STREAMING AUTOREGRESSIVE VIDEO GENERATION VIA DIAGONAL DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Large-scale pretrained diffusion models have significantly enhanced the quality of generated videos, and yet their use in real-time streaming remains limited. Autoregressive models offer a natural framework for sequential frame synthesis but require heavy computation to achieve high fidelity. Diffusion distillation can compress these models into efficient few-step variants, but existing video distillation approaches largely adapt image-specific methods that neglect temporal dependencies. These techniques often excel in image generation but underperform in video synthesis, exhibiting reduced motion coherence, error accumulation over long sequences, and a latency–quality trade-off. We identify two factors that result in these limitations: insufficient utilization of temporal context during step reduction and implicit prediction of subsequent noise levels in next-chunk prediction (exposure bias). To address these issues, we propose *Diagonal Distillation*, which operates orthogonally to existing approaches and better exploits temporal information across both video chunks and denoising steps. Central to our approach is an asymmetric generation strategy: more steps early, fewer steps later. This design allows later chunks to inherit rich appearance information from thoroughly processed early chunks, while using partially denoised chunks as conditional inputs for subsequent synthesis. By aligning the implicit prediction of subsequent noise levels during chunk generation with the actual inference conditions, our approach mitigates error propagation and reduces oversaturation in long-range sequences. We further incorporate implicit optical flow modeling to preserve motion quality under strict step constraints. Our method generates a **5-second video** in just **2.61 seconds** (up to **31 FPS**), achieving a **277.3×** speedup over the undistilled model and doubling the acceleration ratio of the state-of-the-art (140×) without sacrificing visual quality. The source code and trained models will be released and we have uploaded the video to an anonymous hosting service at: https://diagonal-distillation.github.io/.

1 INTRODUCTION

Recent years have witnessed the rapid progress of diffusion models in video generation. A major enabler of such progress has been Diffusion Transformer architectures (Peebles & Xie, 2023), which leverage bidirectional attention to denoise all video frames simultaneously (Blattmann et al., 2023a;b; Brooks et al., 2024; Kong et al., 2024; Polyak et al., 2024; Villegas et al., 2022; Wan et al., 2025; Yang et al., 2024). While effective for offline generation, this design requires the entire video to be generated at once, as each frame can attend to all others, including future ones. As a result, such models face fundamental limitations in real-time applications, including game simulation (Deng et al., 2024; Peebles & Xie, 2023; Song et al., 2023; Vondrick et al., 2016) and robot learning (Ge et al., 2022; Jolicoeur-Martineau, 2018; Wang et al., 2023), where future frames are unavailable when generating the current frame.

Autoregressive (AR) models are well-suited for streaming video generation, as their chunk-by-chunk synthesis naturally aligns with real-time constraints (Bruce et al., 2024; Kondratyuk et al., 2023; Ren et al., 2025; Wang et al., 2024; Weissenborn et al., 2019; Yan et al., 2021). However, traditional GPT-style models (Wang et al., 2024; Yan et al., 2021) often suffer from limited visual quality (Gao et al., 2024a). To address this, recent works (Jin et al., 2024; Weng et al., 2024; Teng et al., 2025) integrate diffusion processes into AR generation. Yet these methods still require multiple denoising steps per chunk, which hinders real-time deployment. To reduce inference latency, step distillation (Yin et al., 2025; Huang et al., 2025; Yin et al., 2024b) has been introduced to distill multi-step diffusion models

Figure 1: Our Diagonal Distillation framework achieves comparable quality to the full-step model while significantly reducing latency. The method yields a 1.88× speedup on 5-second short video generation on a single H100 GPU.

into efficient few-step sampling AR model. Recent training methods (Chen et al., 2024; Gao et al., 2024b; Gu et al., 2025; Hu et al., 2024; Li et al., 2024b; Liu et al., 2024b; Weng et al., 2024; Yin et al., 2025; Zhang et al., 2025a;b) have further improved stability and efficiency, making interactive applications increasingly feasible (Arriola et al., 2025; Liu et al., 2024c).

Despite these advances, existing video distillation methods are largely adapted from image generation, and their direct extension to video often yields suboptimal results. This limitation arises from insufficient consideration of the temporal dimension and the neglect of inter-frame consistency. As a result, multi-step sampling remains essential for maintaining high-quality video generation. For example, while autoregressive frameworks such as Causvid (Yin et al., 2025) and Self Forcing (Huang et al., 2025) can reduce latency, they still require multiple steps per segment, and compressing them to fewer steps leads to noticeable performance degradation.

Our guiding insight is that, in autoregressive video generation, predicting the next chunk inherently requires predicting the next noise level (see Figure 2). This implicit prediction, however, introduces two critical challenges. First, autoregressive video models often suffer from exposure bias. When predicting the next chunk conditioned on previously generated clean frames, the model must implicitly predict the next noise level for subsequent frames. This can lead to progressive degradation, such as over-saturation in later frames, as errors in noise-level prediction accumulate over time. Although techniques like Self Forcing (Huang et al., 2025) have been proposed to mitigate exposure bias by using model-generated content during training, they still struggle to maintain visual quality over long sequences. Second, the same phenomenon implies that if structural priors are captured in early chunks, later chunks can generate relatively clear frames even with fewer denoising steps. However, existing distillation approaches often discard valuable temporal context accumulated across denoising steps in video generation models, which is essential for preserving coherence and detail when reducing the sampling steps.

Motivated by these insights, we introduce a flow-aware diagonal distillation framework – **DiaDistill** that redefines the temporal context incorporation by leveraging information across both time and denoising steps. Departing from standard practices that process chunks in isolation, our method employs a novel diagonal forcing mechanism operating jointly across time and denoising steps. This results in a diagonal denoising trajectory wherein earlier chunks are denoised with more steps, while later chunks use progressively fewer. This strategy improves computational efficiency by using less denoising steps in total and allows each chunk to inherit denoising trajectories from prior chunks as contextual priors—a training paradigm we term Diagonal Forcing. By explicitly simulating diagonal denoising paths during training through controlled noise injection, Diagonal Forcing enhances self-conditioned generation and mitigates error accumulation in long videos. Furthermore, we empirically observe that employing very few steps in later chunks can attenuate motion amplitude. To counteract this, we introduce Flow Distribution Matching, which integrates explicit temporal modeling into the distillation loss. This approach preserves dynamic consistency by ensuring the predicted motion distributions align with those of the full-step model, thus ensuring that the student model not only matches the teacher in image quality but also faithfully preserves motion characteristics. The contributions of this work are:

- We propose **Diagonal Distillation**, a method for high-quality video generation during model distillation and inference. It allocates more denoising steps to earlier chunks and progressively fewer to later ones, rather than keeping the number of steps constant across all chunks. This

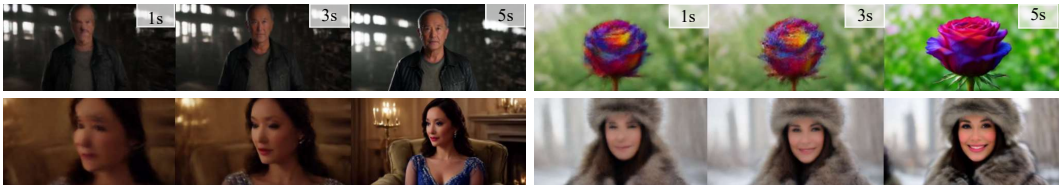


Figure 2: We find that when the training data uses explicit noise frames as conditions in Causvid (Yin et al., 2025), the next chunk prediction essentially functions as an implicit next noise level prediction. It can be observed that even with single-step prediction, the image progressively becomes clearer.

approach achieves an improved trade-off between quality and efficiency by leveraging contextual structured priors in AR video generation.

- We introduce **Diagonal Forcing** based on diagonal distillation, a unified method that operates along both temporal and denoising-step dimensions. It leverages trajectories from prior chunks as contextual priors and explicitly simulates diagonal denoising paths during training through controlled noise injection, thereby reducing long-term error accumulation.
- We present **Flow Distribution Matching** as a complementary component to diagonal distillation, designed to mitigate motion degradation and amplitude attenuation in later chunks. By incorporating explicit temporal modeling into the distillation loss, this approach enhances dynamic consistency and ensures smooth motion transitions.

2 Related Work

Diffusion Distillation. Diffusion distillation accelerates sampling via deterministic or distributional approaches. Deterministic methods (e.g., progressive distillation (Salimans & Ho, 2022), consistency distillation (Li et al., 2023; Song et al., 2023), rectified flow (Lamb et al., 2016)) regress noise-to-sample mappings but often yield blurry outputs with few steps due to optimization challenges (Kingma et al., 2021), typically requiring multiple steps (e.g., eight) for acceptable quality (Li et al., 2023; 2024a). Distributional methods approximate the teacher's distribution using adversarial training (Brooks et al., 2024; Ho et al., 2022), score distillation (Li et al., 2022; Luo et al., 2024), or hybrid objectives. Recent hybrids combine both paradigms but still suffer from one-step artifacts and commonly need multi-step sampling. Representative works include LADD (Sauer et al., 2024a) (relies on expensive pre-generated teacher targets), Lightning (Lin et al., 2024) and Hyper (Ren et al., 2024) (require intermediate timestep supervision), and DMD/DMD2 (Yin et al., 2024b;a) and ADD (Sauer et al., 2024b) (integrate adversarial and score matching losses). While these distillation methods have shown impressive results in image generation, their direct application to video often yields suboptimal results due to insufficient consideration of the temporal dimension and inter-frame consistency. Our work addresses this gap by proposing a flow-aware diagonal distillation framework specifically designed for video generation, which leverages temporal context across both time and denoising steps to maintain coherence while reducing sampling steps.

Autoregressive, Diffusion, and Hybrid Video Generation. Modern video generation is dominated by scalable diffusion and autoregressive (AR) models. Video diffusion models use bidirectional attention to denoise all frames concurrently (Blattmann et al., 2023a;b; Brooks et al., 2024; Deng et al., 2024; Kong et al., 2024; Polyak et al., 2024; Villegas et al., 2022; Wan et al., 2025; Yang et al., 2024), while AR models generate spatiotemporal tokens sequentially via next-token prediction (Bruce et al., 2024; Kondratyuk et al., 2023; Ren et al., 2025; Wang et al., 2024; Weissenborn et al., 2019; Yan et al., 2021; Liu et al., 2025). Hybrid models that merge these two paradigms have recently emerged as a promising direction (Chen et al., 2024; Gao et al., 2024b; Gu et al., 2025; Hu et al., 2024; Jin et al., 2024; Li et al., 2024b; Liu et al., 2024a;b; Weng et al., 2024; Yin et al., 2025; Zhang et al., 2025a;b), also in other sequence domains (Arriola et al., 2025; Liu et al., 2024c). These hybrids typically integrate diffusion into AR generation to boost visual quality, but they still require multiple denoising steps per chunk, hindering real-time deployment. Our work builds on these hybrids, drawing inspiration from Yin et al. (2025) and Huang et al. (2025) to mitigate exposure bias. However, these methods still face challenges with long-term error accumulation and motion degradation when compressed to fewer steps. Our proposed *Diagonal Distillation* framework addresses these issues via a novel diagonal forcing mechanism operating jointly across time and denoising steps, enabling efficient computation while preserving temporal coherence. The *Diagonal*

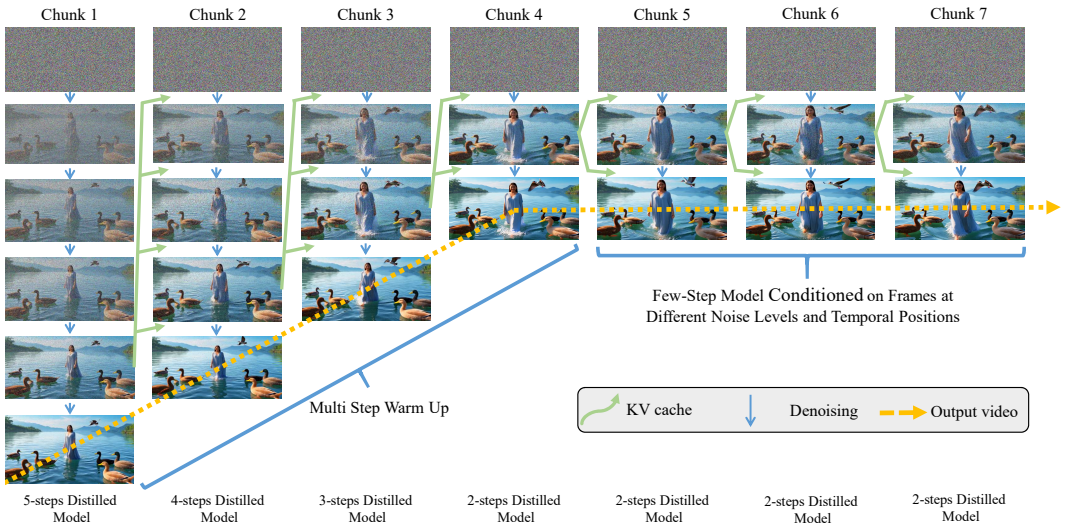


Figure 3: Diagonal Denoising with Diagonal Forcing and Progressive Step Reduction. We illustrate our method starting with 5 denoising steps for the first chunk and progressively reducing them to 2 steps by Chunk 7 (see Section 4.3 for more parameters). For chunks with $k \geq 4$, we use a fixed two-step denoising process, reusing the Key-Value (KV) cache from the previous chunk's last noisy frame. This approach maintains temporal coherence while reducing latency, the pseudo-code is provided in the appendix.

Forcing training paradigm explicitly simulates diagonal denoising paths to enhance self-conditioned generation, and *Flow Distribution Matching* ensures motion consistency with reduced steps.

3 Methodology

3.1 Preliminary and Framework Overview

Diffusion Models generate data through an iterative denoising process. The forward diffusion process progressively corrupts a sample $x \sim p_{\text{real}}$ over T steps, such that at timestep t, the diffused sample follows $p_{\text{real},t}(x_t) = \int p_{\text{real}}(x) q(x_t|x) dx$, with $q_t(x_t|x) \sim \mathcal{N}(\alpha_t x, \sigma_t^2 I)$, where $\alpha_t, \sigma_t > 0$ are determined by the noise schedule. The model learns to reverse this process by predicting a denoised estimate $\mu(x_t, t)$. The score function of the diffused distribution is:

$$s_{\text{real}}(x_t, t) = \nabla_{x_t} \log p_{\text{real},t}(x_t) = -\frac{x_t - \alpha_t \mu_{\text{real}}(x_t, t)}{\sigma_t^2}. \tag{1}$$

Sampling typically requires many iterative steps. Distribution Matching Distillation (DMD) distills a multi-step diffusion model (teacher) into a one-step generator G by minimizing the KL divergence between the diffused real and generated distributions, $p_{\text{real},t}$ and $p_{\text{fake},t}$. The gradient of this loss is:

$$\nabla \mathcal{L}_{\text{DMD}} = \mathbb{E}_t \left(\nabla_\theta \text{KL}(p_{\text{fake},t} \| p_{\text{real},t}) \right) = -\mathbb{E}_t \left(\int \left(s_{\text{real}}(F(G_\theta(z), t), t) - s_{\text{fake}}(F(G_\theta(z), t), t) \right) \frac{dG_\theta(z)}{d\theta} dz \right), \tag{2}$$

where $z \sim \mathcal{N}(0, \mathbf{I})$, F is the forward diffusion process, and s_{real}, s_{fake} are scores from models trained on real and generated data. An additional regression loss is often used for regularization:

$$\mathcal{L}_{\text{reg}} = E_{(z,y)} d(G_\theta(z), y), \tag{3}$$

where y is an image generated by the teacher from z. Directly applying DMD to video generation faces a significant challenge: the regression loss $\mathcal{L}_{\text{reg}}$ primarily ensures per-frame quality but fails to explicitly capture the underlying temporal coherence and long-range dependencies between frames, which are critical for video quality. This often results in degraded fluidity and consistency. To overcome this, we extend the DMD framework with two core innovations: 1) a **Diagonal Denoising with Diagonal Forcing** strategy that manages long-sequence generation and reduces error accumulation (Section 3.2)) a novel **Flow Distribution Matching** objective that explicitly aligns the temporal dynamics of the student and teacher models (Section 3.3).

3.2 Diagonal Denoising with Diagonal Forcing

Building upon the DMD foundation, we present diagonal distillation, a framework for efficient video generation. As illustrated in Figure 3, our approach introduces a Diagonal Denoising strategy that progressively reduces denoising steps across video chunks, combined with a novel Diagonal Forcing mechanism to maintain temporal coherence and mitigate error accumulation.

Diagonal Denoising: Progressive Step Reduction Strategy Our core innovation is a diagonal denoising strategy that allocates computation based on temporal importance. The method assigns more denoising steps to earlier chunks and progressively fewer to later ones, rather than maintaining a constant number of steps across all chunks. This approach achieves an improved trade-off between quality and efficiency by leveraging contextual structured priors in autoregressive video generation. For the first three chunks ($k = 1, 2, 3$), we use distilled models with decreasing steps ($s_k = 5, 4, 3$):

$$\mathbf{X}_k = \mathcal{D}_{s_k}(\mathbf{Z}_k|\tilde{\mathbf{X}}_{<k}), \tag{4}$$

where $\mathbf{X}_k$ is the k-th chunk output, $\mathbf{Z}_k \sim \mathcal{N}(0, \mathbf{I})$ is Gaussian noise, and $\tilde{\mathbf{X}}_{<k}$ contains previously noised chunks. For $k \geq 4$, we employ efficient two-step denoising:

$$\mathbf{C}_k = \mathcal{T}(\tilde{\mathbf{X}}_{k-1}), \mathbf{X}_k \quad = \mathcal{D}_2(\mathcal{D}_1(\mathbf{Z}_k|\mathbf{C}_k)|\mathbf{C}_k), \tag{5}$$

where $\mathbf{C}_k$ is the conditioning signal derived from previous chunks, $\mathcal{T}$ denotes the conditioning module, and $\mathcal{D}_1$, $\mathcal{D}_2$ represent the first and second denoising steps respectively.

Diagonal Forcing: Contextual Prior Propagation The core innovation of Diagonal Forcing lies in its explicit modeling of diagonal denoising trajectories during training through controlled noise injection. This approach ensures temporal coherence across chunks while minimizing error accumulation by conditioning each new chunk on the final noised state from the previous chunk's diffusion process. Specifically, the conditioning input for chunk k is derived from the clean output $\mathbf{X}_{k-1}$ of chunk $k-1$ through a noise injection operation:

$$\tilde{\mathbf{X}}_{k-1} = \sqrt{\alpha_{k-1}}\mathbf{X}_{k-1} + \sqrt{1-\alpha_{k-1}} \cdot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{6}$$

where α_{k-1} controls the noise schedule along the diagonal path and $\boldsymbol{\epsilon}$ is standard Gaussian noise. This formulation explicitly maintains the diagonal denoising trajectory $\mathbf{X}_k \rightarrow \tilde{\mathbf{X}}_{k-1} \rightarrow \mathbf{X}_{k-1}$, where $\tilde{\mathbf{X}}_{k-1}$ serves as the KV cache input for chunk k. By propagating these noised representations across chunks, the method effectively leverages denoising trajectories from prior chunks as contextual priors. The diagonal alignment of these trajectories ensures that error accumulation is minimized while preserving long-range coherence in the generated output.

3.3 Flow Distribution Matching

Motion attenuation in few-step denoising stems from truncated noise estimation paths. We quantify the temporal distribution mismatch through flow-based divergence:

$$\mathcal{E}_{\text{motion}} = D_{\text{KL}}\left(p_{\text{teacher}}(\mathcal{F}(\mathbf{x})|\mathbf{x}_t) \| p_{\text{student}}(\mathcal{F}(\mathbf{x})|\mathbf{x}_t)\right) \tag{7}$$

where $\mathcal{F}(\mathbf{x})$ represents the motion flow field extracted from video sequence $\mathbf{x}$. This measures the distributional divergence between teacher and student in the temporal dimension.

The standard Distribution Matching Distillation (DMD) framework minimizes spatial divergence through reverse KL minimization. We extend this to the temporal domain by defining flow distribution matching:

$$\nabla_\phi \mathcal{L}_{\text{DMD}}^{\text{flow}} \triangleq \mathbb{E}_t\left(\nabla_\phi \text{KL}\left(p_{\text{gen,flow},t} \| p_{\text{data,flow},t}\right)\right) \tag{8}$$

where $p_{\text{data,flow},t} = p(\mathcal{F}(\mathbf{x}) \mid \Psi(\mathbf{x}, t))$ is the smoothed flow distribution from real data, and $p_{\text{gen,flow},t} = p(\mathcal{F}(\mathbf{x}) \mid \Psi(G_\phi(\epsilon), t))$ is the generator's flow distribution. The gradient approximation for flow distribution matching follows the DMD framework:

$$\nabla_\phi \mathcal{L}_{\text{DMD}}^{\text{flow}} \approx -\mathbb{E}_t\left[\int \left(s_{\text{data}}^{\text{flow}}\left(\Psi(G_\phi(\epsilon), t), t\right) - s_{\text{gen},\xi}^{\text{flow}}\left(\Psi(G_\phi(\epsilon), t), t\right)\right)\frac{dG_\phi(\epsilon)}{d\phi}d\epsilon\right], \tag{9}$$

where $s_{\text{data}}^{\text{flow}}$ and $s_{\text{gen},\xi}^{\text{flow}}$ are the flow score functions defined as:

$$s^{\text{flow}}(\mathbf{x}_t, t) = \nabla_{\mathbf{x}_t} \log p(\mathcal{F}(\mathbf{x})|\mathbf{x}_t). \tag{10}$$

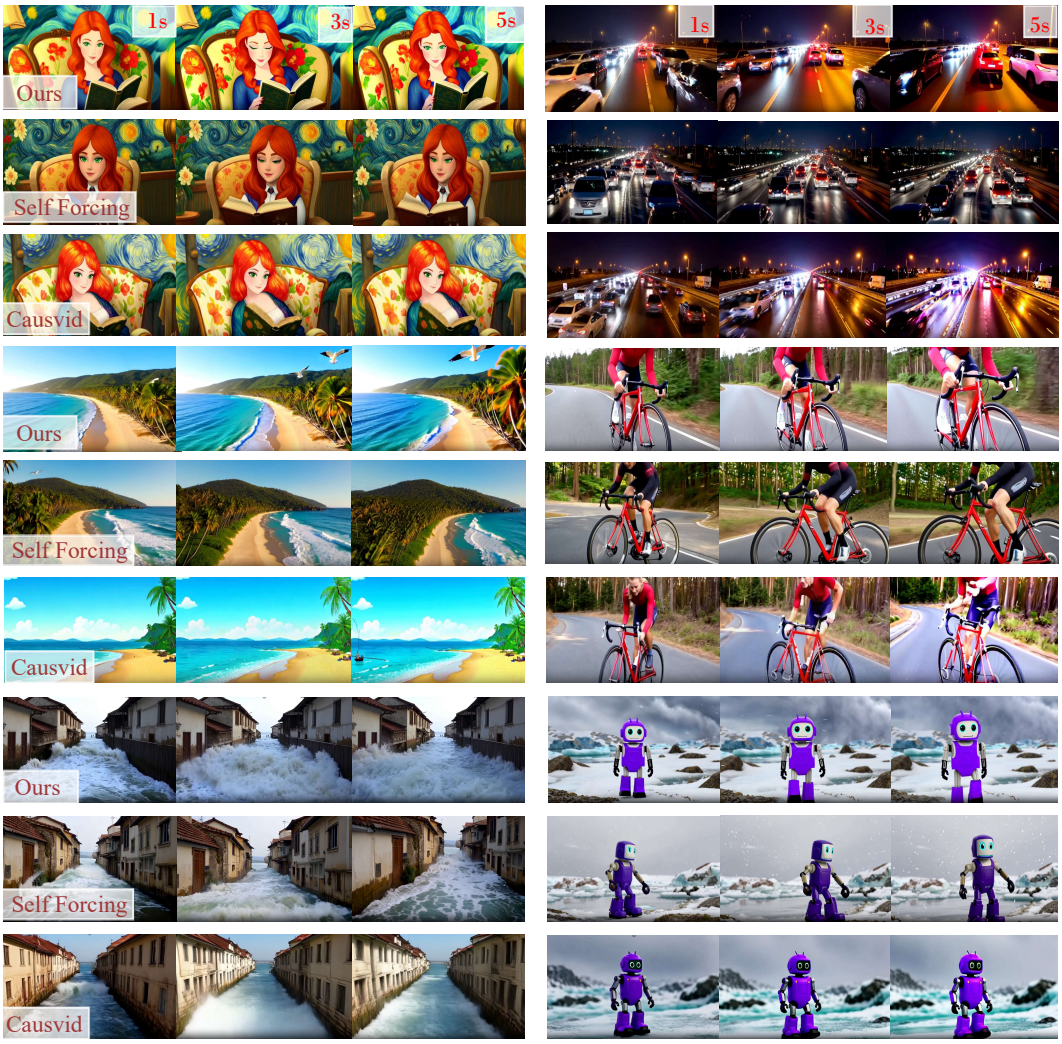


Figure 4: Comparing the results from three different models. For more results, please refer to our supplementary material.

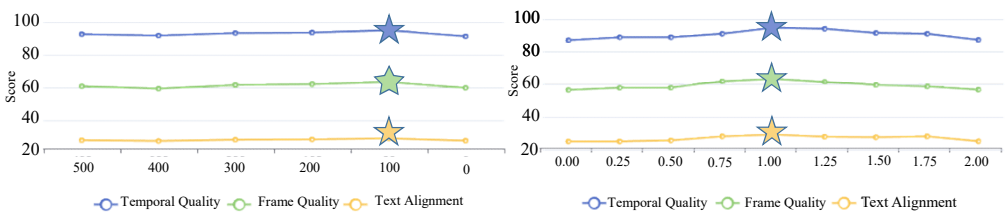
To operationalize this framework, we employ a flow regression loss for feature alignment:

$$\mathcal{L}_{\text{reg}}^{\text{flow}} = \mathbb{E}_{t,\epsilon}\left[\|\mathcal{F}(G_\phi^{\text{teacher}}(\epsilon,t)) - \mathcal{F}(G_\phi^{\text{student}}(\epsilon,t))\|_2^2\right], \tag{11}$$

where $G_\phi(\epsilon, t)$ denotes the generator output at timestep t. Our method uses a lightweight, self-contained motion feature extraction module $\mathcal{F}(\cdot)$ that operates directly on latent representations, avoiding dependencies on external pre-trained optical flow estimators. Specifically, we implement $\mathcal{F}(\cdot)$ as a learnable representation with convolution on latent difference: it first computes the difference between consecutive latent frames, then applies convolutional layers to extract local motion patterns, followed by an MLP for feature adaptation. The student version is trainable with gradient flow, while the teacher components are updated via EMA, ensuring stable and efficient motion representation learning. The overall training objective combines both spatial and temporal distribution matching:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{spatial}}\mathcal{L}_{\text{DMD}} + \mathcal{L}_{\text{reg}} + \gamma\left(\lambda_{\text{flow}}\mathcal{L}_{\text{DMD}}^{\text{flow}} + \mathcal{L}_{\text{reg}}^{\text{flow}}\right), \tag{12}$$

where γ weights the temporal terms. Where we set λ_{spatial}=4 and λ_{flow}=4 This framework jointly minimizes motion distribution divergence while maintaining spatial fidelity in the distilled video model.



(a) Average Quality over Time (b) Flow Loss Weight Evaluation

Figure 5: Ablation study results:(a)Performance evaluation across different diagonal forcing timesteps, demonstrating optimal outcomes at 100 steps (1000 steps correspond to complete noise addition, while 0 steps represent the clean frame);(b)Impact of motion loss weight on model performance.

4 Experiments and Results

4.1 Implementation Details

Training Details We implemented DiaDistill using Wan2.1-T2V-1.3B (Wan et al., 2025), a model based on Flow Matching (Lipman et al., 2022) that is capable of generating 5 videos at 16 FPS with a resolution of 832 × 480. For both ODE initialization and Diagonal Distillation training, we sampled text prompts from a filtered and LLM-extended version of VidProM (Wang & Yang, 2024).

Inference Details To assess real-time applicability, we measured both throughput (frames per second) and first-frame latency, acknowledging that true real-time performance requires exceeding video playback rates while maintaining imperceptible delay. All speed tests were conducted on a single NVIDIA H100 GPU with tiny VAE Boer Bohan (2025). The core component is the rolling KV cache mechanism following Self-Forcing Huang et al. (2025), which operates with a chunk size of 3 frames. Our buffering strategy is implemented using a fixed-size KV cache that maintains context from the most recent 4 chunks, resulting in a consistent memory footprint of 17.5 GB. For detailed ablation analysis please refer to our supplementary.

Evaluation Details We evaluated visual quality and semantic consistency using VBench (Huang et al., 2024). Temporal Quality is the average of Subject Consistency, Background Consistency, Temporal Flickering, Motion Smoothness, and Dynamic Degree. Frame Quality is the average of Aesthetic Quality and Imaging Quality. Text Alignment is the average of Object Class, Multiple Objects, Human Action, Color, Spatial Relationship, Scene, Appearance, Style, and Temporal Style. The aggregation method for each score is a simple arithmetic mean of the normalized scores from its constituent sub-dimensions. This approach is consistent with prior works like Causvid and Self-Forcing for fair comparison.

4.2 Comparison with State-of-the-Art Methods

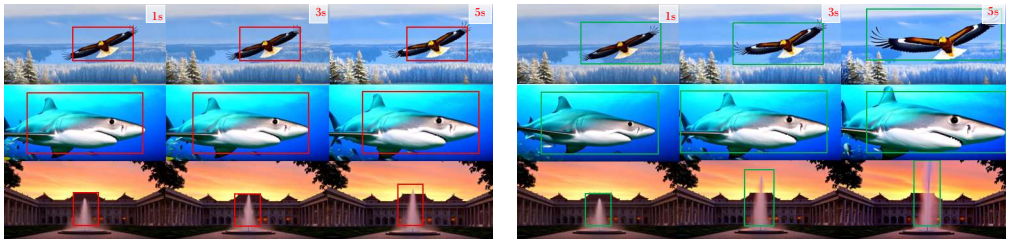
We evaluate DiaDistill against five state-of-the-art video generation methods: Wan2.1 (Wan et al., 2025), SkyReels-V2 (Chen et al., 2025), MAGI-1 (Teng et al., 2025), Causvid (Yin et al., 2025), and Self Forcing (Huang et al., 2025). As shown in Table 1, our method achieves a 277.3× speedup over the Wan2.1 baseline while maintaining competitive visual quality (85.26 vs. 85.3). This represents a 1.53× improvement in latency over the previous fastest method, Self Forcing (149.3×), alongside superior overall performance and semantic consistency . Qualitative results in Figure 4 further demonstrate advantages in temporal consistency, with smoother frame transitions and fewer dynamic artifacts. Visual fidelity improvements are most apparent in complex motions and textures, where baseline methods exhibit blurring or distortion. These findings collectively show that DiaDistill effectively balances the traditional trade-off between generation quality and computational efficiency.

Model	Throughput↑	First-Frame Latency ↓	Speedup	Total↑	Quality↑	Semantic↑
Wan2.1 (Wan et al., 2025)	0.78	103	1.0×	84.26	**85.3**	80.09
SkyReels-V2 (Chen et al., 2025)	0.49	112	0.91×	82.67	84.70	74.53
MAGI-1 (Teng et al., 2025)	0.19	282	0.36×	79.18	82.04	67.74
Causvid (Yin et al., 2025)	17.0	0.69	149.3×	81.20	84.05	69.80
Self Forcing (Huang et al., 2025)	17.0	0.69	149.3×	84.31	85.07	81.28
DiaDistill (Ours)	**31.0**	**0.37**	**277.3×**	**84.48**	85.26	**81.73**

Table 1: Comprehensive comparison of video generation methods

4.3 Ablation Studies

Key Components Diagonal Denoising assigns more denoising steps to early video chunks to establish a high-quality foundation and progressively reduces denoising steps for subsequent chunks, whereas



(a) w/o flow loss (b) with flow loss

Figure 6: **Visual comparison of motion effects:** (a) Without motion loss shows minimal motion amplitude with only slight object movement; (b) With motion loss demonstrates significantly increased motion amplitude throughout the entire frame, validating our method's effectiveness.

Figure 7: Qualitative comparison of long video generation(45s) with Self Forcing and Causvid. The visual results show that other methods suffer from noticeable saturation distortion and quality decay over time, whereas our approach preserves detail and consistency. Additional results are provided in the supplementary material.

without it, the same number of steps is applied uniformly across all chunks. Diagonal Forcing refers to using noisy frames instead of clean frames as the Key-Value (KV) cache in autoregressive generation. Our ablation study shows that removing either flow distribution matching loss or Diagonal Forcing significantly degrades video quality across all metrics (Table 2). Without Diagonal Denoising—which corresponds to the inference cost of Self Forcing in Table 1—we observe that the model achieves performance comparable to ours, but our method achieves a $1.53\times$ speedup. Notably, we find that flow distribution matching loss primarily benefits the few-step denoising regime and helps align its performance with the many-step denoising baseline (i.e., without Diagonal Denoising), and provides limited benefits when applied to a many-step denoising setting.

Diagonal Forcing Timesteps Moreover, we systematically evaluated diagonal forcing using metrics across different noise levels of timesteps for the kv cache. As Figure 5(a) shows, 100 timesteps achieved optimal scores across all evaluation dimensions, including temporal quality, frame quality, and text alignment. The performance peaks at this specific noise level before degrading as timesteps approach complete noise addition (1000 steps) or clean frames (0 steps). This can be attributed to the

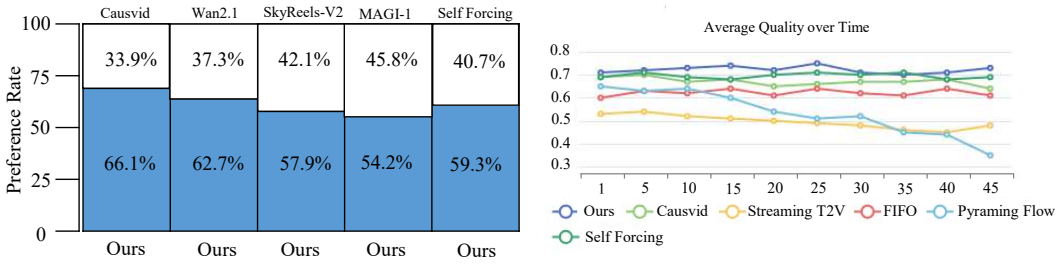


Figure 8: Quantitative evaluation of long video generation. The plot compares human preference scores and quality consistency over time for different methods under identical conditions. Our approach maintains stable quality throughout extended sequences, achieving scores above 50%, and attains a significant reduction in inference latency.

fact that excessive noise (high timesteps) blurs the structural priors in the video context, leading to reduced motion magnitude. This also explains why our method generates larger motion amplitudes compared to MAGI (Teng et al., 2025). Conversely, insufficient noise (low timesteps) causes the next chunk prediction to implicitly perform next noise level prediction, which can result in over-denoising of subsequent chunks and ultimately lead to over-saturated outputs.

Ablation Variant	Temporal Quality ↑	Frame Quality ↑	Text Alignment ↑	Total Score ↑
Without Diagonal Forcing	92.1	60.1	26.9	83.58
Without Flow Loss	92.5	60.8	27.8	84.18
Without Diagonal Denoising	**95.1**	63.2	28.6	84.46
Full Method (Ours)	94.9	**63.4**	**28.9**	**84.48**

Table 2: Ablation Study on Key Components of DiaDistill.

Steps	Temporal Quality ↑	Frame Quality ↑	Text Alignment ↑	NFEs	In-Flight Latency (s) ↓	Throughput (FPS) ↑
4322222	94.9	63.4	28.9	34	0.23 ± 0.02	31.0
5433333	95.1	63.2	29.3	48	0.34 ± 0.02	23.3
5432222	94.8	63.1	29.0	40	0.23 ± 0.02	29.7
5333333	95.0	63.9	29.1	46	0.34 ± 0.02	22.5
4333333	95.0	63.7	28.5	44	0.34 ± 0.02	23.5
4222222	93.4	62.3	27.8	32	0.23 ± 0.02	32.0

Table 3: Performance evaluation of denoising step configurations

Flow Loss Weight We conducted a comprehensive ablation study across eight motion loss weight configurations. Figure 5(b) reveals the crucial balance between motion guidance (via Flow Distribution Matching) and the DMD learning objectives, with optimal performance observed at a weight of 1.0. This balanced weighting scheme ensures the harmonious optimization of temporal consistency, frame quality, and textual alignment metrics.

Denoising Configurations We evaluated six denoising configurations (represented by 7-digit sequences specifying steps per chunk as a 5 seconds video have 7 chunks in our setting) across quality and computational metrics. As shown in Table 3, these configurations exhibit trade-offs between generation quality and efficiency. Among them, configuration 5333333 achieves the highest quality, while 4222222 offers the maximum throughput. To balance video quality and real-time performance, we selected configuration 4322222, as it has the second-lowest number of NFEs and delivers performance comparable to configurations with significantly higher latency and throughput, with only marginal differences.

KV Cache Scaling Analysis We further analyze the trade-offs between attention window size, video quality (Total Score), and in-flight interaction latency—the delay in responding to a new input signal during steady-state generation. As shown in Table 4, a larger window size generally improves quality at the cost of higher memory usage and, crucially, a longer delay in responding to user interactions. Performance plateaus around a window size of 18-27, leading us to select an optimal size that balances responsiveness with quality and efficiency.

Attention Window Size	Total Score	In-Flight Latency (s)	Memory (GB)
3	80.9	0.37 ± 0.01	14.9
6	81.3	0.38 ± 0.01	15.8
9	82.3	0.40 ± 0.01	16.6
12	84.3	0.46 ± 0.02	17.5
15	84.2	0.51 ± 0.02	18.4
18	84.4	0.54 ± 0.02	19.2
21	84.5	0.59 ± 0.02	20.1
24	84.3	0.64 ± 0.02	20.9
27	84.5	0.68 ± 0.02	21.8

Table 4: KV cache scaling analysis

4.4 Long Video Generation Evaluation

We evaluated our long video generation framework using both simple and complex prompts. As shown in Figure 8, our model maintains consistent perceptual quality over time, whereas baseline methods suffer from rapid quality decay due to error accumulation. A large-scale user study (93 participants, 150 comparisons per model pair) on the first 50 prompts from MovieGenBench further validated

Figure 9: Illustration of long video generation with dynamic prompting. This feature allows for the integration of new prompts at arbitrary time points, facilitating the creation of coherent long videos with changing narratives. The specific prompts used for each segment are detailed in the appendix.

our method's superiority in overall visual quality, text faithfulness, and long-term consistency. User study results, consistent with the qualitative comparison in Figure 7, confirm that baseline methods degrade with issues like saturation distortion, while our approach sustains high quality. A key feature of our framework is its support for *dynamic prompting* (Figure 9), allowing users to input new text descriptions at any timeline point to create complex narratives with evolving scenes and actions.

5 Concluding Remarks

In this work, we introduce Diagonal Distillation, a novel framework for efficient autoregressive video generation. It leverages temporal dependencies across video chunks and denoising steps through an asymmetric denoising strategy—allocating more steps to early chunks and progressively fewer to later ones. This design significantly reduces the total number of denoising steps while preserving motion coherence and visual quality. Diagonal Forcing explicitly models the temporal denoising trajectory, reducing error accumulation and aligning training with inference for stable long-range synthesis. Additionally, Flow Distribution Matching ensures dynamic consistency under strict step constraints by aligning the optical flow distributions of generated and real videos. Extensive experiments demonstrate our method's superior trade-off between efficiency and quality.

Ethics Statement

While the real-time video generation technology presented in this study significantly improves generation efficiency (achieving a 277.3× speedup compared to the baseline model), we are fully aware of its dual-use nature. This technology could potentially be misused to create misleading content or deepfake videos. To mitigate this risk, we commit to embedding usage guidelines and restrictions when open-sourcing the code and models, and we advocate for the adoption of traceability technologies such as digital watermarks and content authentication. Concurrently, this technology holds significant positive potential in fields such as education, the creative industries, and assistive tools. We aim to maximize its societal benefits and minimize potential harms through ongoing discussions on technology ethics and responsible release practices.

Reproducibility Statement

For detailed reproducibility information, including full implementation details, training configurations, hyperparameters, and evaluation protocols, please refer to the appendix sections. All source code, trained model weights, and configuration files will be released to ensure the full reproducibility of our results.

References

Marianne Arriola, Aaron Gokaslan, Justin T Chiu, Zhihan Yang, Zhixuan Qi, Jiaqi Han, Subham Sekhar Sahoo, and Volodymyr Kuleshov. Block diffusion: Interpolating between autoregressive and diffusion language models. *arXiv preprint arXiv:2503.09573*, 2025. 2, 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*, 2023a. 1, 3

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*, pp. 22563–22575, 2023b. 1, 3

Ollin Boer Bohan. Taehv: Tiny autoencoder for hunyuan video. https://github.com/madebyollin/taehv, 2025. 7

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):1, 2024. 1, 3

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. 1, 3

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024. 2, 3

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 7

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 1, 3

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, and Jun Xiao. Vid-gpt: Introducing gpt-style autoregressive generation in video diffusion models. *arXiv preprint arXiv:2406.10981*, 2024a. 1

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. *arXiv preprint arXiv:2411.16375*, 2024b. 2, 3

Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pp. 102–118. Springer, 2022. 1

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025. 2, 3

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 3

Jinyi Hu, Shengding Hu, Yuxuan Song, Yufei Huang, Mingxuan Wang, Hao Zhou, Zhiyuan Liu, Wei-Ying Ma, and Maosong Sun. Acdit: Interpolating autoregressive conditional modeling and diffusion transformer. *arXiv preprint arXiv:2412.07720*, 2024. 2, 3

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 1, 2, 3, 7

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. 7

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. 1, 3

A Jolicoeur-Martineau. The relativistic discriminator: A key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 1

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 3

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023. 1, 3

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 1, 3

Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016. 3

Qing Li, Xun Tang, Junkun Peng, Yuanzheng Tan, and Yong Jiang. Latency reducing in real-time internet video transport: A survey. *Available at SSRN 4654242*, 2023. 3

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024a. 3

Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *European conference on computer vision*, pp. 515–534. Springer, 2022. 3

Zongyi Li, Shujie Hu, Shujie Liu, Long Zhou, Jeongsoo Choi, Lingwei Meng, Xun Guo, Jinyu Li, Hefei Ling, and Furu Wei. Arlon: Boosting diffusion transformers with autoregressive models for long video generation. *arXiv preprint arXiv:2410.20502*, 2024b. 2, 3

Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 3

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 7

Haozhe Liu, Shikun Liu, Zijian Zhou, Mengmeng Xu, Yanping Xie, Xiao Han, Juan C Pérez, Ding Liu, Kumara Kahatapitiya, Menglin Jia, et al. Mardini: Masked autoregressive diffusion for video generation at scale. *arXiv preprint arXiv:2410.20280*, 2024a. 3

Jinxiu Liu, Shaoheng Lin, Yinxiao Li, and Ming-Hsuan Yang. Dynamicscaler: Seamless and scalable video generation for panoramic scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6144–6153, 2025. 3

Yaofang Liu, Yumeng Ren, Xiaodong Cun, Aitor Artola, Yang Liu, Tieyong Zeng, Raymond H Chan, and Jean-michel Morel. Redefining temporal modeling in video diffusion: The vectorized timestep approach. *arXiv preprint arXiv:2410.03160*, 2024b. 2, 3

Zhijun Liu, Shuai Wang, Sho Inoue, Qibing Bai, and Haizhou Li. Autoregressive diffusion transformer for text-to-speech synthesis. *arXiv preprint arXiv:2406.05551*, 2024c. 2, 3

Weijian Luo, Zemin Huang, Zhengyang Geng, J Zico Kolter, and Guo-jun Qi. One-step diffusion distillation through score implicit matching. *Advances in Neural Information Processing Systems*, 37:115377–115408, 2024. 3

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023. 1

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 1, 3

Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. Next block prediction: Video generation via semi-autoregressive modeling. *arXiv preprint arXiv:2502.07737*, 2025. 1, 3

Yuxi Ren, Xin Xia, Yanzuo Lu, Jiacheng Zhang, Jie Wu, Pan Xie, Xing Wang, and Xuefeng Xiao. Hyper-sd: Trajectory segmented consistency model for efficient image synthesis. *Advances in Neural Information Processing Systems*, 37:117340–117362, 2024. 3

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022. 3

Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11, 2024a. 3

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024b. 3

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models, 2023. URL https://arxiv.org/abs/2303.01469. 1, 3

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 1, 7, 9

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 1, 3

Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 1

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 1, 3, 7

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024. 7

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024. 1, 3

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023. 1

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. Scaling autoregressive video models. *arXiv preprint arXiv:1906.02634*, 2019. 1, 3

Wenming Weng, Ruoyu Feng, Yanhui Wang, Qi Dai, Chunyu Wang, Dacheng Yin, Zhiyuan Zhao, Kai Qiu, Jianmin Bao, Yuhui Yuan, et al. Art-v: Auto-regressive text-to-video generation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7395–7405, 2024. 1, 2, 3

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1, 3

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. 1, 3

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a. 3

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b. 1, 3

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974, 2025. 1, 2, 3, 7

Tianyuan Zhang, Sai Bi, Yicong Hong, Kai Zhang, Fujun Luan, Songlin Yang, Kalyan Sunkavalli, William T Freeman, and Hao Tan. Test-time training done right. *arXiv preprint arXiv:2505.23884*, 2025a. 2, 3

Yuan Zhang, Jiacheng Jiang, Guoqing Ma, Zhiying Lu, Haoyang Huang, Jianlong Yuan, and Nan Duan. Generative pre-trained autoregressive diffusion transformer. *arXiv preprint arXiv:2505.07344*, 2025b. 2, 3