffusion models offer a powerful framework for generating high-quality videos from conditional inputs. These models generate coherent video sequences by learning to reverse a noise-injection process, progressively refining random noise into structured outputs.

We adopt flow matching [43] which offers a deterministic alternative to stochastic diffusion processes. Flow matching constructs a continuous trajectory from a noise distribution to the target video distribution, conditioned on an input y, representing the conditioning signal (e.g., a text prompt). Formally, given a clean video sample $x_1 \in \mathbb{R}^{T \times C \times H \times W}$, a noise sample $x_0 \sim \mathcal{N}(0, I)$, and the condition y, the interpolation at timestep $t \in [0, 1]$ is: $x_t = tx_1 + (1 - t)x_0$. The model learns the conditional velocity of this trajectory, defined as: $v_t = \frac{dx_t}{dt} = x_1 - x_0$. The training objective minimizes the discrepancy between the predicted velocity $u(x_t, y, t; \theta)$ and the true velocity:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_1, x_0 \sim \mathcal{N}(0, I), y, t \in [0,1]} \left[\|u(x_t, y, t; \theta) - v_t\|_2^2 \right], \quad (1)$$

where θ denotes the model parameters and y is the conditioning input. During inference, an ordinary differential equation (ODE) solver integrates the velocity field to generate videos from
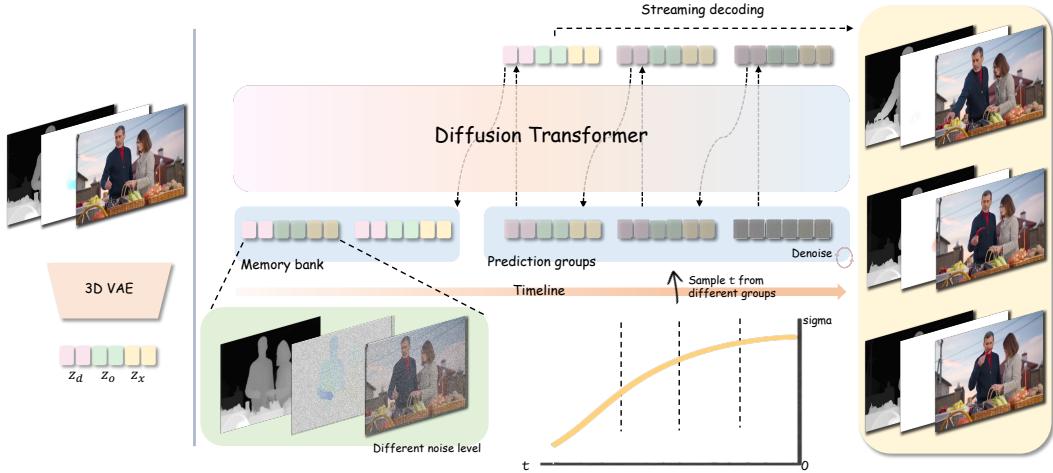
Figure 2: **Framework of WorldWeaver.** Given an input video, RGB, depth, and optical flow signals are encoded into a joint latent representation via a 3D VAE. The latents are split into a memory bank and prediction groups for the Diffusion Transformer. The memory bank stores historical frames and is excluded from loss computation; short-term memory retains a few fully denoised frames for fine details, while long-term memory keeps depth cues noise-free and adds low-level noise to RGB information. During training, prediction groups are assigned different noise levels according to the noise scheduler curve, aligned with the noise scheduling used during inference.

noise, guided by y. To mitigate the computational complexity of high-dimensional video data, models employ a pretrained variational autoencoder (VAE) to compress video sequences into a latent space. Given a video sequence $X \in \mathbb{R}^{F \times C \times H \times W}$, where F, C, H, and W represent the number of frames, channels, height, and width, respectively, the VAE maps it to a latent representation $z_x \in \mathbb{R}^{f \times c \times h \times w}$. This compression reduces the temporal and spatial dimensions while preserving essential features, facilitating efficient optimization in the latent space. Diffusion Transformer serves as the primary network architecture, leveraging attention mechanisms to model long-range spatio-temporal dependencies.

3.2 Enhancing Video Generation with Perceptual Conditions

Recent studies [40, 9] reveal a critical limitation in current diffusion-based video generation approaches: the learning is prioritized towards color information, leading to a neglect of essential attributes such as motion, relative size, shape, and spatial relationships. This bias hampers the ability of models to capture the dynamic and structural complexities of video data, particularly in next-frame prediction frameworks for long video generation. Here, distortions in shape, size, or motion in individual frames can trigger rapid error accumulation, leading to incoherent sequences. To address this, we introduce a unified modeling framework that integrates RGB data with perceptual conditions, specifically video depth and optical flow. Additionally, we evaluate the impact of various perceptual conditions, including video segmentation, with findings detailed in Sec. 4.3.

We first describe the process of preparing depth and optical flow data for our training videos. For depth estimation, we employ Video Depth Anything [68] to generate initial depth maps for each frame of the input video. Then we use DepthAnythingV2 [70] to compute relative depth across frames through least-squares optimization. The resulting depth sequences are normalized to form $D \in \mathbb{R}^{F \times 1 \times H \times W}$. To match the dimensionality of the VAE input, we repeat the single channel three times, giving $D \in \mathbb{R}^{F \times 3 \times H \times W}$. This depth representation is then compressed into a latent space using a variational autoencoder (VAE): $z_d = \mathcal{E}(D)$, where $\mathcal{E}$ denotes the VAE encoder.

Next, we employ SEA-RAFT [66] to extract optical flow between consecutive frames of the input video, producing a displacement field $O \in \mathbb{R}^{(F-1) \times 2 \times H \times W}$, where $O(u, v)$ represents the displacement of the pixels between frames. To encode this into a format compatible with our framework, we convert O into an RGB representation by computing the magnitude and direction of motion of each pixel, defined as $m = \min\left\{1, \frac{\sqrt{u^2+v^2}}{\sigma\sqrt{H^2+W^2}}\right\}$ and $\alpha = \arctan 2(v, u)$, with $\sigma = 0.15$,
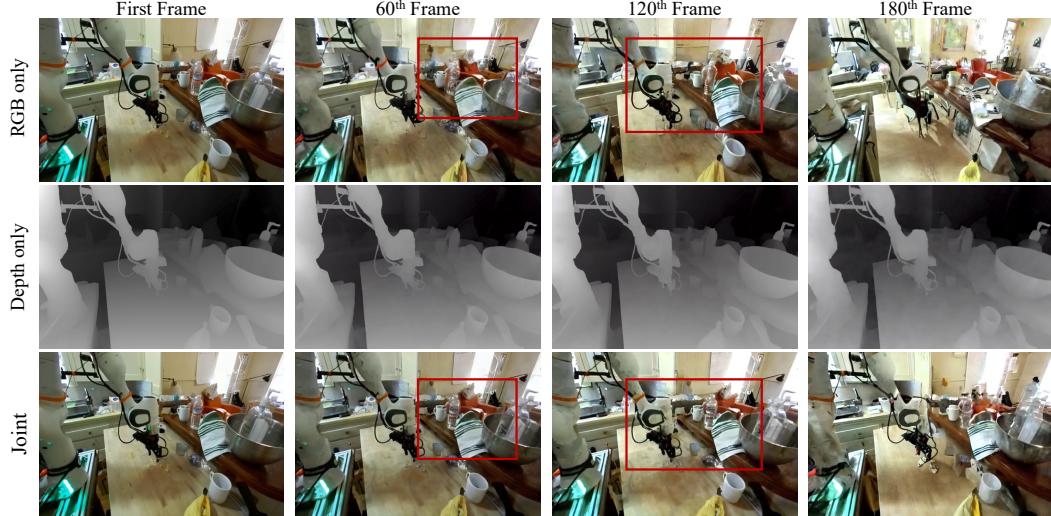
Figure 3: **Visualization of drift resistance.** We construct video data by repeating a single image during training and include "static" in the prompt. When the model is set to output only RGB, severe structural degradation and color distortion occur as the number of frames increases. In contrast, when outputting only depth, the structure remains well preserved. Jointly outputting both RGB and depth significantly alleviates visual drift. See experimental details in supplementary materials.

following VideoJAM [9]. Here, m determines pixel opacity, and α assigns a color based on the motion direction. To align with the VAE input, we pad O with a zero frame to match the temporal dimension of the input video, forming $O \in \mathbb{R}^{F \times 3 \times H \times W}$. This flow representation is then compressed into a latent space using the VAE encoder: $z_o = \mathcal{E}(O)$. Note that we use "synthetic" depth and optical flow data from open source predictive models, since obtaining ground-truth depth/flow is not possible with large-scale web videos.

To integrate appearance, depth, and optical flow into a unified latent representation, we extend the input and output layers of the pretrained model to accommodate the additional channels for depth and flow. For the RGB input channels, we initialize the model branch from the corresponding pretrained weights to reduce training cost, while the channels corresponding to depth and flow are randomly initialized to allow the model to adapt to these modalities during fine-tuning. As shown on the left of Fig. 2, after encoding, the latent representations of the video, depth, and optical flow are concatenated along the channel dimension to form a joint representation $\mathbf{z} = [z_x, z_d, z_o]$. During training, we add noise to this joint latent at timestep ($t \in [0, 1]$) to obtain $\mathbf{z}_t$. We extend the training objective to jointly predict all three modalities, defined as:

$$\mathcal{L}(\theta) = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_0 \sim \mathcal{N}(0, I), t \in [0,1]} \left[\left| u^+ \left(\mathbf{z}_t, t; \theta \right) - v_t^+ \right|_2^2 \right], \tag{2}$$

where $u^+ = [\hat{z}_x, \hat{z}_d, \hat{z}_o]$ denotes the predicted latents for RGB, depth, and optical flow, respectively, and $v_t^+ = [v_t^x, v_t^d, v_t^o]$ represents the target velocities. The loss is computed across the channels corresponding to the modalities. Unlike the complex sampling-stage guidance in VideoJAM [9], we adopt a simpler approach by simultaneously leveraging RGB and current perceptual conditions to guide future frame generation, as detailed in the next section.

3.3 Long-Horizon Generation Framework

Building on the unified modeling, we now describe how to effectively leverage historical frame contexts, including appearance and perceptual conditions, to achieve stable long video generation. Our design is informed by following key insights: (1) During training with teacher forcing [67], the model is conditioned on ground-truth frames as inputs. However, during inference, it must rely solely on its own previously generated frames. This discrepancy between training and inference can lead to the accumulation of errors over time. To mitigate this issue, we adopt a non-causal sampling head combined with delayed denoising [60], which allows the model to refine its predictions using additional context beyond the immediate past. (2) As shown in Fig. 3, depth prediction exhibits

5

greater temporal stability compared to RGB generation. When the model is trained to output only RGB, we observe rapid degradation in both color and shape over time. In contrast, depth outputs alone show significantly less drift. A key reason for this robustness is that depth is invariant to illumination, shadows, and surface textures, which introduce high variability in RGB signals and are difficult to model consistently over long sequences. This invariance reduces the visual complexity the model must reproduce, thereby slowing down error accumulation. Furthermore, jointly predicting RGB and depth substantially improves RGB quality, suggesting that the geometric constraints provided by depth help stabilize the appearance generation; (3) In contrast to the random per-frame noise scheduling used in Diffusion Forcing [10], we adopt a pre-defined, group-wise noise schedule that more closely aligns training with inference behavior. This structured scheduling strategy reduces training instability and improves overall efficiency by providing more consistent temporal dynamics across samples. Detailed results and comparisons are presented in Sec. 4.2.

Specifically, we partition the inputs to the Diffusion Transformer into two parts: *memory bank* and *prediction groups*. The memory bank consists of both short-term and long-term memory. For short-term memory, we select a small set of historical frames that are adjacent to the current frame. Since high-frequency details are typically reconstructed in the later denoising steps, we set the noise level of these frames to zero, allowing model to preserve fine-grained texture information. For long-term memory, these historical frames contribute less to the current frame's fine details. Although depth perception (as shown in Fig. 3) helps stabilize shape and size, RGB color drift can still occur. To bridge the training-inference gap, we apply a low-level noise t_m to the RGB and optical flow channels of long-term memory, while keeping depth frames un-noised to provide stable global structural guidance. Without perceptual conditions, a high noise level (e.g., $t_m = 0.3$) is typically required for these channels, but this provides limited useful information. During training, t_m is randomly sampled from 0.7 to 0.9 to enhance robustness, and during inference, it is set to 0.8. Note that a higher t_m corresponds to a lower noise level, consistent with our preliminary.

For the prediction groups, we assign noise levels in a structured manner as follows: We divide the input frames into G consecutive groups, where G is the total number of groups. For each training sample, we sample an index i uniformly from the interval $\left[1000 - \frac{1000}{G}, 1000\right)$. For each group k ($k = 0, 1, \ldots, G - 1$), the noise index is defined as:

$$\text{index}_k = i - k \cdot \frac{1000}{G}, \quad i \sim \mathcal{U}\left(1000 - \frac{1000}{G}, 1000\right) \tag{3}$$

The corresponding sigma and timestep for each group are obtained from the noise scheduler using index_k, the training timesteps is set to 1000. This scheduling ensures that the noise intervals are evenly distributed among the G groups, and each group receives noise levels sampled from different segments of the scheduler curve. During inference, the groups are denoised sequentially in a streaming manner, with noise levels decreasing as each group is processed.

4 Experiments

4.1 Experimental Setup

Datasets. To comprehensively evaluate the effectiveness and generalizability of our proposed method, we conduct experiments using two publicly available foundation models: Wan2.1-1.3B [63] (flow-based), trained on an internal general-purpose video dataset at a resolution of 480×832; and CogVideoX-2B [72] (diffusion-based), trained on the in-the-wild DROID robotic manipulation dataset [42] at a resolution of 480×720. The general-purpose dataset contains 300K raw videos, segmented into 1 million short clips, while the DROID dataset consists of 200K successful robotic operation clips. This setup allows us to assess the feasibility of our approach on both flow-based and diffusion-based models, as well as its potential as a long-horizon world model in both general and specialized domains.

Training. All training is conducted on 32 NVIDIA A100 GPUs with a learning rate of 1e-4, using the AdamW optimizer [50], with a per-GPU batch size of 4. The Wan2.1-1.3B model is trained for 50K iterations while the CogVideoX-2B model is trained for 15K steps. To preserve text-to-video performance, we apply the same noise level to every frame in 10% of the training steps.

Evaluation. To assess video generation performance, we utilize VBench [36], a framework that evaluates models across disentangled dimensions, grouping metrics into two categories: consistency
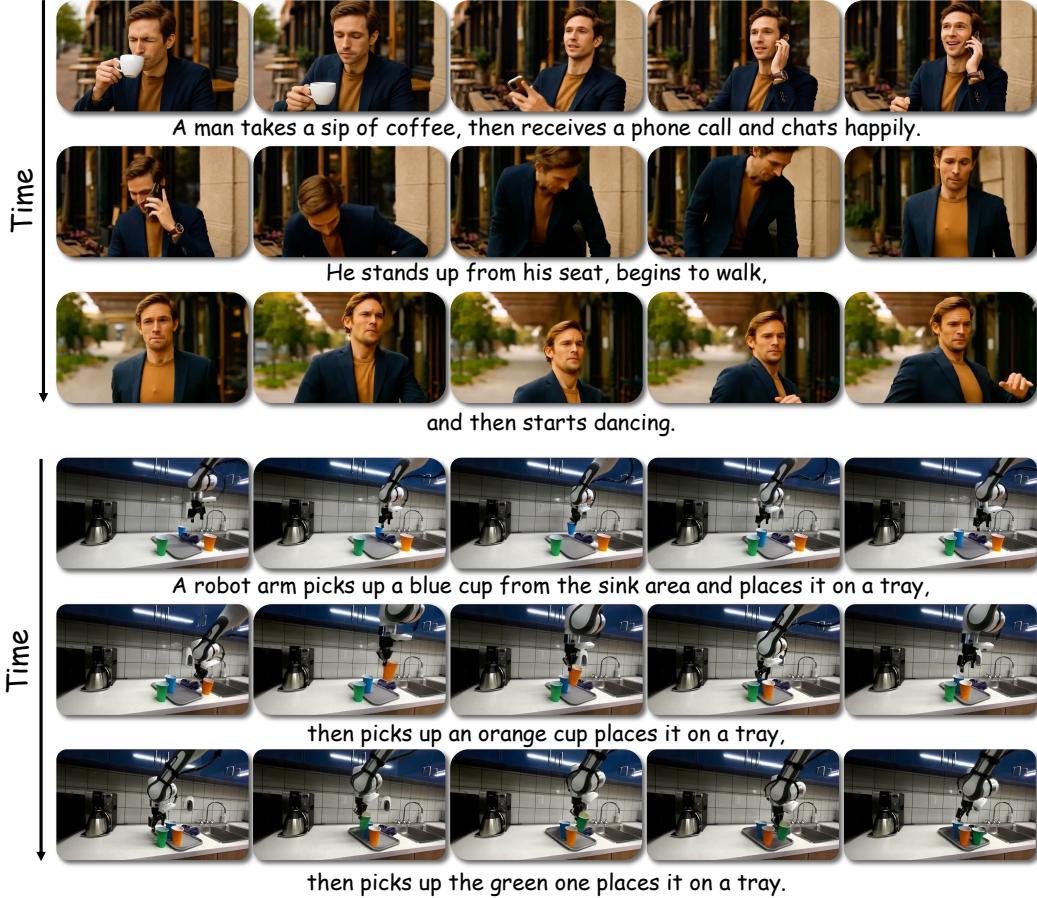
Figure 4: **Generated Long-Horizon Video Results of WorldWeaver.** WorldWeaver demonstrates strong generalization, as it can be applied to various base models. These results highlight not only the effectiveness of our method, but also its potential as a versatile and extensible world model across diverse domains.

and motion. Consistency metrics assess per-frame subject and background coherence, while motion metrics evaluate the amount and temporal coherence of generated motion. We also evaluate a start-end quality contrast metric to measure potential drifting. Specifically, we compute the absolute difference in image quality metric between the first 5 seconds and the last 5 seconds of each video. A larger value of $\Delta_{\text{drift}}^{\text{quality}}$ indicates a greater discrepancy between the beginning and end of the video.

4.2 Comparisons

Visualization of WorldWeaver. We present 30-second and 20-second long video generation results of our model on both general-purpose and robotic manipulation datasets in Fig. 4. Due to space limitations, additional visualizations and comparisons with existing methods are provided in the supplementary materials.

Quantitative experiments. We compare our model with existing open-source methods capable of generating long videos, including StreamingT2V [32], CausVid (Wan) [74], SkyReels-V2 1.3B [11], and Magi 4.5B [59]. Specifically, we use ChatGPT-4o to generate initial images and corresponding text prompts, each containing 4-6 actions, which are used to guide video generation. These prompts correspond to videos of 20-30 seconds in length. As shown in Tab. 1, WorldWeaver surpasses the Magi 4.5B on drift and background consistency metrics. These results are partly influenced by differences in training data and engineering optimizations. In the next section, we provide a more detailed comparison under the same settings. In addition, we conduct a user study to complement the automatic evaluation metrics, with further details available in the supplementary materials.

Table 1: **Comparison of with other open-source models.** Note that the motion degree metric is omitted, as nearly all 20-30 second videos are judged as true for this criterion, making it uninformative.

Methods	Subject consistency	Background consistency	Image quality	Motion smoothness	$\Delta_{\text{drift}}^{\text{Quality}}$
StreamingT2V	79.34	84.32	0.47	0.73	0.23
CausVid	84.71	88.43	0.54	0.82	0.16
SkyReels-V2	87.05	90.07	0.61	0.86	0.11
Magi	**88.71**	90.32	**0.62**	**0.89**	0.09
Ours	87.34	**90.49**	0.59	0.87	**0.07**

Discussion of long context design. Due to differences in the dataset, model architectures, training resources, and the use of post-processing techniques for long videos among existing open-source models, direct comparisons are not entirely fair. Therefore, We evaluate several relevant strategies for long-video generation using the same base model and dataset: *Repeating image-to-video (I2V)*, which applies zero noise to the first frame and identical random noise to all subsequent frames during training, then concatenates the model's final output back into the input sequence to extend video length; *CausVid* [74], which adapts the standard full-attention mechanism to a causal-only attention variant using the official implementation; *History Diffusion* [60], which feeds fully noised historical frames into the unconditional branch of classifier-free guidance to reinforce long-term memory; *Rolling Diffusion* [58], which applies a linearly decreasing, frame-wise noise schedule throughout the sequence; *Diffusion Forcing* [10], which applies random noise to each frame during training and delays the denoising of past frames during inference; and *Ours (RGB-only)* and *Ours (RGB+Perceptual)*, which use the memory bank and the prediction groups to model either RGB information alone or RGB and perceptual channels jointly. Since FramePack [76] requires pre-generating the first and last frames, it is not well suited for scenarios with substantial changes, such as robotic manipulation or dynamic scenes, and is thus not considered in our evaluation. It is worth noting that we also discuss the convergence time required for each algorithm, defined as the training duration after which further improvements in model performance are no longer observed.

Table 2: **Comparison with relevant long-horizon methods.** We compare with other relevant methods across the global metrics, drifting metric. The tests are conducted with CogVideoX-2B on the robotic manipulation dataset, with all experiments performed using 16 A100 GPUs.

Methods	Subject consistency	Background consistency	Image quality	Motion smoothness	$\Delta_{\text{drift}}^{\text{Quality}}$	Training steps
I2V	90.23	91.36	0.51	0.64	0.24	-
CausVid	90.43	91.56	0.53	0.67	0.19	-
History Diffusion	89.78	91.92	0.57	0.68	0.09	70K
Rolling Diffusion	88.74	91.09	0.58	0.70	**0.05**	20K
Diffusion Forcing	88.67	91.23	**0.61**	0.72	0.06	80K
Ours *w/o* perceps	89.57	91.51	0.55	0.67	0.15	30K
Ours	**90.92**	**92.39**	0.60	**0.75**	0.07	30K

As depicted in Tab. 2, our key findings are as follows: (1) Our approach achieves the best performance on most metrics, particularly in terms of consistency, while performing comparably to the best methods on other metrics. (2) Training with independent random noise schedules for each frame, which is adopted by Diffusion Forcing and History Diffusion, requires more training steps for convergence. In contrast, WorldWeaver reduced the training time with the improved training strategies with aligned training and inference noise schedules. (3) While delayed denoising of historical frames mitigates drift, the lack of clean historical frames as guidance compromises video consistency. (4) Compared to RGB-only modeling, unified modeling with perceptual conditions enhances per-frame prediction accuracy, allowing the historical frames of the memory bank to maintain a low noise level, thereby providing effective guidance for subsequent predictions and improving overall consistency.

4.3 Ablation studies

Impact of perceptual conditions. We assess the contributions of various perceptual conditions to improving different aspects of video generation. In addition to the video depth and optical flow

adopted in our final framework, we also explore joint modeling with video segmentation, which is derived using SAM2 [55] by retaining up to 10 masks with the highest confidence scores for each video. To eliminate the influence of drift, we conduct these experiments on short 5-second videos.

Table 3: **Effects of various perceptions.** We evaluate the impact of combining RGB with individual perceptual conditions, as well as jointly using multiple perceptual conditions together.

Perceptions	Subject consistency	Background consistency	Dynamic degree	Motion smoothness
-	92.72	94.76	0.62	0.74
w/ depth	94.87	95.14	0.70	0.82
w/ seg	94.01	94.83	0.65	0.70
w/ flow	92.48	94.22	**0.74**	0.81
w/ depth & seg	**94.93**	95.17	0.68	0.77
w/ depth & flow	94.76	**95.23**	0.73	**0.85**

As shown in Tab. 3, we find that video depth improves both consistency and motion, while optical flow mainly enhances motion. Segmentation contributes more to consistency but slightly reduces motion quality. Since both depth and segmentation capture similar shape and size information, combining them does not further improve consistency. In contrast, jointly modeling depth and optical flow leads to better motion performance. Compared to VideoJAM [9], which relies primarily on optical flow with complex sampling guidance, our approach demonstrates that conditioning on video depth yields greater improvements in temporal coherence and structural fidelity, with optical flow providing complementary benefits.

Noise level of memory bank. We further analyze the impact of the noise levels applied to the long-term memory. As demonstrated in Tab. 4, when perception conditions are not applied, reducing the noise level in the memory bank improves the overall temporal continuity of the video, mitigating abrupt changes in certain frames. However, due to the gap between the generated video during inference and the ground-truth video during training, accumulated errors exacerbate video drift. In contrast, incorporating perception conditions enhances model robustness, allowing clearer past frames to be input into the model as guidance with much lower quality drift. We randomly vary the noise level t_m for RGB and flow latents between 0.7 and 0.9 during training to enhance robustness, and set it to a fixed value of 0.8 during inference. Note that a higher t_m corresponds to a lower noise level, consistent with our preliminary.

Table 4: **Ablation study on noise level.** Adding noise to the memory bank reduces drift but reduces consistency. With perceptual conditions, the model is more robust to drift and less sensitive to noise level. The tests are conducted with CogVideoX-2B on the robotic manipulation dataset.

t_m	Subject consistency		Background consistency		Image quality		Motion smoothness		$\Delta_{drift}^{Quality}$	
	w/o	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*	*w/o*	*w/*
0.9	89.64	**91.78**	92.17	**92.56**	0.52	0.60	0.67	0.73	0.19	0.07
0.7	89.23	90.17	91.45	92.37	0.56	**0.62**	0.69	**0.76**	0.15	0.06
0.3	88.49	88.98	90.62	91.34	0.60	0.61	0.70	0.75	0.10	**0.04**

Ablation on number of groups and length of memory bank. To further analyze the impact of memory bank length and the number of prediction groups, we conduct an ablation study on these hyperparameters. Let F denote the total number of input latent frames, G the number of prediction groups, N the number of frames per group, and M the memory bank length. These parameters are related by $N \cdot G + M = F$. We fix $F = 21$ for all settings, in order to fix the computational cost. Within the memory bank, one frame is reserved for short-term memory (fully denoised frame), while the remaining $M - 1$ frames are dedicated to long-term memory. As shown in Tab. 5, considering the overall performance across all metrics as well as the number of frames generated under the same computation, we set $M = 5$ and $N = G = 4$ in our final training configuration. This experiment is conducted on the robotic manipulation dataset using 16 A100 GPUs, consistent with the ablation studies in the main paper, including the discussion on methods for long videos and the analysis of

Table 5: **Ablation study.** Notations such as `m1_g4_n5` indicate a memory bank length $M = 1$, prediction groups $G = 4$, and frames per group $N = 5$, and so forth, while `Frameoutput` refers to the number of video frames produced by a single complete denoising pass.

Setting	Subject consistency	Background consistency	Image quality	Motion smoothness	$\Delta_{drift}^{Quality}$	frame output
`m1_g4_n5`	88.85	91.30	0.61	**0.75**	**0.05**	20
`m5_g4_n4`	90.92	92.39	0.60	**0.75**	0.07	16
`m13_g4_n2`	**91.04**	**92.51**	0.58	0.74	0.09	8
`m5_g8_n2`	90.64	91.87	**0.62**	0.73	0.06	16
`m1_g2_n10`	90.55	91.23	0.56	0.68	0.12	20

noise levels in the memory bank. For these analyses, we extract the first frame from videos where DROID operations fail, using these frames as input images. This subset of data is entirely unseen during training. Based on these first frames, we employ ChatGPT o3 to generate 30 prompts for 15–20 second robotic operations, each containing 3–5 action commands (e.g., move, grasp, switch, push/pull). For the ablation study on the contribution of perceptual conditions, we use 60 short 5-second videos to isolate the effects from long video frameworks.

Drift resistance experiment details. To investigate the drift resistance of perceptual conditions and color information, we conduct the following experiment after model training. For the depth-only and RGB-only variants, we retain only the weights corresponding to the relevant input and output channels, while loading all other parameters from the fully trained model. We then select a single image or depth map and replicate it across 81 frames (to match the model's input sequence length), appending "static" to the training caption. This process generates 10,000 such samples, which we fine-tune for 1,000 steps each, enabling the model to perform a simple task: consistently outputting static images in a streaming manner. Notably, since optical flow represents motion between frames and lacks temporal continuity as a signal, its isolated output lacks meaningful consistency, so we exclude it from this experiment. We assess the drift resistance by measuring the normalized Mean Squared Error (MSE) between the subsequently generated frames and the first generated frame. As shown in Fig. 5, depth output alone demonstrates superior drift resistance. Jointly outputting both RGB and depth mitigates the drift phenomenon in RGB.
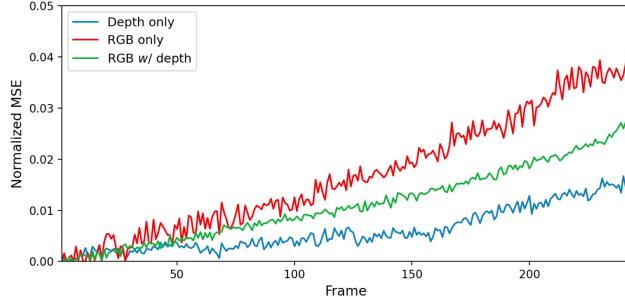
Figure 5: **Normalized mse for drift resistance.**

5 Conclusion

We present WorldWeaver, a framework for long video generation that jointly model RGB information and perceptual conditions under a unified long-context modeling pipeline. We find that jointly predicting RGB and perceptual signals leads to improvements in both consistency and motion quality. Leveraging perceptual conditions, especially depth, which is more resistant to drift than color, enables us to better preserve historical context and enhance temporal consistency. Furthermore, by adopting non-causal prediction groups and a group-wise noise strategy that ensures alignment between training and inference, we are able to alleviate drift and reduce the overall training cost. Extensive experiments on robotic-manipulation and in-the-wild datasets show that WorldWeaver improves long-horizon stability and visual fidelity. It achieves superior consistency and drift mitigation, confirming its effectiveness.

References

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

[2] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos J Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *Advances in Neural Information Processing Systems*, 37:58757–58791, 2024.

[3] Kevin Black, Mitsuhiko Nakamoto, Pranav Atreya, Homer Walke, Chelsea Finn, Aviral Kumar, and Sergey Levine. Zero-shot robotic manipulation with pretrained image-editing diffusion models. *arXiv preprint arXiv:2310.10639*, 2023.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023.

[5] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.

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

[7] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. *arXiv preprint arXiv:2411.00769*, 2024.

[8] Chi-Lam Cheang, Guangzeng Chen, Ya Jing, Tao Kong, Hang Li, Yifeng Li, Yuxiao Liu, Hongtao Wu, Jiafeng Xu, Yichu Yang, et al. Gr-2: A generative video-language-action model with web-scale knowledge for robot manipulation. *arXiv preprint arXiv:2410.06158*, 2024.

[9] Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

[10] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *NIPS*, 37:24081–24125, 2024.

[11] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, Weiming Xiong, Wei Wang, Nuo Pang, Kang Kang, Zhiheng Xu, Yuzhe Jin, Yupeng Liang, Yubing Song, Peng Zhao, Boyuan Xu, Di Qiu, Debang Li, Zhengcong Fei, Yang Li, and Yahui Zhou. Skyreels-v2: Infinite-length film generative model, 2025. URL https://arxiv.org/abs/2504.13074.

[12] Shoufa Chen, Chongjian Ge, Yuqi Zhang, Yida Zhang, Fengda Zhu, Hao Yang, Hongxiang Hao, Hui Wu, Zhichao Lai, Yifei Hu, et al. Goku: Flow based video generative foundation models. *arXiv preprint arXiv:2502.04896*, 2025.

[13] Karan Dalal, Daniel Koceja, Gashon Hussein, Jiarui Xu, Yue Zhao, Youjin Song, Shihao Han, Ka Chun Cheung, Jan Kautz, Carlos Guestrin, et al. One-minute video generation with test-time training. *arXiv preprint arXiv:2504.05298*, 2025.

[14] Etched Decart, Q McIntyre, S Campbell, Xinlei Chen, and R Wachen. Oasis: A universe in a transformer, 2024.

[15] Zihan Ding, Amy Zhang, Yuandong Tian, and Qinqing Zheng. Diffusion world model: Future modeling beyond step-by-step rollout for offline reinforcement learning. *arXiv preprint arXiv:2402.03570*, 2024.

[16] Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Josh Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. *NIPS*, 36:9156–9172, 2023.

[17] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. Vivid: Video virtual try-on using diffusion models. *arXiv preprint arXiv:2405.11794*, 2024.

[18] Zixun Fang, Kai Zhu, Zhiheng Liu, Yu Liu, Wei Zhai, Yang Cao, and Zheng-Jun Zha. Panoramic video generation with pretrained diffusion models. *arXiv preprint arXiv:2506.23513*, 2025.

[19] Ruili Feng, Han Zhang, Zhantao Yang, Jie Xiao, Zhilei Shu, Zhiheng Liu, Andy Zheng, Yukun Huang, Yu Liu, and Hongyang Zhang. The matrix: Infinite-horizon world generation with real-time moving control. *arXiv preprint arXiv:2412.03568*, 2024.

[20] Ruiyuan Gao, Kai Chen, Bo Xiao, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrivedit: High-resolution long video generation for autonomous driving with adaptive control. *arXiv preprint arXiv:2411.13807*, 2024.

[21] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *arXiv preprint arXiv:2405.17398*, 2024.

[22] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144, 2020.

[23] Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

[24] 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.

[25] Gunshi Gupta, Karmesh Yadav, Yarin Gal, Dhruv Batra, Zsolt Kira, Cong Lu, and Tim GJ Rudner. Pre-trained text-to-image diffusion models are versatile representation learners for control. *NIPS*, 37: 74182–74210, 2024.

[26] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

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

[28] Danijar Hafner, Timothy Lillicrap, Mohammad Norouzi, and Jimmy Ba. Mastering atari with discrete world models. *arXiv preprint arXiv:2010.02193*, 2020.

[29] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

[30] Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, and Xuelong Li. Learning an actionable discrete diffusion policy via large-scale actionless video pre-training. *arXiv preprint arXiv:2402.14407*, 2024.

[31] Yingqing He, Tianyu Yang, Yong Zhang, Ying Shan, and Qifeng Chen. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*, 2022.

[32] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.

[33] Daniel Ho, Kanishka Rao, Zhuo Xu, Eric Jang, Mohi Khansari, and Yunfei Bai. Retinagan: An object-aware approach to sim-to-real transfer. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10920–10926. IEEE, 2021.

[34] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

[35] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[36] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[37] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023.

[38] 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.

[39] Anssi Kanervisto, Dave Bignell, Linda Yilin Wen, Martin Grayson, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Tabish Rashid, Tim Pearce, Yuhan Cao, et al. World and human action models towards gameplay ideation. *Nature*, 638(8051):656–663, 2025.

[40] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

[41] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, pages 8110–8119, 2020.

[42] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, Peter David Fagan, Joey Hejna, Masha Itkina, Marion Lepert, Yecheng Jason Ma, Patrick Tree Miller, Jimmy Wu, Suneel Belkhale, Shivin Dass, Huy Ha, Arhan Jain, Abraham Lee, Youngwoon Lee, Marius Memmel, Sungjae Park, Ilija Radosavovic, Kaiyuan Wang, Albert Zhan, Kevin Black, Cheng Chi, Kyle Beltran Hatch, Shan Lin, Jingpei Lu, Jean Mercat, Abdul Rehman, Pannag R Sanketi, Archit Sharma, Cody Simpson, Quan Vuong, Homer Rich Walke, Blake Wulfe, Ted Xiao, Jonathan Heewon Yang, Arefeh Yavary, Tony Z. Zhao, Christopher Agia, Rohan Baijal, Mateo Guaman Castro, Daphne Chen, Qiuyu Chen, Trinity Chung, Jaimyn Drake, Ethan Paul Foster, Jensen Gao, David Antonio Herrera, Minho Heo, Kyle Hsu, Jiaheng Hu, Donovon Jackson, Charlotte Le, Yunshuang Li, Kevin Lin, Roy Lin, Zehan Ma, Abhiram Maddukuri, Suvir Mirchandani, Daniel Morton, Tony Nguyen, Abigail O'Neill, Rosario Scalise, Derick Seale, Victor Son, Stephen Tian, Emi Tran, Andrew E. Wang, Yilin Wu, Annie Xie, Jingyun Yang, Patrick Yin, Yunchu Zhang, Osbert Bastani, Glen Berseth, Jeannette Bohg, Ken Goldberg, Abhinav Gupta, Abhishek Gupta, Dinesh Jayaraman, Joseph J Lim, Jitendra Malik, Roberto Martín-Martín, Subramanian Ramamoorthy, Dorsa Sadigh, Shuran Song, Jiajun Wu, Michael C. Yip, Yuke Zhu, Thomas Kollar, Sergey Levine, and Chelsea Finn. Droid: A large-scale in-the-wild robot manipulation dataset. 2024.

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

[44] Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *arXiv preprint arXiv:2402.08268*, 2024.

[45] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023.

[46] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Customizable image synthesis with multiple subjects. *Advances in neural information processing systems*, 36:57500–57519, 2023.

[47] Zhiheng Liu, Ka Leong Cheng, Qiuyu Wang, Shuzhe Wang, Hao Ouyang, Bin Tan, Kai Zhu, Yujun Shen, Qifeng Chen, and Ping Luo. Depthlab: From partial to complete. *arXiv preprint arXiv:2412.18153*, 2024.

[48] Zhiheng Liu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jie Xiao, Kai Zhu, Nan Xue, Yu Liu, Yujun Shen, and Yang Cao. Infusion: Inpainting 3d gaussians via learning depth completion from diffusion prior. *arXiv preprint arXiv:2404.11613*, 2024.

[49] Zhiheng Liu, Ka Leong Cheng, Xi Chen, Jie Xiao, Hao Ouyang, Kai Zhu, Yu Liu, Yujun Shen, Qifeng Chen, and Ping Luo. Manganinja: Line art colorization with precise reference following. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5666–5677, 2025.

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[51] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. In *ECCV*, pages 329–345. Springer, 2024.

[52] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Freelong: Training-free long video generation with spectralblend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024.

[53] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *arXiv preprint arXiv:2310.15169*, 2023.

[54] Kanishka Rao, Chris Harris, Alex Irpan, Sergey Levine, Julian Ibarz, and Mohi Khansari. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11157–11166, 2020.

[55] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

[56] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. *arXiv preprint arXiv:2303.07109*, 2023.

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[58] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. *arXiv preprint arXiv:2402.09470*, 2024.

[59] Sand-AI. Magi-1: Autoregressive video generation at scale, 2025. URL https://static.magi.world/static/files/MAGI_1.pdf.

[60] Kiwhan Song, Boyuan Chen, Max Simchowitz, Yilun Du, Russ Tedrake, and Vincent Sitzmann. History-guided video diffusion. *arXiv preprint arXiv:2502.06764*, 2025.

[61] Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. Video-infinity: Distributed long video generation. *arXiv preprint arXiv:2406.16260*, 2024.

[62] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. *arXiv preprint arXiv:2408.14837*, 2024.

[63] 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*, 2025.

[64] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.

[65] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *NIPS*, 36:7594–7611, 2023.

[66] Yihan Wang, Lahav Lipson, and Jia Deng. Sea-raft: Simple, efficient, accurate raft for optical flow. *arXiv preprint arXiv:2405.14793*, 2024.

[67] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280, 1989.

[68] Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024.

[69] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, et al. Generalized predictive model for autonomous driving. In *CVPR*, pages 14662–14672, 2024.

[70] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024.

[71] Mingyu Yang, Junyou Li, Zhongbin Fang, Sheng Chen, Yangbin Yu, Qiang Fu, Wei Yang, and Deheng Ye. Playable game generation. *arXiv preprint arXiv:2412.00887*, 2024.

[72] 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. *arXiv preprint arXiv:2408.06072*, 2024.

[73] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, et al. Nuwa-xl: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023.

[74] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv preprint arXiv:2412.07772*, 2024.

[75] 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*, 2025.

[76] Lvmin Zhang and Maneesh Agrawala. Packing input frame context in next-frame prediction models for video generation. *arXiv preprint arXiv:2504.12626*, 2025.

[77] Deyu Zhou, Quan Sun, Yuang Peng, Kun Yan, Runpei Dong, Duomin Wang, Zheng Ge, Nan Duan, Xiangyu Zhang, Lionel M Ni, et al. Taming teacher forcing for masked autoregressive video generation. *arXiv preprint arXiv:2501.12389*, 2025.

[78] Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: Learning interactive real-robot action simulators. *arXiv:2406.12802*, 2024.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (12 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims we made.

   Guidelines:
   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in the supplemental material.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper is not theoretical.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Ours results is reproducible, and we will release the codes.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release the codes and in supplementary materials

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so No is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specify all the training and test details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The methodologies employed in this work involve computational models that yield consistent and repeatable outputs without variability under the same conditions. These tests are based on established simulations that deterministically produce the same results each time they are run, provided the input parameters remain unchanged.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Sufficient information on the computer resources is provided in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We understand the importance of ethical standards in research and take this matter seriously.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We discuss the potential societal impacts of our work in the supplemental material.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: We commit to continuously monitoring the usage of our released models and codes and will take action to restrict access or provide additional guidance if we identify concerning patterns of misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All the assets we used have been properly cited.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

21

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

The content of this supplementary PDF is organized as follows:

- Provided more visualizations and comparisons.
- Systematically conducted a user study to compare our method with open-source models.
- Additional ablation and analysis experiments.
- Discussion on limitations and broader impact.

Contents

6 Additional Visual Results

6.1 More visualizations

As shown in Fig. 6, we present additional videos generated by WorldWeaver. Additionally, we recommend accessing the *viewer.html* file for a visual comparison between our approach and current state-of-the-art methods

7 User Study

To conduct a comprehensive comparison, we evaluate our model (based on Wan2.1 1.3B [63]) against recently released state-of-the-art long-video models of similar scale, including SkyReels-V2 1.3B [11] and MAGI 4.5B [59]. Specifically, we use 48 prompts, each containing 4–6 actions, to generate videos lasting 20–30 seconds. All prompts are provided in the prompts.txt file. We engage 15 annotators to complete a questionnaire, with the UI screenshot presented in Fig. 7. The questionnaire comprises five questions: (1) Which video has the highest overall image quality? (2) Which video exhibits better consistency (considering both subject and background)? (3) Which video shows the smallest difference between the first 5 seconds and the last 5 seconds (in terms of quality and

The two people walk towards each other,

then they have a conversation,

they each raise one hand and put them together, and the girl smiles happily.

A college student studies in a campus cafe, taking notes in his textbook,

checking his laptop, glancing at messages on his phone,

and finally packing up his things when a friend arrives.

A young man jogs around a peaceful lake at dawn,

he stops to catch his breath and stretch,

he takes a photo of the sunrise, then continues running.

Figure 6: **More visualizations.**

consistency)? (4) Which video has smoother motion? (5) Which video aligns best with the overall action described in the prompt.

Results are shown in Fig. 8. Our method demonstrates performance comparable to MAGI in motion smoothness and consistency metrics, while achieving superior results in quality drift. However, a slight gap in image quality remains, with overall performance generally surpassing SkyReels-V2. These findings are generally consistent with our measurements on the VBench benchmark.
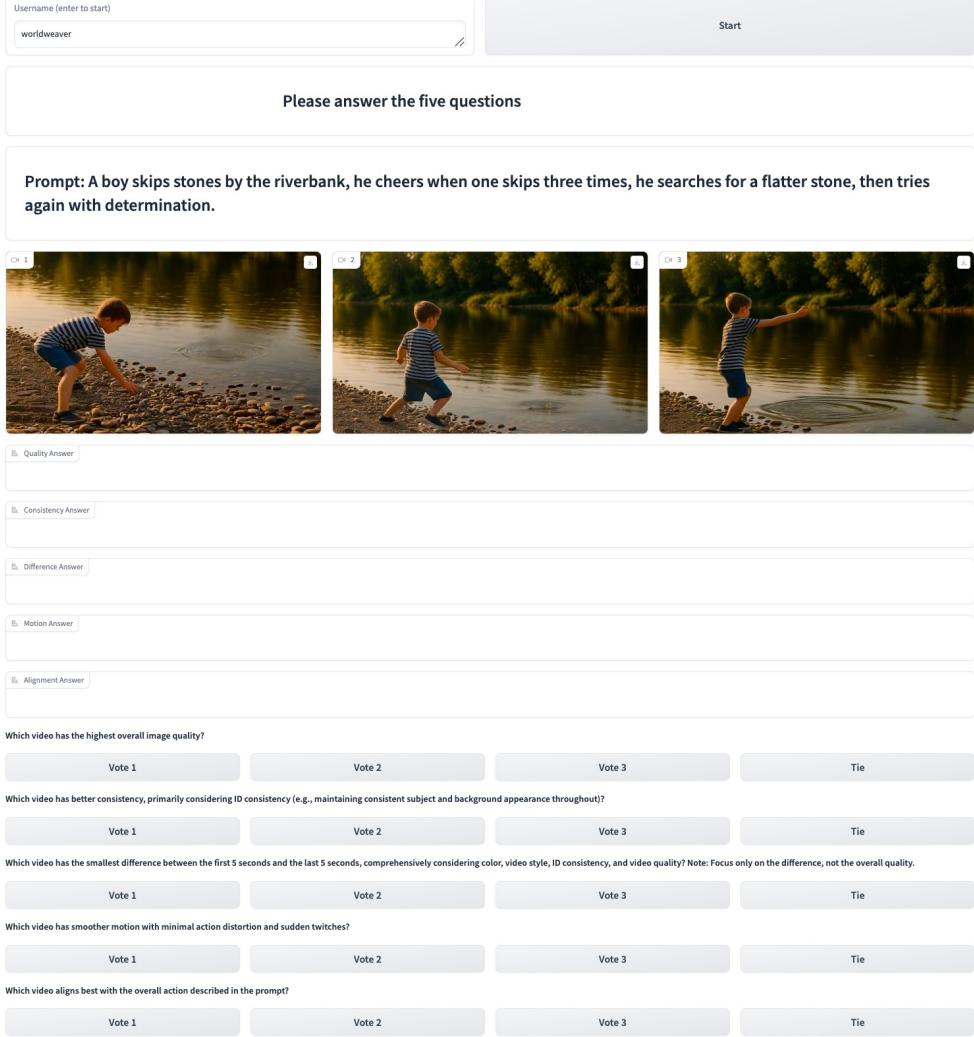
Figure 7: **Visualization of the user study interface.**

7.1 Limitations and future works

Limitations and future works. Despite its strengths, our work has several limitations. First, learning stable physical dynamics from the complexity of the real world remains a long way off, and our model is far from perfect. Operations involving very small objects can still exhibit sudden disappearances, since even depth-based cues cannot fully capture these small objects. Second, although fine-tuning existing models allows us to generate 10∼20s videos, the success rate of generation still decreases as the video length increases, and error accumulation persists over longer horizons, which we leave as an important direction for future work. Finally, while video depth is shown to be the most effective perceptual condition in our experiments, investigating additional or complementary cues to further exploit the rich information in real-world data also represents a promising avenue for further study.
Broader impacts. As video generation technology continues to advance, the development of robust authentication and forgery detection methods becomes increasingly critical. The ability to create highly realistic videos, such as those generated by our model, underscores the need for parallel advancements in counterfeit identification to mitigate potential misuse, ensuring the integrity of digital content in an era of rapid technological evolution.

25
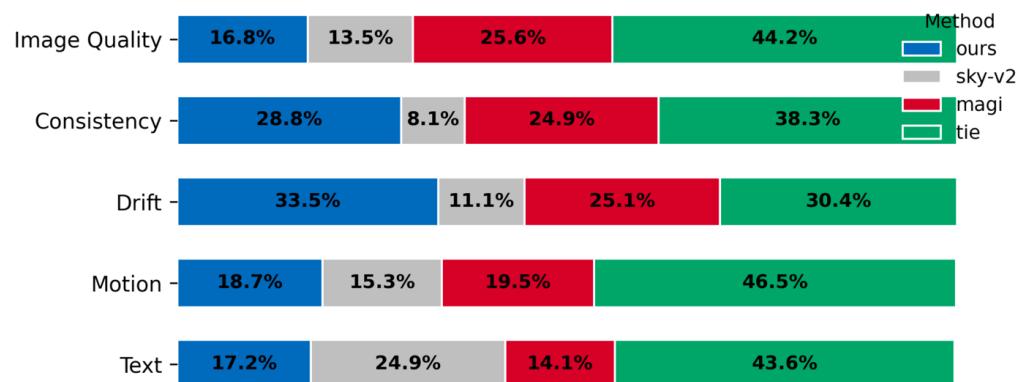
Figure 8: **Results of user study.**