ScalingNoise: Scaling Inference-Time Search for Generating Infinite Videos

Haolin Yang[1,3*], Feilong Tang[1,2,3*], Ming Hu[1,2,3], Yulong Li[1,3], Junjie Guo[4],
Yexin Liu[5], Zelin Peng[6], Junjun He[3], Zongyuan Ge[2†], Imran Razzak[1†]
[1]MBZUAI, [2]Monash University, [3]Shanghai AI Lab, [4]Nanjing University, [5]HKUST, [6]Shanghai Jiaotong University,
Project Page: *https://yanghlll.github.io/ScalingNoise.github.io/*

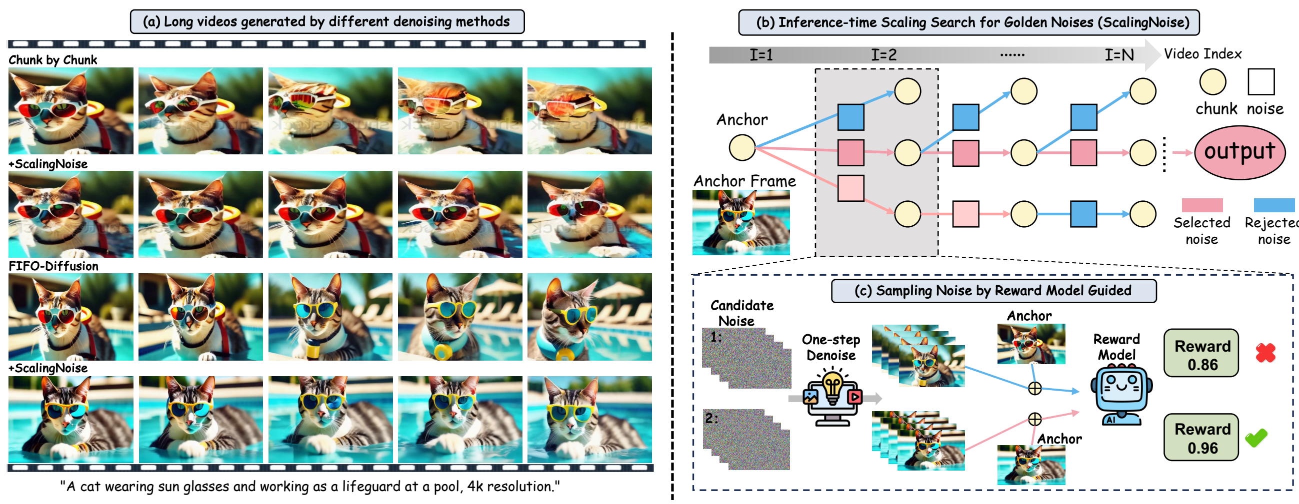


Figure 1. An overview of how ScalingNoise improves long video generation through inference-time search. (a) Chunk-by-chunk and FIFO-Diffusion methods often suffer from accumulated errors and visual degradation over long sequences. (b) ScalingNoise mitigates this by conducting a tailored step-by-step beam search for suitable initial noises, guided by a reward model that incorporates an anchor frame to ensure a long-term signal. (c) At each step, we perform one-step denoising on candidate noises to obtain a clearer clip for evaluation; the reward model then predicts the long-term value of each candidate, helping avoid noises that could introduce future inconsistencies.

Abstract

*Video diffusion models (VDMs) facilitate the generation of high-quality videos, with current research predominantly concentrated on scaling efforts during training through improvements in data quality, computational resources, and model complexity. However, inference-time scaling has received less attention, with most approaches restricting models to a single generation attempt. Recent studies have uncovered the existence of "golden noises" that can enhance video quality during generation. Building on this, we find that guiding the scaling inference-time search of VDMs to identify better noise candidates not only evaluates the quality of the frames generated in the current step but also preserves the high-level object features by referencing the anchor frame from previous multi-chunks, thereby delivering long-term value. Our analysis reveals that diffusion models inherently possess flexible adjustments of computation by varying denoising steps, and even a one-step denoising approach, when guided by a reward signal, yields significant long-term benefits. Based on the observation, we propose **ScalingNoise**, a plug-and-play inference-time search strategy that identifies golden initial noises for the diffusion sampling process to improve global content consistency and visual diversity. Specifically, we perform one-step denoising to convert initial noises into a clip and subsequently evaluate its long-term value, leveraging a reward model anchored by previously generated content. Moreover, to preserve diversity, we sample candidates from a tilted noise distribution that up-weights promising noises. In this way, ScalingNoise significantly reduces noise-induced errors, ensuring more coherent and spatiotemporally consistent video generation. Extensive experiments on benchmark datasets demonstrate that the proposed ScalingNoise effectively improves both content fidelity and subject consistency for resource-constrained long video generation.*

[1]Equal Contribution. † Corresponding Authors.

1. Introduction

Long video generation has a significant impact on various applications, including film production, game development, and artistic creation [39, 87, 94]. Compared to image generation [16, 41, 80], video generation demands significantly greater data scale and computational resources due to the high-dimensional nature of video data. This necessitates a trade-off between limited resources and model performance for Video Diffusion Models (VDMs) [35, 48, 65, 82].

Recent VDMs typically adopt two main methods: one is the chunked autoregressive strategy [4, 23, 26, 43, 72], which predicts several frames in parallel conditioned on a few preceding ones, consequently reducing the computational burden, and "diagonal denoising" from FIFO-diffusion [11, 31], which re-plans the time schedule and maintains a queue with progressively increasing noise levels for denoising. However, video generation is induced by both the diffusion strategies and the noise. Variations in the noises can lead to substantial changes in the synthesized video, as even minor alterations in the noise input can dramatically influence the output [51, 89, 115]. This sensitivity underscores that noises affect both the overall content and the subject consistency of long video generation.

The key to enhancing the quality of long video generation lies in identifying "golden noises" for the diffusion sampling process. Recent studies employ the approach of increasing data [8, 17, 88, 98], computational resources [30, 83, 97, 106, 108, 111], and model size [20, 37, 44, 49, 114] to reduce the truncation errors during the sampling process, but these methods often incur substantial additional costs. Conversely, other approaches focus on training-free denoising strategies such as FreeNoise [42, 53, 81, 115] and Gen-L-Video [74]. They aim to enhance the consistency of generated video by refining local denoising processes to mitigate noise-induced discrepancies, thereby ensuring smoother temporal transitions. Recently, in Large Language Models (LLMs), the study on improving their capability has expanded to inference-time [1, 36, 62, 93]. By allocating more compute during inference, often through sophisticated search processes, these works show that LLMs can produce higher-quality and more appropriate responses. As a result, inference-time scaling opens new avenues for improving model performance when additional resources become available after training. Similarly, recent explorations in diffusion models have leveraged the extra inference-time compute to refine noise search and enhance denoising, thereby improving sample quality and consistency [45, 50, 75]. However, while previous works have shown effectiveness, they focus solely on information within a local window, overlooking long-term feedback and accumulated errors. In this study, we argue that scaling inference-time search of VDMs to identify golden noises enhances long-term consistencies in long video generation.

Search	Representative Methods	Type	Advantage
Greedy	Chunk-Wise Generation [107]	Video	E
Tree	Scaling Denoising Steps [45] Steering Generation [60]	Image Image	S E
	ScalingNoise (Ours)	Video	G S E

Table 1. Greedy decoding is efficient but easily trapped in local optima. Tree methods, better for global optimal decisions, is suitable for inference-time scaling. Our ScalingNoise achieves: G lobal-Optimality of solution, S caling to long-range planning, and E fficiency.

To this end, we propose **ScalingNoise**, a plug-and-play inference-time search strategy that identifies golden initial noises by leveraging a reward model to steer the diffusion process, as illustrated in Fig. 1 (b). Specifically, we employ beam search [86] tailored for intermediate steps and mitigate the accumulated error at each step, while progressively selecting the golden initial noises by choosing the initial noises. Moreover, rather than relying solely on the short-term reward of locally noised clips, we predict the long-term consequences of the initial noises to maintain high coherence. A key challenge lies in the impracticality of directly assessing the initial noises, as it typically requires multiple denoising steps to produce a clear image for assessment, resulting in an exponential increase in computational cost. To address this, we introduce a one-step evaluation strategy that employs the predicted clearer clip from the first DDIM step as an efficient proxy of the quality of a fully denoised clip, as illustrated in Fig. 1 (c). Then, the predicted clip is fed into the reward model while preceding image serve as anchor, providing subject contextual information that preserves appearance consistency and supports long-term value estimation beyond the immediate search step. Our integrated search strategy achieves a practical balance between global optimality, scalability, and efficiency, as shown in Table 1. Moreover, while greedy decoding easily becomes trapped in local optima, the tree search strategy retains multiple candidate sequences, thus exploring the search space more comprehensively and enhancing both the quality and diversity of the generated results.

Given limited computational resources, our strategy selects from a finite pool of candidate noises. Moreover, the quality of candidate noises constitutes a critical factor, complementing the robust reward model that provides long-term supervisory signals. To ensure the candidate pool comprises noise that enhances video consistency, we construct it by sampling from a skewed distribution. Specifically, the weight of high-reward samples is increased, while samples are still drawn from the standard normal distribution to preserve diversity. Through this iterative search process, we significantly reduce accumulated errors and avoid inconsistencies arising from the randomness of initial noises.

On multiple benchmarks, we verify the effectiveness of our method. Our main contributions are as follows:

- We propose a plug-and-play inference-time scaling strategy for long video generation by incorporating long-term supervision signals into the reward process.
- We introduce a one-step denoising approach that transforms the evaluation of initial noises into the evaluation of a clearer image without extra computational overhead.
- Extensive experiments demonstrate that our proposed ScalingNoise can be effectively applied to various video diffusion models. It consistently improves the inference quality of generated videos.

2. Related Work

Long viedo generation. Video generation has advanced significantly [3, 59, 69, 95, 112, 113], yet producing high-quality long videos remains challenging due to the scarcity of such data and the high computational resources required [3, 10, 77]. This limits training models for direct long video generation, leading to the widespread use of autoregressive approaches built on pre-trained models [15, 91]. Current solutions fall into two categories: training-based [38, 58, 76, 109] and training-free methods [5, 102, 107, 110]. Training-based methods [19, 84] like NUWA-XL [96] use a divide-and-conquer strategy to generate long videos layer by layer, while FDM [22] and SEINE [12] combine frame interpolation and prediction. Other approaches, such as LDM [4], MCVD [72], Latent-Shift [2], and StreamingT2V [24], incorporate short-term and long-term information as additional inputs. Despite their success, these methods demand high-quality long video data and substantial computation. Training-free methods address these challenges. Gen-L-Video [74] and Freenoise [53] use a chunk-by-chunk approach, linking segments with final frames, but this risks degradation and inconsistency. Freelong [42] blends global and local data via high-low frequency decoupling, while FreeInit [82] refines initial noises for better consistency. FIFO-Diffusion [31] introduces a novel paradigm, reorganizing denoising with a noise-level queue, dequeuing clear frames, enqueuing noise, for efficient, flexible long video generation without training. In this work, we propose a plug-and-play inference-time strategy that can improve the consistency of videos based on these long video generation methods.

Inference-time Search. A variety of inference-time search strategies have been proven crucial in long context generation within the field of LLMs [29, 34, 40, 66, 92, 99]. The advent of DeepSeek-R1 [14] has further advanced inference-time search. By applying various search techniques in the language space, such as controlled decoding [7, 18, 90, 100], best of N [33, 36], and Monte Carlo tree search [67, 73, 79, 105], LLMs achieve better step-level responses, thus enhancing performance. During inference-time search, leveraging a good process reward model (PRM) [13, 27, 70, 78] is essential to determine the quality of the responses. Ma et al. [45] proposed using supervised verifiers as a signal to guide generating trajectory within Diffusion Models (DMs), but did not investigate its impact on video generation inference-time search. Furthermore, some work has preliminarily explored inference-time search in VDMs [68, 69, 85, 104], however, there is still a lack of investigation into the inference-time scaling law and long-term signals in the process of long video generation. In this work, we explore the effectiveness of scaling inference-time budget utilizing beam search to enhance the consistency of generated long videos.

3. Methodology

3.1. Preliminaries: Video Denoising Approach

Long Video Generation. We introduce the formulation of VDM for generating long videos. There are two approaches to create a long video: the chunk-by-chunk method and the FIFO diffusion method. Both paradigms can be seamlessly integrated into the subsequent framework. Consider extending a pre-trained model that generates fixed-length videos into a long video generation model with a distribution of p_θ, based on the two approaches mentioned above. This model processes an input consisting of both prompt and image to generate a video $\mathbf{v} = [v_1, v_2, \dots, v_N]$, where $\mathbf{v}$ consists of N step-level responses. Each step-level response v_i is a frame or video clip of the long video, treated as a sample drawn from a conditional probability distribution:

$$v_i = p_\theta(v_i|\mathbf{z}_{<i}, c), \quad i = 1, 2, \dots, T, \tag{1}$$

where c can be denoted as a single prompt or a prompt, image pair and $\mathbf{z}_{<i} = [v_1, v_2, \dots, v_i]$ is the concatenated video. In the following, we provide the specific step-level formulations for these two paradigms, respectively:

• **Chunk by Chunk**: Chunk-by-chunk is a generation paradigm [47, 74, 103] that operates through a sliding window approach, using the last few frames generated in the previous chunk as the starting point for the next chunk to continuously produce content. In this paradigm, $v_i = \{v_i^f\}_{f=1}^M$ denotes a video clip of a fixed length M produced by the generated model, while $v_{i,t}$ denotes t noise level of the video clip. A chunk-by-chunk step can be formalized as:

$$v_{i,t-1} = \Psi\left(v_{i,t}, t, \boldsymbol{\epsilon}_\theta(v_{i,t}, t, c)\right), \tag{2}$$

where Ψ and $\boldsymbol{\epsilon}_\theta$ denote a sampler such as DDIM and a noise predict network, respectively.

• **FIFO-Diffusion**: Different from the aforementioned process, FIFO-Diffusion [31] introduces a diagonal denoising paradigm [9, 56] by rescheduling noise steps during the inference process. FIFO-Diffusion achieves autoregressive generation by maintaining a queue where the noise level increases step by step. We define the queue as $Q = \{v_{i,t}\}_{t=1}^T$, where t denotes the noise level, and i indicates the i-th frame in the queue. In this paradigm, i is equal to t, and v_i denotes a frame. The length of T is $M \times P$ where P denotes the partition of the queue. The procedure of FIFO step can be described as follows:

$$Q = \Psi\left(Q, \{t\}_{t=1}^{T}, \epsilon_\theta(Q, \{t\}_{t=1}^{T}, c)\right). \tag{3}$$

DDIM. DDIM [63] introduces a new sampling paradigm by de-Markovizing the process, which remaps the time steps of DDPM [25] from $[0; \ldots; T]$ to $[\tau_0; \ldots; \tau_T]$, which is a subset of the initial diffusion scheduler, thereby accelerating the sampling process of DDPM. Here, DDIM(v_{τ_t}) consists of three distinct components, which can be formulated as:

$$\begin{aligned} v_{\tau_{t-1}} &= \mathrm{DDIM}(v_{\tau_t}) \\ &= \alpha_{\tau_{t-1}} \left(\frac{v_{\tau_t} - \sigma_{\tau_t} \epsilon_\theta(v_{\tau_t}, \tau_t)}{\alpha_{\tau_t}} \right) + \sigma_{\tau_{t-1}} \epsilon_\theta(v_{\tau_t}, \tau_t), \end{aligned} \tag{4}$$

where α and σ_{τ_t} denote predefined schedules.

3.2. Formulation of Video Generation Inference

The long video generation framework can be formulated as a Markov Decision Process (MDP) problem defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R})$. $\mathcal{S}$ is the state space. Each state is defined as a combination of the generated video and the condition. The initial state $\boldsymbol{s}_0$ only corresponds to the input prompt x and guided image I. $\boldsymbol{S}_i$ is the combination of the currently generated videos. $\mathcal{A}$ denotes the action space, where each action encompasses a two-part process: sampling an initial noise from a tilted distribution, followed by denoising the current video clip based on the noise. We also have the reward function $\mathcal{R}$ to evaluate the reward of each action, which is also known as the process reward model (PRM) in LLMs. With this MDP modeling, we can search for additional states by increasing the inference-time compute, thereby obtaining a better VDM response $\mathbf{v}$. Specifically, we can take different actions in each state, continuously explore, and then make choices based on the reward model to achieve a better state. The general formulation of the selection process is:

$$\boldsymbol{a}_{t+1} = \arg\max_{\mathcal{A}} \Phi(\boldsymbol{s}_t, \mathcal{A}), \tag{5}$$

where Φ denotes the reward model $\mathcal{R} : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$. The core of our method focuses on efficiently and accurately estimating then selecting initial noises, improving video generation with better guidance.

3.3. Reward Design

During the search process, our objective is to evaluate the consistency and quality of the video at each step, using these

Algorithm 1 ScalingNoise Inference-time Search

Require: Diffusion Model D, Reward Function Φ, Sample Tilted Distribution $Sample$, Condition c, Beam Size k, Step Size n, DDIM Steps τ_t, Generated Video $\boldsymbol{V} = [\,]$ Anchor Frame $\boldsymbol{v}_a$

while Generation is not Done **do**
  for i in [1, 2, ..., k] **do**
    $\boldsymbol{r} = [\,]$
    for j in [1, 2, ..., n] **do**
      $\boldsymbol{\epsilon}_{ij} \leftarrow Sample(\boldsymbol{V})$
      $\hat{\boldsymbol{v}}_{ij} \leftarrow D(\boldsymbol{\epsilon}_{ij}, c, \text{num\_steps} = \tau_t)$
      $r_{ij} \leftarrow \Phi(\hat{\boldsymbol{v}}_{ij}, \boldsymbol{v}_a)$
      $\boldsymbol{r}$.append(r_{ij})
    end for
  end for
  $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k] \leftarrow$ Select the best k elements from $\boldsymbol{r}$
  for i in $[\tau_0, \tau_1, ..., \tau_t]$ **do**
    $\boldsymbol{v} \leftarrow D(\boldsymbol{\epsilon}, c, \text{num\_steps} = i)$
  end for
  $\boldsymbol{v}_a \leftarrow \boldsymbol{v}_{i0}$
  Append current clip $[\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k]$ to $\boldsymbol{V}$
end while
return $\boldsymbol{V}$

insights to guide subsequent searches, as illustrated in Fig. 1 (c). The evaluation of each search step, defined as applying an action a_t to the state $\boldsymbol{s}_t$, is performed by a reward function $r_t = \Phi(s_t, a_t) \in \mathbb{R}$. Below, we elaborate on the specific design of this reward function.

One-Step Denoising. Throughout our evaluation process, the actions in the MDP sample initial noises, which is typically Gaussian noise. Therefore, it is difficult to assess. A key challenge arises from the difficulty of directly assessing this initial noises. To address this, we propose a one-step evaluation approach. Specifically, we utilize the predicted $\hat{v}_{\tau_0}$ component from Eq. 4 above as the target for evaluation. The detailed formulation is presented as:

$$\hat{v}_{\tau_0} = \frac{v_{\tau_t} - \sigma_{\tau_t} \epsilon_\theta(v_{\tau_t}, \tau_t)}{\alpha_{\tau_t}}. \tag{6}$$

Our evaluation method leverages a single DDIM step to accurately and efficiently assess the initial noise. In contrast to the original brute-force approach, which fully denoises the initial noises into a clear video, our technique is less computationally intensive, though it may yield suboptimal results. The brute-force method, while thorough, requires significant computational resources, rendering it impractical for scaling to long video generation.

Consistency Reward. After obtaining an evaluable object, we need to design a long-term reward to evaluate the overall consistency of the video and prevent the accumulation of errors. The reward function requires careful design.

Specifically, we take into account not only the video clip currently being generated by the action but also the states of previous nodes. To this end, we select fully denoised, clear video frames v_a from several frames prior as anchor points and assess the consistency of the video within the current window after one-step denoising. This approach substantially reduces inconsistencies in video generation. In our experiments, we employ a DINO [6] model and calculate the reward using the following formula:

$$\Phi := \frac{1}{2(T-1)} \sum_{i=1}^{M} \left(\langle d_{\mathrm{a}} \cdot d_i \rangle + \langle d_i \cdot d_{i-1} \rangle \right), \tag{7}$$

where d_a and d_i denote the image features of the anchor frame and the i-th frame in the current clip, respectively. And $\langle \cdot \rangle$ is the dot product operation for calculating cosine similarity.

3.4. Action Design

Search Framework. Once equipped with a rewards model (Section 3.3), VDM can leverage any planning algorithm for search, as demonstrated in [21]. We employ beam search (Fig. 1 (b)), a robust planning technique that efficiently navigates the search tree space, effectively balancing exploration and exploitation to identify high-reward trajectories. Specifically, beam search constructs a search tree iteratively, with each node representing a state and each edge denoting an action and the resulting state transition. To steer VDM toward the most promising nodes, the algorithm relies on a state-action reward function, Φ in Eq. 7. To promote diversity among response candidates at each step, we maintain K distinct trajectories. From a tilted distribution, we sample N initial noise instances, generating $K \times N$ candidates for the current step. The reward model evaluates each candidate, and the top-K candidates with the highest scores are selected as responses for that step. This sampling and selection process iterates until the full response sequence is generated. Further details and the pseudo-code for this planning algorithm are provided in Algorithm 1.

Tilted Distribution Sampling. During the search process, we need to sample the initial noises. Due to computational constraints, exhaustively searching the entire noise space is infeasible. To increase the likelihood of generating superior results, we sample initial noise from a tilted distribution [52, 60]. Specifically, we construct a high-quality candidate pool by employing four distinct tilted distributions to sample the initial noises. These operations are detailed as follows:

$\diamond$ ***A1*** *Random Sampling: Directly sample noise from a Gaussian distribution* $v_i \sim \mathcal{N}(0, I)$.

$\diamond$ ***A2*** *FFT Sampling: Utilize 2D/3D Fast Fourier Transform (FFT) [11, 42] to blend Gaussian noise with the last few frames, denoted as:*

$$v_i = \boldsymbol{F}^{r}_{low}(v_i) + \boldsymbol{F}^{r}_{high}(\eta)\ , \tag{8}$$

where $\boldsymbol{F}$ denotes the FFT function, and η is the Gaussian noise.

$\diamond$ ***A3*** *DDIM Inversion: Apply DDIM Inversion to renoise the previous frame. It can be formulated as:*

$$v_{t-1} = \alpha_{t-1} \left(\frac{v_t - \sigma_t \epsilon_\theta(v_t, t)}{\alpha_t} \right) + \sigma_{t-1} \epsilon_\theta(v_t, t). \tag{9}$$

$\diamond$ ***A4*** *Inversion Resampling: Building on DDIM Inversion, sample new noise within its δ-neighborhood defined as* $\{v' : d(v, v') < \delta\}$. Through these strategies, we enhance the quality of candidate noise, enabling us, within limited resources, to maximize the leveraging of actions with higher rewards.

4. Experiment

In this section, we conduct experiments to answer the following two questions: *1. Does the long-term reward guided search yield higher-quality and consistent video compared with other inference-time scaling methods? 2. Does the one-step evaluation provide an efficient and accurate assessment of initial noises?*

4.1. Baseline and Implementation details

To evaluate the effectiveness and generalization capacity of our proposed method, we implement it on the text-to-video FIFO-Diffusion [31] and image-to-video chunk by chunk long video generation, based on existing open-source diffusion models trained on short video clips. These models are limited to producing videos with a fixed length of 16 frames. In our experiments, the evaluations are conducted on an NVIDIA A100 GPU.

Vbench Dataset. Our approach is systematically evaluated using VBench [28], a video generation benchmark featuring 16 metrics crafted to thoroughly evaluate motion quality and semantic consistency. We select 40 representative prompts spanning all categories, generate 100 video frames, and analyze the model's performance accordingly. We choose five VBench evaluation metrics for performance comparison: Subject Consistency, Background Consistency, Motion Smoothness, Temporal Flickering, and Imaging Quality. We maintain k distinct candidates in the beam and sample n completions for each one. Specifically, we set k = 2, 3 for beam size, and n = 5 with FIFO-Diffusion based on VideoCraft2 [10], n = 10 with ConsistI2V [55] chunk by chunk to balance the quality and efficiency. For the I2V model, we use the first image of the video generated by FIFO-Diffusion to guide the video generation. We consider two types of baselines: (1) **Base Model**: This serves as our baseline approach, employing a naive method that deliberately avoids any form of inference-time scaling. (2) **Best of N (BoN)**: A widely adopted technique to improve model response quality during inference. For each prompt,

Method	N_{can}	Subjection Consistency↑	Background Consistency↑	Motion Smoothing↑	Time Flicking↑	Imaging Quality↑	Overall Score↑
Streamingt2v [24]	1	84.03	91.01	96.58	95.47	61.64	85.74
OpenSora v1.1 [32]	1	86.92	93.18	97.50	**98.72**	53.07	85.87
FreeNoise [53]	1	92.30	**95.16**	96.32	94.94	67.14	89.17
ConsistI2V(Chunk-wise) [55]	1	89.57 ↓ +0.00	93.22 ↓ +0.00	97.62 ↓ +0.00	96.63 ↑ +0.00	55.21 ↓ +0.00	86.45 ↓ +0.00
+BoN	3	89.92 ↑ +0.35	93.64 ↑ +0.42	97.59 ↓ −0.03	96.66 ↑ +0.03	55.74 ↑ +0.50	86.71 ↑ +0.26
+BoN	5	90.56 ↑ +0.99	93.59 ↑ +0.37	97.73 ↑ +0.11	96.24 ↓ −0.39	56.24 ↑ +1.03	86.87 ↑ +0.42
+ScalingNoise (Ours)	2	91.58 ↑ +2.01	94.36 ↑ +1.14	97.85 ↑ +0.25	96.79 ↑ +0.16	56.82 ↑ +1.61	87.48 ↑ +1.03
+ScalingNoise (Ours)	3	92.02 ↑ +2.45	94.44 ↑ +1.22	**97.91** ↑ +0.29	96.97 ↑ +0.34	58.12 ↑ +2.91	87.89 ↑ +1.18
FIFO-Diffusion [31]	1	90.26 ↓ +0.00	93.53 ↓ +0.00	95.86 ↓ +0.00	92.78 ↓ +0.00	65.52 ↓ +0.00	87.59 ↓ +0.00
+BoN	3	90.92 ↑ +0.66	94.49 ↑ +0.96	95.20 ↓ −0.66	93.76 ↑ +0.98	64.13 ↓ −1.39	87.70 ↑ +0.11
+BoN	5	91.26 ↑ +1.00	94.91 ↑ +1.38	95.97 ↑ +0.11	93.97 ↑ +1.19	64.37 ↓ −1.15	88.10 ↑ +0.51
+ScalingNoise (Ours)	2	91.60 ↑ +1.34	93.74 ↑ +0.21	96.67 ↑ +0.81	94.96 ↑ +2.18	65.67 ↑ +0.15	88.53 ↑ +0.94
+ScalingNoise (Ours)	3	**93.14** ↑ +2.88	94.61 ↑ +1.08	97.01 ↑ +1.15	95.34 ↑ +2.56	**67.91** ↑ +2.39	**89.60** ↑ +2.01

Table 2. **Quantitative comparison results.** Comparison of performance metrics for various video generation methods as benchmarked by VBench. We calculate the average performance in the last column, demonstrating its effectiveness in producing fidelity and consistent long videos. Bold indicates the highest value, and underlined indicates the second highest.

we generate multiple responses, specifically 3 and 5 distinct outputs. We also select three state-of-the-art methods as baselines, namely StreamingT2V [24], Openasora v1.1 [32], and FreeNoise [53].

UCF-101 Dataset. UCF-101 [64] is a large-scale human action dataset containing 13,320 videos across 101 action classes. We utilize the following metrics:

- Frechet Video Distance (FVD) [71] for temporal coherence and motion realism.
- Inception Score (IS) [57] for frame-level quality and diversity.

We measure the two metrics using Latte [46] as a base model, which is a DiT-based video model trained on UCF-101 [64], employing FIFO-Diffusion as the paradigm for long video generation, configured with k = 2 and m = 5. We generate 2,048 videos with 128 frames each to calculate FVD_{128}, a version of FVD which uses 128-frames-long videos to compute the statistics, and randomly sample a 16-frame clip from each video to measure the IS score, following evaluation guidelines in StyleGAN-V [61]. As the base model, we choose StyleGAN-V, PVDM [101], FIFO-Diffusion, as they are representative open-sourced models.

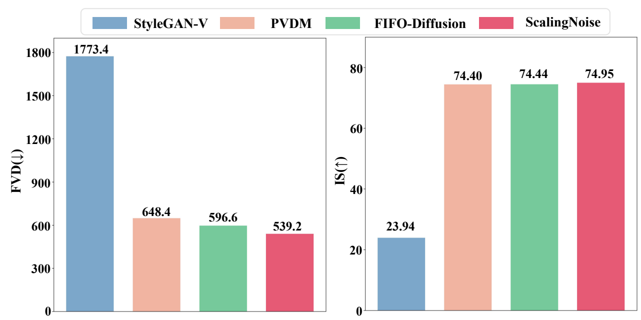
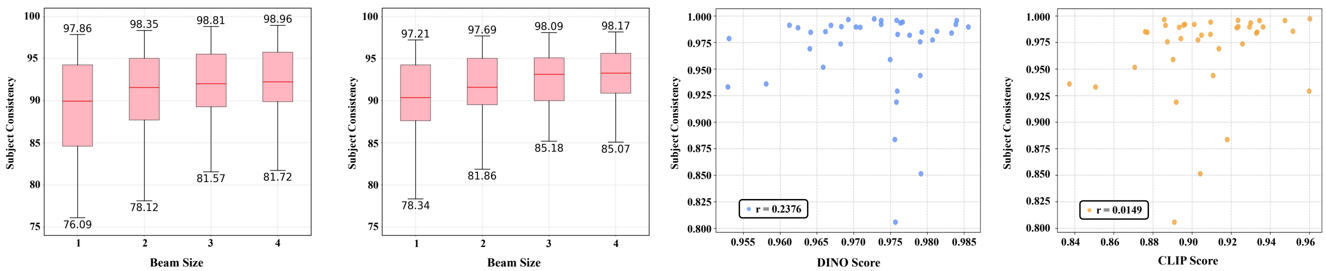


Figure 2. Comparisons of FVD_{128} and IS scores on UCF-101. ScalingNoise utilizes Latte [46] as its baseline, where the number of beam sizes is 2, and noise candidates are 5. The FVD and IS scores of the other algorithms are obtained from their respective papers, and PVDM [101] denotes PVDM-L (400-400s).

4.2. Quantitative results

Scale Search Improves Video Consistency. We compare ScalingNoise with the baselines trained for long video generation in terms of multiple benchmarks. **Vbench**: As shown in Table 2, we find that the videos generated by ScalingNoise are significantly more preferred compared with the baseline. Videos generated by the base model decoding are of the lowest quality. While increasing inference compute via BoN shows improvement, they still fall short compared with ScalingNoise search. The computational cost of BoN(3) is comparable to that of our method when configured with k = 3. In contrast, the computational cost of BoN(5) significantly exceeds that of our approach. Although its consistency has improved, our method outperforms BoN(5) across all evaluated aspects. The long videos obtained using ScalingNoise search not only significantly *ehnance consistency*, but also provide *high quality* video frames. **FVD and IS**: As illustrated in Fig. 2, our approach outperforms all the compared methods including PVDM-L (400-400s) [101], which employs a chunked autoregressive generation strategy. Note that PVDM-L (400-400s) iteratively generates 16 frames conditioned on the previous outputs over 400 diffusion steps.

Benefits from Further Scaling Up Inference Compute. We next explore the effect of increasing inference-time computation on the response quality of the VDM at each



(a) Scaling tendency of search beam size.

(b) Correction of reward score and consistency.

Figure 3. (a) The two figures are boxplots showing the tendency of scaling beam sizes for ScalingNoise based on two paradigms, in order of FIFO-Diffusion and Chunk by chunk. (b) From Left to Right: Correction of reward model DINO and CLIP feature similarity score and final subject consistency. All points are generated by VideoCraft2.

Method	Subjection Consistency↑	Overall Score↑	Inference Time↓
BoN	**97.87**	**92.06**	477.75
10-Step	97.14	91.59	79.67
ScalingNoise (Ours)	97.71	91.83	**12.34**

Table 3. Video consistency and inference times of different evaluation methods. ScalingNoise utilizes one-step evaluation to significantly improve efficiency.

Method	Subjection Consistency↑	Image Quality↑	Overall Score↑
Local	92.16	66.83	88.94
Anchor	92.67	66.39	89.28
ScalingNoise (Ours)	**93.14**	**67.91**	**89.60**

Table 4. **Reward Function Studies.** Video consistency and quality of different reward function guided inference time search.

step by varying the beam sizes. To ensure experimental fairness, we set n = 5 for FIFO-Diffusion, while for chunk-by-chunk methods, we maintain n = 10. This distinction arises because FIFO-Diffusion requires initializing only one initial noise per step, whereas the chunk-wise search process initializes one clip per step. Consequently, after a single DDIM denoising step, the chunk-wise approach needs to select the combination of multiple frames, resulting in a search space complexity significantly larger than that of FIFO-Diffusion. Setting n = 5 would yield an overly constrained search space and fail to align with the computational load of FIFO-Diffusion. We adopt the subject consistency from V-Bench, as mentioned earlier, as our evaluation criteria. We report the scores for long video generation achieved through ScalingNoise search, based on both FIFO-Diffusion and chunk-by-chunk paradigms, with beam sizes set to 1, 2, 3, and 4. The experimental results are illustrated in Fig. 3 (a). Since some prompts are static while others correspond to video actions with very large movements, this results in a significant variance. Our observations reveal that the performance of ScalingNoise, for both FIFO-Diffusion and chunk-wise strategies, improves steadily as the search beam size increases. This trend suggests that scaling inference-time computation effectively enhances the visual consistency capabilities of VDM.

One-Step Evaluation's Efficiency and Accuracy. To evaluate computational efficiency and accuracy of our evaluation method, we examine the VBench evaluation metrics and inference time across different assessment methods. Using the VideoCraft2 [10] model (configured with DDIM=64), we generate videos with a fixed length, adopting 16 frames. We sample 10 candidate initial noises and employ our one-step evaluation method to select one. To ensure a fair comparison of initial noise quality, we fully denoise the selected noise and assess the consistency and quality of the resulting video. In addition, we test two baseline approaches: (1) **BoN**: selection after complete denoising for clarity, and (2) **m-Step Evaluation**: the initial noises are denoised for m steps via DDIM, followed by selection using the same reward model. Here we set m to 10. As shown in Table 3, our one-step evaluation method generates videos in just 12.34 seconds, enabling the assessment and selection of initial noises without compromising baseline performance. In contrast, the other two baselines require 79.67 seconds and 477.75 seconds, respectively. This efficiency allows for scalable search within the recent long video generation paradigm.

4.3. Qualitative results

User Study. We start with human evaluation with results shown in Fig. 4. We utilize generated videos from the evaluation dataset, allowing human annotators to assess and compare the output quality and consistency across different methods. The win rate is then calculated based on their judgments, providing a clear metric for performance comparison. The robust performance of our method, ScalingNoise, underscores its capability to produce videos that are not only more natural and visually coherent but also maintain a high level of consistency throughout. Compared

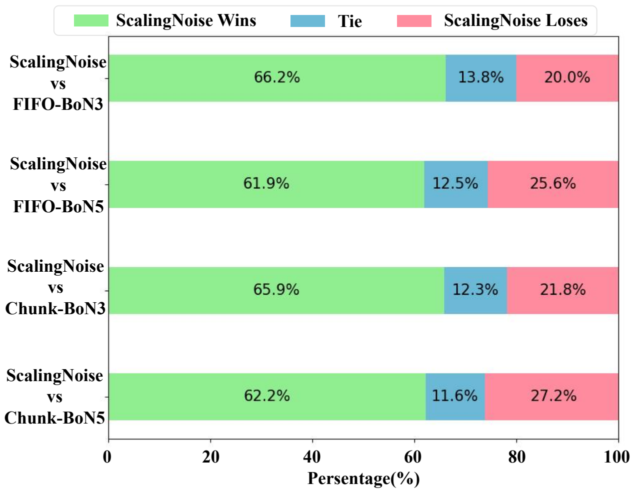


Figure 4. **User Study.** Win rate of videos generated using ScalingNoise compared with other inference-time scaling methods.

to the naive inference time scaling method, BoN, ScalingNoise distinctly showcases its superior efficiency.

Case Study of Search Trajectory. As shown in Fig. 5, ScalingNoise demonstrates a clear search process based on consistI2V, which illustrates how it evolves from starting state (contains a prompt and a guided image) into a complete long video. In each step, ScalingNoise employs a selection, steering by the long-term reward function. For example, in the first step, the reward model assigns a higher score to images where red wine is not spilled, thus avoiding subsequent cumulative errors. At the same time, it can be seen that due to our long-term reward strategy, even if a bad case has already occurred, our method can still make corrections to subsequent frames based on the anchor frame.

4.4. Ablation Study

In this section, we conduct ablation studies to evaluate the impact of each design component in ScalingNoise for long video generation. Unless otherwise specified, all experiments follow previous settings for fair comparison.

Long-Term Reward. First, we present the performance of different variants explored in the design of the reward function based on FIFO. Table 4 details the results of two additional runs, while still using the DINO model, with different reward functions: *(1) local reward*: using only local clip during the denoising process, and *(2) anchor reward*: using only the initial noise and anchor frame, leading to a drop of 0.66% and 0.32%, respectively. The specific calculation formula is as follows:

$$\Phi_{local} = \sum < d_i \cdot d_{i-1} >, \quad \Phi_{anchor} =< d_a \cdot d_n > .$$

As shown in Table 4, ScalingNoise achieves the best performance. As illustrated in Fig. 6, we present videos guided by different reward models. Our reward function not only considers the long-term consistency between the initial noise and anchor frame, but also effectively accounts for the cross-temporal influence of initial noise propagation across video frames within the denoising window.

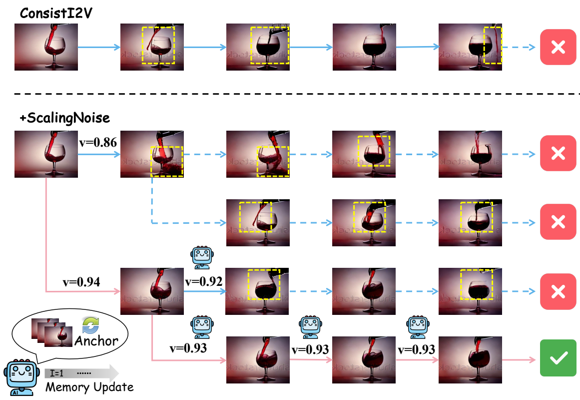


Figure 5. The upper part of this figure represents the original method, which involves a greedy approach to generating long videos. In contrast, the tree-structured searching process of ScalingNoise is outlined below: We highlight the path from the initial state to the best-performing generation, reporting the score for each node. Our prompt is "Red wine is poured into a glass. highly detailed, cinematic, arc shot, high contrast, soft lighting".

Method	Subjection Consistency↑	Image Quality↑	Overall Score↑
Random	92.64 ↓ +0.00	65.98 ↓ +0.00	89.07 ↓ +0.00
2D FFT	92.67 ↑ +0.03	65.79 ↓ −0.19	89.24 ↑ +0.17
Resample	92.94 ↑ +0.30	65.71 ↓ −0.27	88.83 ↓ −0.11
Reverse	**93.27** ↑ +0.63	64.83 ↓ −1.15	88.96 ↓ −0.66
All (Ours)	93.14 ↑ +0.50	**67.91** ↑ +1.93	**89.60** ↑ +0.53

Table 5. **Tilted Distribution studies.** Comparsion of sampling from the different tilted distribution.

Different Reward Model. Then we investigated the effects of using different reward models (*i.e.*, DINO [6] and CLIP [54]) to guide the search process. For each reward model, we generated a 16-frame video and, after one denoising step, scored it using DINO and CLIP, respectively. As shown in Fig. 3 (b), the vertical axis still represents the subject consistency score, while the horizontal axis represents the reward model scores. It can be observed that DINO's similarity scores demonstrate a stronger alignment with the final video's subject consistency compared to CLIP. Specifically, there exists a robust correlation between the consistency of the main subject in the final video and the feature space extracted by DINO, as measured by the cosine similarity of the video frames. In contrast to DINO,

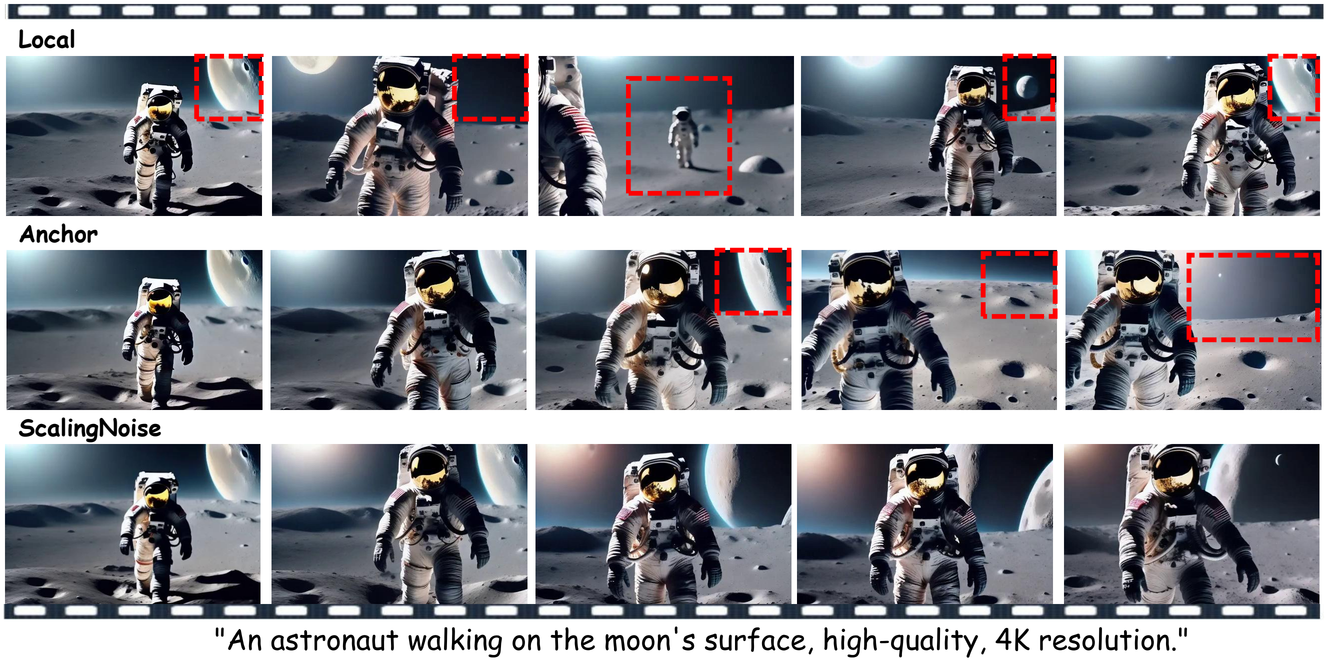


Figure 6. **Example Videos.** Illustrations of long videos guided by different reward function. **Row 1**: Local reward only consider the quality of current denoised clip. **Row 2**: Anchor reward calculate the similarity of the anchor frame and the initial noise. **Row 3**: Ours combines the best of both, achieving long-term reward.

which effectively captures the features of the primary subject in each frame, CLIP tends to focus on extracting the overall features of the background. During video generation, inconsistencies predominantly stem from variations in the subject, while changes in the background remain relatively minor. Consequently, DINO provides a more accurate and reliable evaluation of subject consistency, making it a superior choice over CLIP for this purpose.

Effectiveness of Tilted Distribution. We investigate how each component of the tilted distribution impacts the quality of video generation. Table 5 summarizes the performance results for long video generation. We tested the performance of these sampling distributions separately, including *(1) Random Distribution*, the standard normal Gaussian distribution. *(2) 2D FFT*, which leverages frequency domain analysis to improve generation quality. *(3) DDIM Inversion*, a method that significantly enhances subject consistency across frames. *(4) Inversion Resampling*, an approach introduced to address specific limitations observed in earlier methods. The 2D FFT is an effective method for improving video quality. However, as the generated length increases, it can lead to a degradation in video quality. Although the DDIM reverse markedly enhances the subject consistency, it results in a significant reduction in the range of motion in the generated video. Therefore, we introduce Inversion Resampling to maintain the diversity. Integrating all methods into the base model yields a performance boost of +0.53%.

5. Conclusion & Limitation

In this work, we introduce ScalingNoise, a novel inference-time search strategy that significantly enhances the consistency of VDMs by identifying golden initial noises to optimize video generation. Utilizing a guided one-step denoising process and a reward model anchored to prior content, ScalingNoise achieves superior global content coherence while maintaining high-level object features across multi-chunk video sequences. Furthermore, by integrating a tilted noise distribution, it facilitates more effective exploration of the state space, further elevating generation quality. Our findings show that scaling inference-time computations enhances both video consistency and the quality of individual frames. Experiments on benchmarks validate that ScalingNoise substantially enhances content fidelity and subject consistency in resource-constrained long video generation.

Limitation: We clarify the limitations of our proposed ScalingNoise: *(i):* Our ScalingNoise may struggle with scenes involving highly complex or abrupt motion, where accurate alignment across frames becomes challenging, potentially affecting temporal coherence. *(ii)*: ScalingNoise cannot completely eliminate accumulated error, while through long-term signal guidance, it can to some extent alleviate this phenomenon. To entirely address this issue, we need to conduct an extra in-depth analysis of the causes of error accumulation.

6. VideoCrafter2

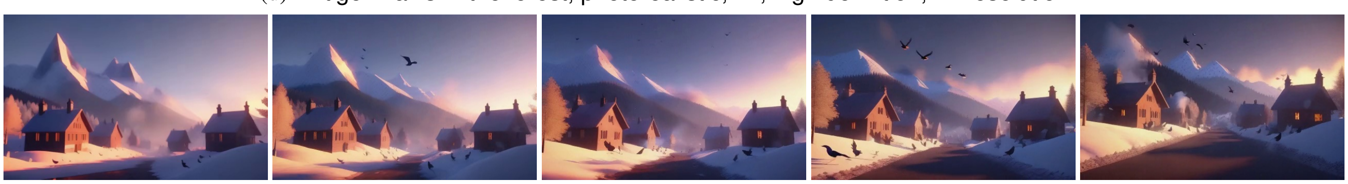
In Fig. 7 and Fig. 8, we provide more qualitative results with VideoCrafter2 [10].

(a) ”A lone figure in a hooded cloak stands atop a skyscraper, city lights sprawling below like glowing circuit board.”

(b) ”A mystical, low-poly enchanted garden with glowing plants and a small pond where fireflies dance under a starry sky.”

(c) ”A pair of tango dancers performing in Buenos Aires, 4K, high resolution.”

(d) ”A tiger walks in the forest, photorealistic, 4k, high definition, 4k resolution.”



(e) ”A tranquil, low-poly mountain village at dawn, with smoke rising from chimneys and birds flying over snowy peaks.”

(f) ”Cinematic closeup and detailed portrait of a reindeer in a snowy forest at sunset.”

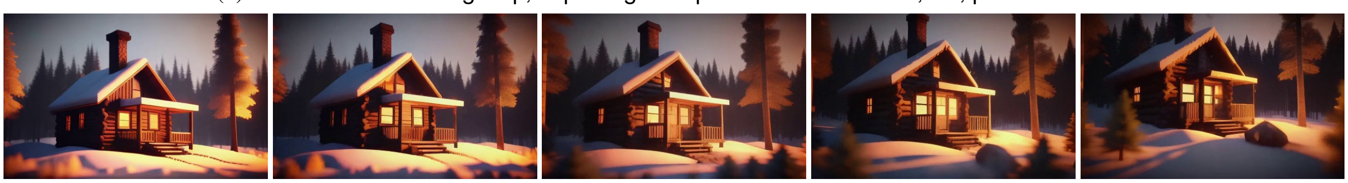
Figure 7. Videos generated by ScalingNoise with VideoCrafter2 based on the paradigm of FIFO-Diffusion.

(a) "A dynamic, low-poly cityscape at night, with neon lights reflecting off wet streets and a lone cyclist riding through the rain."

(b) "A bustling cityscape at sunset with skyscrapers reflecting golden light, people walking, and traffic moving swiftly."

(c) "A peaceful, low-poly countryside with rolling hills, a windmill, and a farmer tending to his crops under a golden sunset."

(d) "A horse race in full gallop, capturing the speed and excitement, 2K, photorealistic."



(e) "A cozy, low-poly cabin in the woods surrounded by tall pine trees, with a warm light glowing from the windows and smoke curling from the chimney, 4k resolution."

(f) "Impressionist style, a yellow rubber duck floating on the wave on the sunset, 4k resolution."

Figure 8. Videos generated by ScalingNoise with VideoCrafter2 based on the paradigm of FIFO-Diffusion.

In Fig. 9, we present the last individual frame of the generated videos.

Figure 9. The last individual video frame generated by ScalingNoise with VideoCrafter2 based on the paradigm of FIFO-Diffusion.

References

[1] Learning to reason with llms. https://openai.com/index/learning-to-reason-with-llms/, 2024. 2

[2] Jie An, Songyang Zhang, Harry Yang, Sonal Gupta, Jia-Bin Huang, Jiebo Luo, and Xi Yin. Latent-shift: Latent diffusion with temporal shift for efficient text-to-video generation, 2023. 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 3

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 2, 3

[5] Minghong Cai, Xiaodong Cun, Xiaoyu Li, Wenze Liu, Zhaoyang Zhang, Yong Zhang, Ying Shan, and Xiangyu Yue. Ditctrl: Exploring attention control in multi-modal diffusion transformer for tuning-free multi-prompt longer video generation, 2024. 3

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 5, 8

[7] Souradip Chakraborty, Soumya Suvra Ghosal, Ming Yin, Dinesh Manocha, Mengdi Wang, Amrit Singh Bedi, and Furong Huang. Transfer q star: Principled decoding for llm alignment. *arXiv preprint arXiv:2405.20495*, 2024. 3

[8] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 2

[9] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, pages 24081–24125. Curran Associates, Inc., 2024. 3

[10] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 3, 5, 7, 10

[11] Jingyuan Chen, Fuchen Long, Jie An, Zhaofan Qiu, Ting Yao, Jiebo Luo, and Tao Mei. Ouroboros-diffusion: Exploring consistent content generation in tuning-free long video diffusion. *arXiv preprint arXiv:2501.09019*, 2025. 2, 5

[12] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. *arXiv preprint arXiv:2310.20700*, 2023. 3

[13] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 3

[14] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 3

[15] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation with-

out vector quantization, 2025. 3

[16] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023. 2

[17] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization, 2024. 2

[18] Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. Cold-attack: Jailbreaking llms with stealthiness and controllability, 2024. 3

[19] Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation, 2025. 3

[20] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. 2

[21] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model, 2023. 5

[22] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. In *NeurIPS*, 2022. 3

[23] 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. 2

[24] 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. 3, 6

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Neural Information Processing Systems*, pages 6840–6851, Virtual Event, 2020. NeurIPS. 4

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 2

[27] Arian Hosseini, Xingdi Yuan, Nikolay Malkin, Aaron Courville, Alessandro Sordoni, and Rishabh Agarwal. V-star: Training verifiers for self-taught reasoners, 2024. 3

[28] 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 (CVPR)*, pages 21807–21818, 2024. 5

[29] Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. O1 replication journey – part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson?, 2024. 3

[30] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling, 2024. 2

[31] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training. *Advances in Neural Information Processing Systems*, 37:89834–89868, 2025. 2, 3, 5, 6

[32] PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, 2024. 6

[33] Chen Li, Weiqi Wang, Jingcheng Hu, Yixuan Wei, Nanning Zheng, Han Hu, Zheng Zhang, and Houwen Peng. Common 7b language models already possess strong math capabilities, 2024. 3

[34] Dacheng Li, Shiyi Cao, Tyler Griggs, Shu Liu, Xiangxi Mo, Eric Tang, Sumanth Hegde, Kourosh Hakhamaneshi, Shishir G. Patil, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Llms can easily learn to reason from demonstrations structure, not content, is what matters!, 2025. 3

[35] Wenhao Li, Yichao Cao, Xiu Su, Xi Lin, Shan You, Mingkai Zheng, Yi Chen, and Chang Xu. Training-free long video generation with chain of diffusion model experts. *arXiv preprint arXiv:2408.13423*, 2024. 2

[36] Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step, 2023. 2, 3

[37] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, Tanghui Jia, Junwu Zhang, Zhenyu Tang, Yatian Pang, Bin She, Cen Yan, Zhiheng Hu, Xiaoyi Dong, Lin Chen, Zhang Pan, Xing Zhou, Shaoling Dong, Yonghong Tian, and Li Yuan. Open-sora plan: Open-source large video generation model, 2024. 2

[38] Dongyang Liu, Shicheng Li, Yutong Liu, Zhen Li, Kai Wang, Xinyue Li, Qi Qin, Yufei Liu, Yi Xin, Zhongyu Li, Bin Fu, Chenyang Si, Yuewen Cao, Conghui He, Ziwei Liu, Yu Qiao, Qibin Hou, Hongsheng Li, and Peng Gao. Lumina-video: Efficient and flexible video generation with multi-scale next-dit, 2025. 3

[39] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024. 2

[40] Yexin Liu, Zhengyang Liang, Yueze Wang, Xianfeng Wu, Feilong Tang, Muyang He, Jian Li, Zheng Liu, Harry Yang, Sernam Lim, and Bo Zhao. Unveiling the ignorance of mllms: Seeing clearly, answering incorrectly. In *CVPR*, 2025. 3

[41] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videodrafter: Content-consistent multi-scene video generation with llm. *arXiv preprint arXiv:2401.01256*, 2024. 2

[42] Yu Lu, Yuanzhi Liang, Linchao Zhu, and Yi Yang. Free-long: Training-free long video generation with spectral-blend temporal attention. *arXiv preprint arXiv:2407.19918*, 2024. 2, 3, 5

[43] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *CVPR*, 2023. 2

[44] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, Kaijun Tan, Kang An, Mei Chen, Wei Ji, Qiling Wu, Wen Sun, Xin Han, Yanan Wei, Zheng Ge, Aojie Li, Bin Wang, Bizhu Huang, Bo Wang, Brian Li, Changxing Miao, Chen Xu, Chenfei Wu, Chenguang Yu, Dapeng Shi, Dingyuan Hu, Enle Liu, Gang Yu, Ge Yang, Guanzhe Huang, Gulin Yan, Haiyang Feng, Hao Nie, Hao-nan Jia, Hanpeng Hu, Hanqi Chen, Haolong Yan, Heng Wang, Hongcheng Guo, Huilin Xiong, Huixin Xiong, Jia-hao Gong, Jianchang Wu, Jiaoren Wu, Jie Wu, Jie Yang, Jiashuai Liu, Jiashuo Li, Jingyang Zhang, Junjing Guo, Junzhe Lin, Kaixiang Li, Lei Liu, Lei Xia, Liang Zhao, Liguo Tan, Liwen Huang, Liying Shi, Ming Li, Mingliang Li, Muhua Cheng, Na Wang, Qiaohui Chen, Qinglin He, Qiuyan Liang, Quan Sun, Ran Sun, Rui Wang, Shaoliang Pang, Shiliang Yang, Sitong Liu, Siqi Liu, Shuli Gao, Tiancheng Cao, Tianyu Wang, Weipeng Ming, Wenqing He, Xu Zhao, Xuelin Zhang, Xianfang Zeng, Xiaojia Liu, Xuan Yang, Yaqi Dai, Yanbo Yu, Yang Li, Yineng Deng, Yingming Wang, Yilei Wang, Yuanwei Lu, Yu Chen, Yu Luo, Yuchu Luo, Yuhe Yin, Yuheng Feng, Yuxiang Yang, Zecheng Tang, Zekai Zhang, Zidong Yang, Binxing Jiao, Jiansheng Chen, Jing Li, Shuchang Zhou, Xiangyu Zhang, Xinhao Zhang, Yibo Zhu, Heung-Yeung Shum, and Daxin Jiang. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. 2

[45] Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps, 2025. 2, 3

[46] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024. 6

[47] Yongjia Ma, Junlin Chen, Donglin Di, Qi Xie, Lei Fan, Wei Chen, Xiaofei Gou, Na Zhao, and Xun Yang. Tuning-free long video generation via global-local collaborative diffusion, 2025. 3

[48] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, et al. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7038–7048, 2024. 2

[49] NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. 2

[50] Yuta Oshima, Masahiro Suzuki, Yutaka Matsuo, and Hiroki Furuta. Inference-time text-to-video alignment with diffusion latent beam search. *arXiv preprint arXiv:2501.19252*, 2025. 2

[51] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally:diffusion noise selection and optimization, 2024. 2

[52] Zhenting Qi, Mingyuan Ma, Jiahang Xu, Li Lyna Zhang, Fan Yang, and Mao Yang. Mutual reasoning makes smaller llms stronger problem-solvers, 2024. 5

[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. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 3, 6

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[55] Weiming Ren, Huan Yang, Ge Zhang, Cong Wei, Xinrun Du, Wenhao Huang, and Wenhu Chen. Consisti2v: Enhancing visual consistency for image-to-video generation. *arXiv preprint arXiv:2402.04324*, 2024. 5, 6

[56] David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *ICML*, 2024. 3

[57] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016. 6

[58] Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. Repvideo: Rethinking cross-layer representation for video generation, 2025. 3

[59] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual,

Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *The Eleventh International Conference on Learning Representations*, 2023. 3

[60] Raghav Singhal, Zachary Horvitz, Ryan Teehan, Mengye Ren, Zhou Yu, Kathleen McKeown, and Rajesh Ranganath. A general framework for inference-time scaling and steering of diffusion models, 2025. 2, 5

[61] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, 2022. 6

[62] Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024. 2

[63] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 4

[64] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 6

[65] Zhenxiong Tan, Xingyi Yang, Songhua Liu, and Xinchao Wang. Video-infinity: Distributed long video generation, 2024. 2

[66] Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[67] Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing. *arXiv preprint arXiv:2404.12253*, 2024. 3

[68] Ye Tian, Ling Yang, Xinchen Zhang, Yunhai Tong, Mengdi Wang, and Bin Cui. Diffusion-sharpening: Fine-tuning diffusion models with denoising trajectory sharpening, 2025. 3

[69] Masatoshi Uehara, Yulai Zhao, Chenyu Wang, Xiner Li, Aviv Regev, Sergey Levine, and Tommaso Biancalani. Inference-time alignment in diffusion models with reward-guided generation: Tutorial and review, 2025. 3

[70] Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022. 3

[71] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. 6

[72] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD: Masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022. 2, 3

[73] Ante Wang, Linfeng Song, Ye Tian, Baolin Peng, Dian Yu, Haitao Mi, Jinsong Su, and Dong Yu. Litesearch: Efficacious tree search for llm. *arXiv preprint arXiv:2407.00320*, 2024. 3

[74] 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. 2, 3

[75] Guanghan Wang, Yair Schiff, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Remasking discrete diffusion models with inference-time scaling. *arXiv preprint arXiv:2503.00307*, 2025. 2

[76] Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, Peter Vajda, Niraj K. Jha, and Xiaoliang Dai. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity, 2024. 3

[77] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023. 3

[78] Peiyi Wang, Lei Li, Zhihong Shao, R. X. Xu, Damai Dai, Yifei Li, Deli Chen, Y. Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations, 2024. 3

[79] Xiyao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. Towards self-improvement of llms via mcts: Leveraging stepwise knowledge with curriculum preference learning. *arXiv preprint arXiv:2410.06508*, 2024. 3

[80] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 2

[81] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. FreeInit: Bridging initialization gap in video diffusion models. *arXiv preprint arXiv:2312.07537*, 2023. 2

[82] Tianxing Wu, Chenyang Si, Yuming Jiang, Ziqi Huang, and Ziwei Liu. Freeinit: Bridging initialization gap in video diffusion models. In *European Conference on Computer Vision*, pages 378–394. Springer, 2025. 2, 3

[83] Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation, 2025. 2

[84] Junfei Xiao, Feng Cheng, Lu Qi, Liangke Gui, Jiepeng Cen, Zhibei Ma, Alan Yuille, and Lu Jiang. Videoauteur: Towards long narrative video generation, 2025. 3

[85] Enze Xie, Junsong Chen, Yuyang Zhao, Jincheng Yu, Ligeng Zhu, Yujun Lin, Zhekai Zhang, Muyang Li, Junyu Chen, Han Cai, Bingchen Liu, Daquan Zhou, and Song Han. Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer, 2025. 3

[86] Yuxi Xie, Kenji Kawaguchi, Yiran Zhao, Xu Zhao, Min-Yen Kan, Junxian He, and Qizhe Xie. Self-evaluation guided beam search for reasoning, 2023. 2

[87] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 2023. 2

[88] Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions, 2024. 2

[89] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models, 2024. 2

[90] Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment. *arXiv preprint arXiv:2410.08193*, 2024. 3

[91] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model, 2023. 3

[92] Haochen Xue, Feilong Tang, Ming Hu, Yexin Liu, Qidong Huang, Yulong Li, Chengzhi Liu, Zhongxing Xu, Chong Zhang, Chun-Mei Feng, et al. Mmrc: A large-scale benchmark for understanding multimodal large language model in real-world conversation. *arXiv preprint arXiv:2502.11903*, 2025. 3

[93] An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024. 2

[94] Jiahui Yang, Donglin Di, Baorui Ma, Xun Yang, Yongjia Ma, Wenzhang Sun, Wei Chen, Jianxun Cui, Zhou Xue, Meng Wang, et al. Tv-3dg: Mastering text-to-3d customized generation with visual prompt. *arXiv preprint arXiv:2410.21299*, 2024. 2

[95] 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. 3

[96] Shengming Yin, Chenfei Wu, Huan Yang, Jianfeng Wang, Xiaodong Wang, Minheng Ni, Zhengyuan Yang, Linjie Li, Shuguang Liu, Fan Yang, Jianlong Fu, Gong Ming, Lijuan Wang, Zicheng Liu, Houqiang Li, and Nan Duan. NUWA-XL: Diffusion over diffusion for extremely long video generation. *arXiv preprint arXiv:2303.12346*, 2023. 3

[97] Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models, 2025. 2

[98] Yuanyang Yin, Yaqi Zhao, Mingwu Zheng, Ke Lin, Jiarong Ou, Rui Chen, Victor Shea-Jay Huang, Jiahao Wang, Xin Tao, Pengfei Wan, Di Zhang, Baoqun Yin, Wentao Zhang, and Kun Gai. Towards precise scaling laws for video diffusion transformers, 2024. 2

[99] Jaesik Yoon, Hyeonseo Cho, Doojin Baek, Yoshua Bengio, and Sungjin Ahn. Monte carlo tree diffusion for system 2 planning, 2025. 3

[100] Fangxu Yu, Lai Jiang, Haoqiang Kang, Shibo Hao, and Lianhui Qin. Flow of reasoning:training llms for divergent problem solving with minimal examples, 2025. 3

[101] Sihyun Yu, Kihyuk Sohn, Subin Kim, and Jinwoo Shin. Video probabilistic diffusion models in projected latent space. In *CVPR*, 2023. 6

[102] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition, 2024. 3

[103] Yunlong Yuan, Yuanfan Guo, Chunwei Wang, Hang Xu, and Li Zhang. Brick-diffusion: Generating long videos with brick-to-wall denoising, 2025. 3

[104] Oussama Zekri and Nicolas Boullé. Fine-tuning discrete diffusion models with policy gradient methods, 2025. 3

[105] Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. Rest-mcts*: Llm self-training via process reward guided tree search. *arXiv preprint arXiv:2406.03816*, 2024. 3

[106] Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization, 2025. 2

[107] Siyang Zhang and Ser-Nam Lim. Towards chunk-wise generation for long videos, 2024. 2, 3

[108] Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation, 2025. 2

[109] Zhongwei Zhang, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Ting Yao, Yang Cao, and Tao Mei. Trip: Temporal residual learning with image noise prior for image-to-video diffusion models, 2024. 3

[110] Min Zhao, Guande He, Yixiao Chen, Hongzhou Zhu, Chongxuan Li, and Jun Zhu. Riflex: A free lunch for length extrapolation in video diffusion transformers, 2025. 3

[111] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast, 2025. 2

[112] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, and Ser-Nam Lim. Videogen-of-thought: A collaborative framework for multi-shot video generation, 2024. 3

[113] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient video generation with latent diffusion models. *arXiv preprint arXiv:2211.11018*, 2022. 3

[114] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. Allegro: Open the black box of commercial-level video generation model, 2024. 2

[115] Zikai Zhou, Shitong Shao, Lichen Bai, Zhiqiang Xu, Bo Han, and Zeke Xie. Golden noise for diffusion models: A learning framework. *arXiv preprint arXiv:2411.09502*, 2024. 2