S^{2}Flow: Towards Fast and Authentic Training-Free High-Resolution Video Generation

Chaoqun Wang[1], **Shaobo Min**[2], **Xu Yang**[3*]
[1]School of Artificial Intelligence, South China Normal University, Guangzhou 510665, China
[2]University of Science and Technology of China, Hefei 230026, China
[3]School of Electronic Engineering, Xidian University, Xi'an 710071, China
cq23@scnu.edu.cn, mbobo2ml@gmail.com, xuyang.xd@gmail.com

Abstract

Rectified flow models have shown strong potential in high-fidelity video generation, yet extending them to high-resolution remains challenging due to the high cost of full attention and error accumulation in the ODE-solving process. In this paper, we propose S^{2}Flow, a training-free framework that enables efficient and authentic high-resolution video generation by jointly exploring **Flow**-guided **S**parse attention and **S**econd-order ODE solution. Specifically, S^{2}Flow exploits and transfers the semantic and structural information from the low-resolution flow trajectory to guide the high-resolution flow in two aspects. First, S^{2}Flow dynamically captures the sparse patterns of the spatio-temporal attention maps from low-resolution videos to construct localized 3D windows, enabling efficient window attention in high-resolution inference. This can significantly reduce redundant computation while preserving contextual dependencies. Second, S^{2}Flow adopts a second-order ODE solver based on Taylor expansion, where the high-order derivative is approximated via central difference from the low-resolution flow, facilitating accurate high-resolution denoising. Extensive experiments on VBench dataset demonstrate that S^{2}Flow outperforms prior methods in both visual quality and inference speed, enabling $4\times$ acceleration on 2560×1536 video generation.

Introduction

Diffusion/rectified flow (RF) models have achieved remarkable success in visual content generation, demonstrating strong capabilities in text-to-image (Rombach et al. 2022; Saharia et al. 2022; Podell et al. 2024) and text-to-video (Ho et al. 2022; Kong et al. 2024) synthesis. However, most existing models operate at moderate resolutions (typically up to 1280×768 for videos), which limits their applicability in real-world high-resolution scenarios such as scientific visualization and industrial design. Retraining large-scale models at higher resolutions is computationally prohibitive, while direct inference-time upscaling suffers from structural artifacts and repetition (He et al. 2023; Huang et al. 2024a).

Recent training-free methods offer a promising solution by reusing pretrained models without additional retraining (Du et al. 2024; Bar-Tal et al. 2023; Haji-Ali, Balakrishnan, and Ordonez 2024). Most existing methods focus on image upscaling, and propose a cascade denoising strategy, *e.g.,* HiFlow (Bu et al. 2025) leverages flow-aligned guidance. However, video generation poses a significantly higher computational burden due to the quadratic complexity of spatio-temporal attention, when scaling to high resolution videos. Thus, current efforts on training-free high-resolution video generation (He et al. 2023; Qiu et al. 2024; Guo et al. 2024) only build upon lightweight backbones such as LVDM (He et al. 2022) and VideoCrafter2 (Chen et al. 2024), which offer limited generative capacity. In contrast, recent large-scale backbones (Kong et al. 2024; Wan et al. 2025) achieve stronger fidelity but suffer from extremely expensive for high-resolution inference. For instance, the leading video DiT model, HunyuanVideo (Kong et al. 2024), takes about 4.3 minutes to generate a 1280×768 video, and the cost increases rapidly to about 1 hours for 2560×1536 distributed on $8\times$H800 GPUs. Although brand-new acceleration strategies, such as sparse attention (Zhang et al. 2025; Lu et al. 2025), have shown promising performance, they usually require an additional online/offline profiling step to determine their sparse strategy. Crucially, existing methods treat training-free video upscaling and efficient inference as independent problems, overlooking their potential synergy.

In this paper, we propose S^{2}Flow, a unified training-free framework for high-resolution video generation that addresses both inference efficiency and visual fidelity through sparse attention and second-order ODE solver. Specifically, S^{2}Flow adopts a cascade generation paradigm, decomposing the flow trajectory into low- and high-resolution stages. These two stages exhibit strong semantic and structural alignment, as they describe the same underlying video content at different granularities. Leveraging this consistency, our core insight is that the low-resolution trajectory encodes sufficient guidance to inform both the sparse attention mechanism and high-order denoising required for high-resolution synthesis. First, for inference efficiency, S^{2}Flow extracts spatio-temporal attention patterns from the low-resolution flow to estimate token relevance. These patterns are transferred to the high-resolution stage to construct adaptive sparse attention windows, which attend to semantically meaningful local regions. This significantly reduces redundant computation while preserving contextual coherence across dense video tokens. Second, for fidelity, the commonly-used first-order Euler solver introduces error ac-

*Corresponding author.

Figure 1: S^{2}Flow is a training-free framework for efficient and authentic high-resolution video generation. Built upon the leading video DiT model pretrained at 1280×768, S^{2}Flow enables 2560×1536 synthesis with $4\times$ faster inference while preserving visual fidelity.

cumulation over time. To enhance accuracy, S^{2}Flow employs a second-order ODE solver based on Taylor expansion, where the high-order derivative is approximated via central difference estimiated from the low-resolution trajectory. This allows for more precise high-resolution denoising with negligible increase in inference cost. Consequently, S^{2}Flow enables efficient and accurate scaling of large RF models to high-resolution video generation without retraining. Extensive experiments on the VBench dataset (Huang et al. 2024b) demonstrate that S^{2}Flow consistently outperforms existing training-free approaches in both visual quality and inference efficiency. Our contributions are summarized as follows:

- We propose S^{2}Flow, a unified training-free framework for high-resolution video generation, simultaneously addressing inference efficiency and generation fidelity via sparse attention and second-order solver.
- S^{2}Flow designs a cascaded flow pipeline that distills semantic and structural cues from low-resolution trajectories to guide efficient sparse attention and accurate second-order ODE solving in high-resolution inference.
- Extensive experiments on the challenging VBench benchmark demonstrate that S^{2}Flow achieves superior visual quality and faster inference compared to existing approaches.

Related Works

Text-to-Video Generation

Video generation has become an increasingly important topic in generative modeling. Recent advances in diffusion and rectified flow models (Blattmann et al. 2023; Chen et al. 2024; Skorokhodov et al. 2024; Wang et al. 2024a; Esser et al. 2024; Liu, Gong, and Liu 2022) have led to substantial improvements in text-to-video generation, with training on large-scale video datasets (Bain et al. 2021; Wang et al. 2024b). Subsequent methods are proposed to enhance both motion coherence and visual quality (Qing et al. 2024; Xing et al. 2025; Menapace et al. 2024). To improve temporal consistency, MotionBooth (Wu et al. 2024) incorporates motion-aware mechanisms for customized video generation. MoVideo (Liang et al. 2024) leverages optical flow to model cross-frame correspondences, enabling more faithful detail preservation and improved temporal alignment. In terms of spatial fidelity, CustomVideo (Wang et al. 2024c) introduces co-occurrence constraints and attention control mechanisms guided by spatial masks to preserve subject identity.

High-Resolution Visual Generation

Diffusion and rectified models for image and video generation are trained at moderate resolutions, which are insufficient for many commercial and professional applications. When directly applied to higher resolutions, these models often suffer from structural artifacts and a loss of fine-grained details (Huang et al. 2024a; Zheng et al. 2024). Retraining these models with high-resolution datasets is a straightforward solution, yet it demands substantial computational resources and high-quality annotated datasets, which are expensive to collect (He et al. 2023). To address these limitations, recent research has explored training-free strategies that leverage pretrained models without requiring additional fine-tuning or data collection (Du et al. 2024;

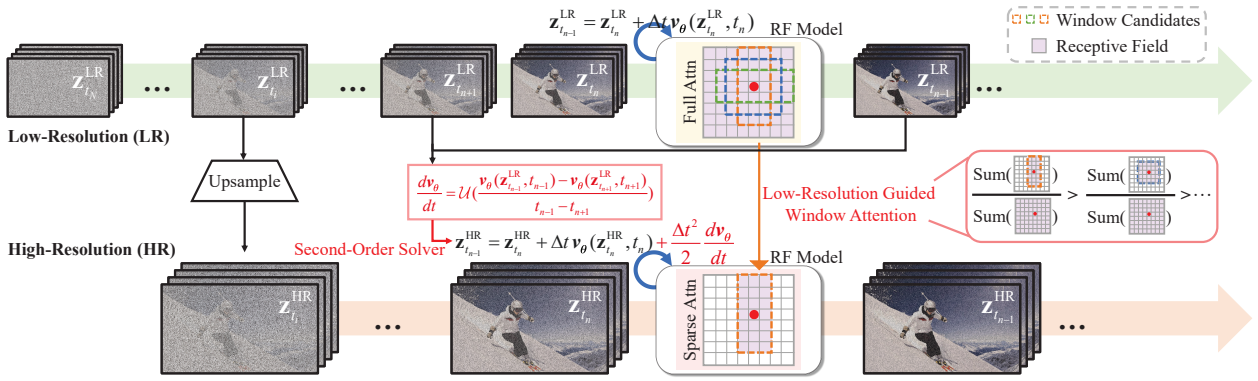


Figure 2: Pipeline overview of S^{2}Flow. The low-resolution (LR) trajectory guides both sparse attention for efficient inference and high-order correction for accurate high-resolution (HR) denoising. 2D attention maps are shown for clear illustration, and 3D implementation can be inferred.

Vontobel et al. 2025; Bar-Tal et al. 2023; Haji-Ali, Balakrishnan, and Ordonez 2024; Guo et al. 2024). These methods offer a more practical and accessible alternative, especially in real-world deployment scenarios.

One major line of training-free methods focuses on patch-wise generation, in which high-resolution outputs are synthesized as overlapping patches that are later merged (Bar-Tal et al. 2023; Haji-Ali, Balakrishnan, and Ordonez 2024; Lee et al. 2023). For instance, MultiDiffusion (Bar-Tal et al. 2023) smoothly fuses the image patches during the denoising process to eliminate incoherent patches. These methods often suffer from spatial discontinuities and semantic inconsistencies (Cao et al. 2024). Another category of approaches aims to improve high-resolution synthesis by modifying the architecture of diffusion models (He et al. 2023; Jin et al. 2023; Huang et al. 2024a). ScaleCrafter (He et al. 2023) expands the receptive field by incorporating dilated convolutions into the denoising U-Net. FreeScale (Qiu et al. 2024) applies restricted dilated convolutions in the downsampling and middle blocks. However, altering the structure of the model can introduce undesirable artifacts or deformation (Kim et al. 2025).

Preliminary: Rectified Flow Model

Rectified flow (RF) model (Liu, Gong, and Liu 2022) learns a continuous transport field that deterministically maps a Gaussian distribution to the data distribution via an ordinary differential equation (ODE):

$$\frac{d\mathbf{z}(t)}{dt} = \boldsymbol{v}_\theta(\mathbf{z}(t), t), \quad t \in [0, 1], \tag{1}$$

where $\mathbf{z}(t)$ is the latent representation of a video at time t, and $\boldsymbol{v}_\theta$ is a learnable time-dependent velocity field parameterized by θ. In practice, the ODE is numerically solved using the first-order Euler method over N discrete timesteps $\{t_N, \ldots, t_0\}$:

$$\mathbf{z}_{t_{n-1}} = \mathbf{z}_{t_n} + \Delta t\, \boldsymbol{v}_\theta(\mathbf{z}_{t_n}, t_n), \quad n = N, \ldots, 1, \tag{2}$$

where $\Delta t = t_{n-1} - t_n$.

RF model is trained by minimizing the expected squared error between the linear transport vector and the learned velocity field:

$$\min_\theta \int_0^1 \mathbb{E}\left[\|(\mathbf{z}_1 - \mathbf{z}_0) - \boldsymbol{v}_\theta(\mathbf{z}_t, t)\|_2^2\right] dt, \tag{3}$$

where $\mathbf{z}_0$ is sampled from the standard Gaussian prior, $\mathbf{z}_1$ is a real data sample, and $\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\mathbf{z}_1$ is a linear interpolation.

Each latent variable $\mathbf{z}_t \in \mathbb{R}^{H' \times W' \times T \times C}$ corresponds to the spatio-temporal latent of a video, where H', W' are spatial resolutions, T is the number of frames, and C is the channel dimension. The latent is obtained from a pretrained video VAE encoder $\mathcal{E} : \mathbf{x} \mapsto \mathbf{z}$, and decoded via $\mathcal{D} : \mathbf{z} \mapsto \mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^{H \times W \times T \times 3}$ denotes the RGB video.

At inference, a noise latent $\mathbf{z}_{t_N}$ is sampled and progressively denoised to $\mathbf{z}_{t_0}$ using the RF model, then decoded to a final video. Existing RF models are typically trained on low-resolution video datasets (*e.g.*, up to 1280×768). Applying low-resolution RF models to high-resolution synthesis faces two challenges: 1) prohibitively high attention computation due to quadratic complexity, and 2) degraded quality, including repetitive patterns and missing details.

Method

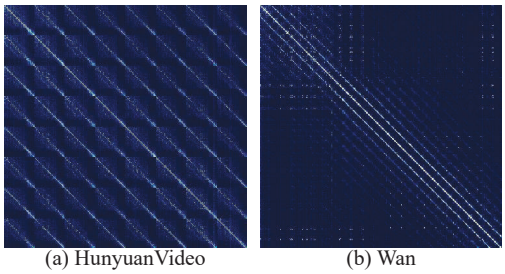
Given a pretrained RF model θ trained on low-resolution (LR) videos $\mathbf{x}^{\mathrm{LR}} \in \mathbb{R}^{H^{\mathrm{LR}} \times W^{\mathrm{LR}} \times T \times 3}$, we aim to generate high-resolution (HR) videos $\mathbf{x}^{\mathrm{HR}} \in \mathbb{R}^{H^{\mathrm{HR}} \times W^{\mathrm{HR}} \times T \times 3}$, where $H^{\mathrm{HR}} > H^{\mathrm{LR}}$, $W^{\mathrm{HR}} > W^{\mathrm{LR}}$.

The baseline is a cascaded high-resolution video generation process. First, we generate a latent trajectory in the LR space using the pretrained RF model θ. Starting from noise $\mathbf{z}^{\mathrm{LR}}_{t_N} \sim \mathcal{N}(0, \mathbf{I})$, the latent trajectory $\{\mathbf{z}^{\mathrm{LR}}_{t_n}\}_{n=N}^{0}$ is generated via the RF ODE:

$$\mathbf{z}^{\mathrm{LR}}_{t_{n-1}} = \mathbf{z}^{\mathrm{LR}}_{t_n} + \Delta t\, \boldsymbol{v}_\theta(\mathbf{z}^{\mathrm{LR}}_{t_n}, t_n), \tag{4}$$

where $\Delta t = t_{n-1} - t_n$. Second, at an intermediate timestep t_i, we decode $\mathbf{z}^{\mathrm{LR}}_{t_i}$ to the pixel space, perform spatial upsampling, and re-encode the result into the HR latent space:

$$\mathbf{z}^{\mathrm{HR}}_{t_i} = \mathcal{U}_p(\mathbf{z}^{\mathrm{LR}}_{t_i}, \theta_e), \tag{5}$$



(a) HunyuanVideo (b) Wan

Figure 3: Attention maps in HunyuanVideo and Wan exhibit strong spatial-temporal sparsity and locality. All the spatio-temporal tokens are flattened, so the adjacent tokens present periodicity here.

where $\mathcal{U}_p(\cdot,\cdot)$ denotes the upsampling operation and θ_e are VAE parameters that bridges latent and pixel spaces. Finally, starting from $\mathbf{z}^{\mathrm{HR}}_{t_i}$, we complete the remaining trajectory in the HR space:

$$\mathbf{z}^{\mathrm{HR}}_{t_{n-1}} = \mathbf{z}^{\mathrm{HR}}_{t_n} + \Delta t\ \boldsymbol{v}_\theta(\mathbf{z}^{\mathrm{HR}}_{t_n}, t_n),\ n \le i. \tag{6}$$

Building on the above baseline, the use of an LR-trained model θ to process HR latents introduces distributional mismatches and computational inefficiencies. To this end, we propose S^{2}Flow that consists of two main components: 1) LR-Guided Window Attention (LRG-WinAttn) that reduces computational cost while maintaining spatio-temporal fidelity, and 2) LR-Induced High-Order Solver (LRI-HiSolver) that explores high-order Tayler expansion terms to improve the ODE solution accuracy. The framework overview is shown in Fig. 2.

Low-Resolution Guided Window Attention

Full spatio-temporal attention incurs quadratic computational complexity with respect to the number of tokens. This becomes prohibitive for high-resolution video generation involving dense grids, *e.g.*, 0.46M tokens for a 2560×1536 video (Kong et al. 2024). However, as illustrated in Fig. 3, we observe that attention maps from state-of-the-art models (HunyuanVideo (Kong et al. 2024), Wan (Wan et al. 2025)) remain largely localized, despite being trained with full attention. Such attention sparsity has also been consistently observed in (Zhang et al. 2025), suggesting substantial redundancy in full attention. Motivated by this, we propose Low-Resolution Guided Window Attention (LRG-WinAttn), a lightweight attention mechanism that exploits sparse attention patterns derived from the LR trajectory to guide HR attention computation. Specifically, LRG-WinAttn restricts each HR query token to attend only within a local spatio-temporal window, thereby significantly reducing computation while maintaining essential context.

Formally, the HR attention at timestep t_n, transformer layer l, and attention head m is computed as:

$$\mathbf{A}^{\mathrm{HR}}_{n,l,m} = \mathrm{Softmax}\left(\frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d}} \odot M_{n,l,m}\right), \tag{7}$$

where $M_{n,l,m} \in \{0,1\}^{N_q \times N_k}$ is a binary attention mask indicating valid key-value positions, and $\odot$ denotes element-wise multiplication. For full attention, $M_{n,l,m} = \mathbf{1}^{N_q \times N_k}$. For sparse attention, a sparse $M_{n,l,m}$ is critical to balance efficiency and quality.

We observe that $\mathbf{z}^{\mathrm{LR}}_{t_n}$ and $\mathbf{z}^{\mathrm{HR}}_{t_n}$ are sampled from the same flow trajectory at the same timestep t_n, sharing similar spatio-temporal semantics and structural layouts. This alignment suggests that attention patterns observed in the LR domain can serve as reliable priors for guiding sparse attention at HR.

LRG-WinAttn targets to infers a sparse $M_{n,l,m}$ from $\mathbf{A}^{\mathrm{LR}}_{n,l,m}$. Notably, $\mathbf{A}^{\mathrm{LR}}_{n,l,m}$ is calculated via full attention during $\boldsymbol{v}_\theta(\mathbf{z}^{\mathrm{LR}}_{t_n}, t_n)$ and exhibits highly spatio-temporal sparsity. We expect $M_{n,l,m}$ to fully capture such sparsity with negligible quality drop. To this end, a set of 3D attention window candidates $\mathcal{W} = \{w_1, w_2, \ldots, w_J\}$ is first constructed, where each window occupies a fixed proportion s of the latent volume. For each candidate w_j, we evaluate its coverage ratio over high-response attention regions:

$$r^{\mathrm{LR}(w_j)}_{n,l,m} = \frac{\sum \mathbf{A}^{\mathrm{LR}}_{n,l,m}\, M^{\mathrm{LR}(w_j)}_{n,l,m}}{\sum \mathbf{A}^{\mathrm{LR}}_{n,l,m}}, \tag{8}$$

where $M^{\mathrm{LR}(w_j)}_{n,l,m}$ is the binary mask corresponding to w_j in LR coordinates. Then the window with the highest coverage is selected:

$$w_{n,l,m} = \arg\max_{w_j \in \mathcal{W}} r^{\mathrm{LR}(w_j)}_{n,l,m}. \tag{9}$$

ensuring that most attention mass for the query token is obtained. The selected window $w_{n,l,m}$ is subsequently mapped to the HR space and upsampled to produce the final sparse attention mask $\tilde{M}_{n,l,m}$. In this way, LRG-WinAttn constructs head-specific sparse attention masks for different transformer layers and different timesteps.

By leveraging structural alignment between LR and HR trajectories, LRG-WinAttn effectively transfers sparsity priors to the HR domain. This leads to significant inference acceleration due to reduced quadratic complexity, with negligible visual quality drop.

Low-Resolution Induced High-Order Solver

While LRG-WinAttn improves inference efficiency by reducing attention complexity, the accuracy of HR generation remains challenging due to numerical integration errors. In particular, existing RF models typically adopt the first-order Euler ODE solver (Eq. (2)), which accumulates errors over time, leading to visual quality degradation, especially in HR settings that demand precise trajectory estimation.

To this end, we propose a high-order ODE solver, termed Low-Resolution Induced High-Order Solver (LRI-HiSolver). Given that the RF model is governed by the continuous-time ODE $\frac{d\mathbf{z}_t}{dt} = v_\theta(\mathbf{z}_t, t)$, and assuming that the trajectory $\mathbf{z}_t$ is smooth with respect to time, we apply a Taylor expansion to approximate the latent at $t + \Delta t$:

$$\mathbf{z}^{\mathrm{HR}}_{t+\Delta t} = \mathbf{z}^{\mathrm{HR}}_t + \sum_{k=1}^{K} \frac{\Delta t^k}{k!} \frac{d^{(k-1)} \boldsymbol{v}_\theta(\mathbf{z}^{\mathrm{HR}}_t, t)}{dt^{(k-1)}}, \tag{10}$$

Method	Speedup	Time (Minutes)	Temporal Flickering	Motion Smoothness	Background Consistency	Overall Consistency	Aesthetic Quality	Imaging Quality	Mean
HunyuanVideo	1×	61	**0.9960**	**0.9956**	**0.9823**	0.1843	0.4334	0.5865	0.6963
DiffuseHigh	1.4×	44	0.9893	0.9933	0.9721	<u>0.2217</u>	0.5736	0.6557	0.7342
HiFlow	1.3×	46	0.9912	0.9940	<u>0.9741</u>	0.2167	<u>0.5799</u>	0.6616	0.7363
MACS	1.4×	44	0.9904	0.9938	0.9734	0.2215	0.5767	<u>0.6639</u>	<u>0.7366</u>
S^{2}Flow (Ours)	**4×**	**15**	<u>0.9914</u>	<u>0.9942</u>	0.9738	**0.2264**	**0.5815**	**0.6660**	**0.7389**

Table 1: Comparison with state-of-the-art methods on the VBench dataset. The best result is bolded, and the second best is underlined.

where K denotes the solver order. For brevity, we omit the time step subscript n. In this work, we adopt the second-order case (*i.e.*, $K = 2$), which is empirically sufficient to yield high-fidelity synthesis. However, computing the time derivative $\frac{d\boldsymbol{v}_\theta}{dt}$ directly is intractable due to the implicit and highly nonlinear form of $\boldsymbol{v}_\theta$. To this end, we adopt a central difference scheme for approximation:

$$\frac{d\boldsymbol{v}_\theta(\mathbf{z}_t^{\mathrm{HR}}, t)}{dt} \approx \frac{\boldsymbol{v}_\theta(\mathbf{z}_{t+\Delta t}^{\mathrm{HR}}, t+\Delta t) - \boldsymbol{v}_\theta(\mathbf{z}_{t-\Delta t}^{\mathrm{HR}}, t-\Delta t)}{2\Delta t}. \tag{11}$$

In Eq. (11), the future latent $\mathbf{z}_{t+\Delta t}^{\mathrm{HR}}$ is unavailable at time t during inference. To overcome this, we exploit the temporal alignment and semantic consistency between LR and HR trajectories, and substitute the unavailable HR states with their LR counterparts: $\boldsymbol{v}_\theta(\overset{\mathrm{HR}}{\mathbf{z}_{t\pm\Delta t}}, t\pm\Delta t) \approx \mathcal{U}(\boldsymbol{v}_\theta(\overset{\mathrm{LR}}{\mathbf{z}_{t\pm\Delta t}}, t\pm\Delta t))$, where $\mathcal{U}(\cdot)$ is an unsampling operation in the latent space. This enables efficient estimation of high-order derivatives without introducing additional HR forward passes. Accordingly, Eq. (10) becomes:

$$\begin{aligned}\mathbf{z}_{t+\Delta t}^{\mathrm{HR}} &= \mathbf{z}_t^{\mathrm{HR}} + \Delta t\, \boldsymbol{v}_\theta(\mathbf{z}_t^{\mathrm{HR}}, t) + \frac{\Delta t^2}{2}\,\frac{d\boldsymbol{v}_\theta}{dt},\\ \frac{d\boldsymbol{v}_\theta}{dt} &= \mathcal{U}\Big(\frac{\boldsymbol{v}_\theta(\mathbf{z}_{t+\Delta t}^{\mathrm{LR}}, t+\Delta t) - \boldsymbol{v}_\theta(\mathbf{z}_{t-\Delta t}^{\mathrm{LR}}, t-\Delta t)}{2\Delta t}\Big).\end{aligned} \tag{12}$$

Eq. (12) provides a second-order correction to the latent update. This reduces the local truncation error from $O(\Delta t^2)$ in Euler method to $O(\Delta t^3)$ per step.

After denoising from t_i to t_0 using both LRI-HiSolver and LRG-WinAttn, the final HR latent $\mathbf{z}_{t_0}^{\mathrm{HR}}$ is decoded into pixel space via the VAE decoder $\mathcal{D}$:

$$\mathbf{x}_0^{\mathrm{HR}} = \mathcal{D}(\mathbf{z}_0^{\mathrm{HR}}). \tag{13}$$

In summary, LRI-HiSolver exploits the semantic consistency of LR trajectories to estimate high-order temporal derivatives, thereby improving HR ODE solving accuracy with negligible additional cost. Combined with sparse attention control via LRG-WinAttn, our S^{2}Flow achieves high-fidelity and efficient high-resolution video generation without retraining.

Experimental Results

Experimental Settings

Datasets and Evaluation Metrics. We evaluate our method on the VBench benchmark (Huang et al. 2024b), which includes 946 diverse prompts with rich semantics and motion. Due to the slow inference speed of some baselines, evaluating all methods on the full set is impractical. We therefore randomly sample 473 prompts for fair and feasible comparison. Evaluation spans three levels: temporal-level (temporal flickering, motion smoothness, background consistency), video-level (overall consistency), and frame-level (aesthetic quality, imaging quality). The mean of all metrics is reported for overall comparison. In ablation study, we aggregate the above metrics into temporal quality and visual fidelity for the convenience of analysis. Details of all metrics are provided in the supplementary material.

To complement automatic evaluation, which may not fully capture human perception, we also conduct a user study to assess visual quality based on human preferences. Detailed results are provided in the supplementary material.

Implementation Details. We adopt HunyuanVideo (Kong et al. 2024) as our baseline model, originally trained with videos of up to 1280×768 resolution. The inference timestep number is set as 50, and the frame number is 117. The hyper-parameters of the upscaling step t_i and window sparsity s are set to 30 and 50%, which are analyzed in the ablation studies. If not specified, other hyper-parameters are set as default. The window attention kernels are implemented based on ThunderKittens (Spector et al. 2025). All experiments are conducted on $8\times$ NVIDIA H800 GPUs.

Comparison with State-of-the-Art Methods

We compare S^{2}Flow with recent state-of-the-art training-free methods, including DiffuseHigh (Kim et al. 2025), HiFlow (Bu et al. 2025), and Make-A-Cheap-Scaling (MACS) (Guo et al. 2024), which are representative and compatible with HunyuanVideo. Other methods tightly coupled with specific backbones are excluded for fairness. All approaches are evaluated on the same backbone, HunyuanVideo pretrained at 1280×768, and assessed by generating videos at 2560×1536. As shown in Table 1, S^{2}Flow consistently outperforms existing methods across multiple metrics. In addition, it achieves substantially lower inference latency, offering significant acceleration while preserving generation quality. These results validate the effectiveness of leveraging low-resolution latent trajectories to guide high-resolution synthesis. By distilling semantic and motion cues from low-resolution dynamics, and integrating sparse attention with a second-order ODE solver, S^{2}Flow enables precise and efficient video generation without retraining. Notably, while HunyuanVideo baseline exhibits high temporal consistency due to its near-static outputs, it suffers from de-

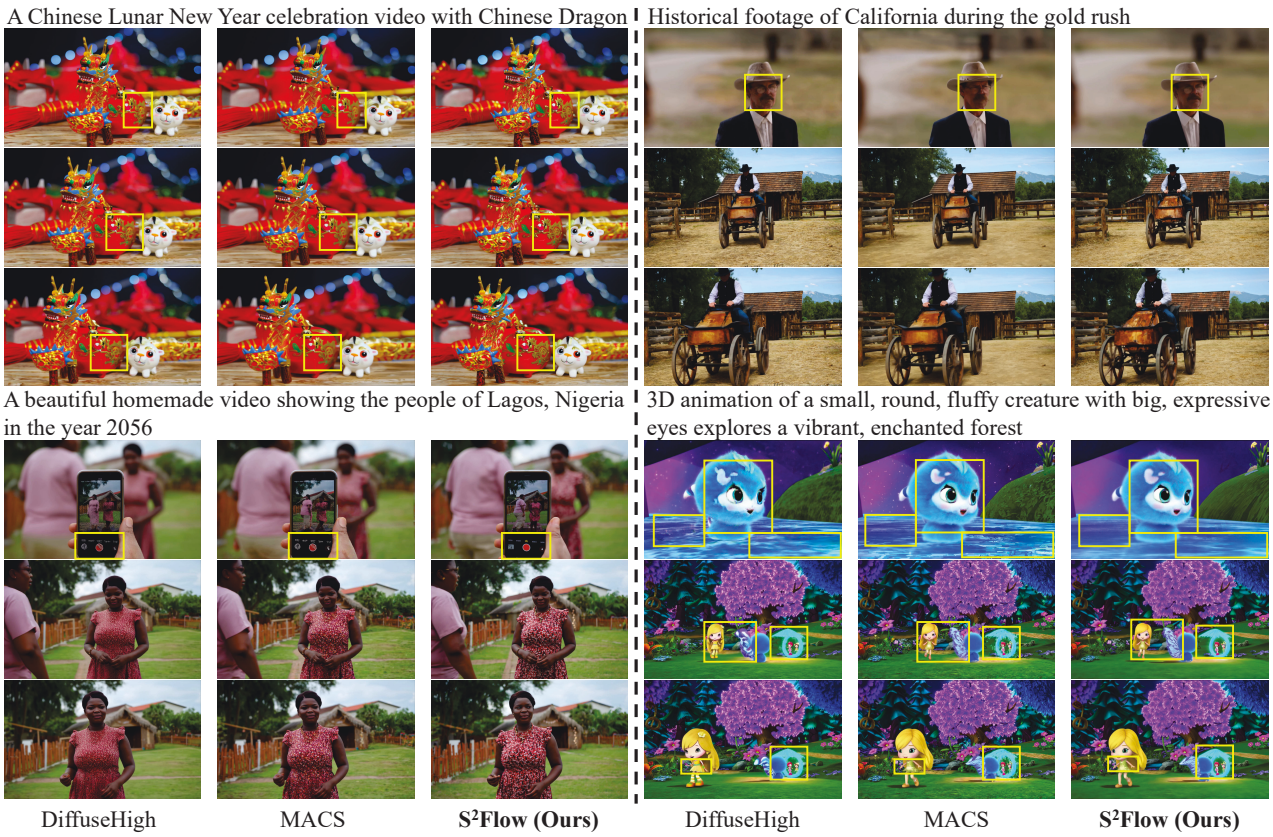


Figure 4: Qualitative comparison with SOTA training-free methods using the HunyuanVideo backbone. Yellow boxes indicate key differences.

Sparsity s	Temporal	Visual	Mean	Speedup
25%	0.7894	0.4605	0.6797	5×
50%	0.7964	0.6238	0.7389	4×
75%	0.7959	0.6293	0.7404	1.8×
100%	0.7960	0.6328	0.7416	1×
50%∗	Fail to Generate			4×

Table 2: Effect of varying window sparsity ratio s.

graded visual quality, primarily due to the distribution mismatch introduced by performing high-resolution inference on a low-resolution-trained model.

Qualitative results are shown in Fig. 4. S^{2}Flow produces more coherent and temporally stable videos, while competing methods often suffer from structural distortions and blurring. In summary, both quantitative and qualitative evaluations demonstrate the generalizability and practical effectiveness of S^{2}Flow.

Ablation Study

Effect of Low-Resolution Guided Window Attention. To reduce the quadratic cost of full attention in high-resolution (HR) inference, we introduce LRG-WinAttn, which restricts each HR query to a localized spatio-temporal window adaptively guided by low-resolution (LR) attention patterns. For each attention head and timestep, the window is dynamically selected based on LR attention responses, enabling content-aware sparsity. We compare three variants: 1) full attention ("100%" in Table 2), where each HR query attends to all tokens, 2) fixed-window attention ("50%*"), using the same sparsity ratio $s = 50\%$ as LRG-WinAttn but with uniform, non-adaptive windows, and 3) our LRG-WinAttn ("50%"). As shown in Table 2, full attention achieves marginally better quality but at high cost. Fixed-window attention fails completely across tested shapes, producing black frames due to its inability to adapt to varying structure and motion. In contrast, our LRG-WinAttn achieves a superior balance between efficiency and fidelity. Its dynamic windowing mechanism preserves essential semantic and structural dependencies while significantly reducing inference time.

Effect of Window Sparsity Ratio s. We further evaluate the impact of sparsity ratio s in LRG-WinAttn, which determines the proportion of spatio-temporal tokens each HR query attends to. As shown in Table 2, increasing s introduces more global context and improves generation quality, at the cost of higher computation. Conversely, reducing s accelerates inference but may compromise fidelity. Notably, performance drops significantly when s decreases from 50% to 25%, indicating loss of critical contextual information. Therefore, we adopt $s = 50\%$ as the default, which con-

Order	Temporal	Visual	Mean
1st	0.7891	0.6175	0.7319
2nd	0.7964	0.6238	0.7389
3rd	0.7963	0.6243	0.7390

Table 3: Effect of LRI-HiSolver on high-res generation.

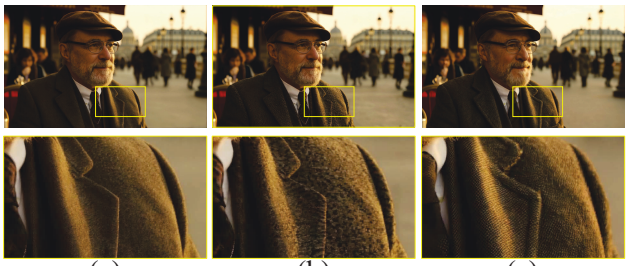


Figure 5: Visual comparison across different configurations: (a) full attention, (b) our LRG-WinAttn, and (c) LRG-WinAttn with LRI-HiSolver.

t_i	Temporal	Visual	Mean
40	0.6774	0.4694	0.6080
30	0.7964	0.6238	0.7389
20	0.7910	0.6175	0.7332
10	0.7900	0.5016	0.6939

Table 4: Effect of the transition step t_i from low- to high-resolution generation.

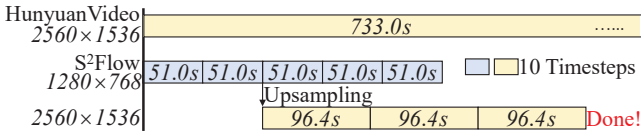


Figure 6: Runtime comparison of S^{2}Flow and HunyuanVideo. Each block represents 10 denoising steps. S^{2}Flow significantly accelerates inference at 2560×1536 resolution. To focus on model efficiency, data preprocessing and upsampling time are excluded.

sistently provides strong generation quality under realistic computational constraints.

Effect of Low-Resolution Induced High-Order Solver. To improve HR denoising accuracy, we introduce a high-order solver guided by LR trajectory estimates. We compare first-, second-, and third-order solvers under identical settings. As shown in Table 3, the second-order solver significantly enhances visual fidelity over the commonly-used first-order solver, owing to its more accurate modeling of latent dynamics enabled by LR-guided correction. Importantly, this improvement incurs no additional HR forward passes, as higher-order terms are derived from the precomputed LR trajectory. The third-order solver yields only marginal further gains, indicating that second-order correction already captures the primary error components.

Visual Effect of LRG-WinAttn and LRI-HiSolver. To better understand the individual contributions of our two core components, LRG-WinAttn and LRI-HiSolver, we conduct a visual comparison across three configurations: 1) full attention during HR inference, 2) applying LRG-WinAttn, and 3) further incorporating LRI-HiSolver. As shown in Fig. 5, replacing full attention with LRG-WinAttn slightly reduces local detail, particularly in fine textures such as fabric patterns, which become marginally coarser. However, this trade-off yields substantial inference acceleration. Upon introducing LRI-HiSolver, the high-frequency details are significantly recovered, restoring texture fidelity without noticeably increasing inference time. These results highlight that our full pipeline effectively balances visual quality and computational efficiency.

Effect of Upsampling Step t_i. In our framework, the transition from LR to HR latent space occurs at a predefined denoising step t_i. We evaluate the impact of different choices of t_i, with results summarized in Table 4. As shown, both excessively early and overly delayed upsampling degrade performance. Upsampling too early leads to unstable HR synthesis, as the latent representation has not yet captured sufficient semantic structure. Conversely, delaying upsampling causes a loss of spatial detail, since the high-frequency components are not fully recovered in the later stages. Empirically, setting $t_i = 30$ (out of 50 total steps) provides the best trade-off.

Inference Efficiency Analysis. We evaluate the efficiency of S^{2}Flow by measuring the average runtime per 10 denoising steps at different resolutions. As shown in Fig. 6, S^{2}Flow takes 51.0s at 1280×768 and 96.4s at 2560×1536, achieving a $7.6\times$ transformer-level speedup over HunyuanVideo (733.0s). This gain is primarily attributed to sparse attention and reduced GPU data communication. Including data processing, cascaded framework, and upsampling operations, the overall pipeline speedup reaches $4\times$ (Table1).

Conclusion

In this paper, we propose S^{2}Flow, a training-free framework for high-resolution video generation that leverages sparse attention and second-order ODE solving, both guided by low-resolution flow trajectories. To improve efficiency, S^{2}Flow introduces Low-Resolution Guided Window Attention that adaptively constructs localized spatio-temporal attention windows from low-resolution features, significantly reducing computational redundancy while preserving contextual dependencies. To enhance accuracy, we incorporate a Low-Resolution Induced High-Order Solver where the high-order derivative term is approximated using central difference derived from low-resolution estimates, enabling more precise high-resolution denoising. Extensive experiments on VBench dataset demonstrate that S^{2}Flow achieves superior visual quality and notable acceleration compared to existing approaches, effectively scaling video generation up to 2560×1536 resolution. Despite its effectiveness, S^{2}Flow may underperform on extremely fine-grained details due to sparsity. Future work will explore training-based extensions to enhance fidelity under such conditions.

Acknowledgments

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation (Nos. 2024A1515140109 and 2023A1515110695), in part by National Natural Science Foundation of China (No. 62571393), and in part by Key Research and Development Program of Shaanxi (No. 2024GX-YBXM-127).

References

Bain, M.; Nagrani, A.; Varol, G.; and Zisserman, A. 2021. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, 1728–1738.

Bar-Tal, O.; Yariv, L.; Lipman, Y.; and Dekel, T. 2023. Multidiffusion: Fusing diffusion paths for controlled image generation. In *International Conference on Machine Learning*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 22563–22575.

Bu, J.; Ling, P.; Zhou, Y.; Zhang, P.; Wu, T.; Dong, X.; Zang, Y.; Cao, Y.; Lin, D.; and Wang, J. 2025. HiFlow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*.

Cao, B.; Ye, J.; Wei, Y.; and Shan, H. 2024. AP-LDM: Attentive and progressive latent diffusion model for training-free high-resolution image generation. *arXiv preprint arXiv:2410.06055*.

Chen, H.; Zhang, Y.; Cun, X.; Xia, M.; Wang, X.; Weng, C.; and Shan, Y. 2024. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7310–7320.

Du, R.; Chang, D.; Hospedales, T.; Song, Y.-Z.; and Ma, Z. 2024. Demofusion: Democratising high-resolution image generation with no $. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6159–6168.

Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. 2024. Scaling rectified flow Transformers for high-resolution image synthesis. In *International Conference on Machine Learning*, 1–16.

Guo, L.; He, Y.; Chen, H.; Xia, M.; Cun, X.; Wang, Y.; Huang, S.; Zhang, Y.; Wang, X.; Chen, Q.; et al. 2024. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European Conference on Computer Vision*, 39–55.

Haji-Ali, M.; Balakrishnan, G.; and Ordonez, V. 2024. Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6603–6612.

He, Y.; Yang, S.; Chen, H.; Cun, X.; Xia, M.; Zhang, Y.; Wang, X.; He, R.; Chen, Q.; and Shan, Y. 2023. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *International Conference on Learning Representations*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022. Video diffusion models. In *Advances in Neural Information Processing Systems*, 8633–8646.

Huang, L.; Fang, R.; Zhang, A.; Song, G.; Liu, S.; Liu, Y.; and Li, H. 2024a. Fouriscale: A frequency perspective on training-free high-resolution image synthesis. In *European Conference on Computer Vision*, 196–212.

Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024b. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 21807–21818.

Jin, Z.; Shen, X.; Li, B.; and Xue, X. 2023. Training-free diffusion model adaptation for variable-sized text-to-image synthesis. In *Advances in Neural Information Processing Systems*, 70847–70860.

Kim, Y.; Hwang, G.; Zhang, J.; and Park, E. 2025. DiffuseHigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 4338–4346.

Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. HunyuanVideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.

Lee, Y.; Kim, K.; Kim, H.; and Sung, M. 2023. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *Advances in Neural Information Processing Systems*, 50648–50660.

Liang, J.; Fan, Y.; Zhang, K.; Timofte, R.; Van Gool, L.; and Ranjan, R. 2024. Movideo: Motion-aware video generation with diffusion model. In *European Conference on Computer Vision*, 56–74.

Liu, X.; Gong, C.; and Liu, Q. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*.

Lu, E.; Jiang, Z.; Liu, J.; Du, Y.; Jiang, T.; Hong, C.; Liu, S.; He, W.; Yuan, E.; Wang, Y.; Huang, Z.; Yuan, H.; Xu, S.; Xu, X.; Lai, G.; Chen, Y.; Zheng, H.; Yan, J.; Su, J.; Wu, Y.; Zhang, Y.; Yang, Z.; Zhou, X.; Zhang, M.; and Qiu, J. 2025. MoBA: Mixture of block attention for long-context LLMs. *arXiv preprint arXiv:2502.13189*.

Menapace, W.; Siarohin, A.; Skorokhodov, I.; Deyneka, E.; Chen, T.-S.; Kag, A.; Fang, Y.; Stoliar, A.; Ricci, E.; Ren, J.; et al. 2024. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7038–7048.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, 1–21.

Qing, Z.; Zhang, S.; Wang, J.; Wang, X.; Wei, Y.; Zhang, Y.; Gao, C.; and Sang, N. 2024. Hierarchical spatio-temporal decoupling for text-to-video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6635–6645.

Qiu, H.; Zhang, S.; Wei, Y.; Chu, R.; Yuan, H.; Wang, X.; Zhang, Y.; and Liu, Z. 2024. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. *arXiv preprint arXiv:2412.09626*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 36479–36494.

Skorokhodov, I.; Menapace, W.; Siarohin, A.; and Tulyakov, S. 2024. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7569–7579.

Spector, B. F.; Arora, S.; Singhal, A.; Parthasarathy, A.; Fu, D. Y.; and Re, C. 2025. ThunderKittens: Simple, fast, and adorable kernels. In *International Conference on Learning Representations*.

Vontobel, T.; Sadat, S.; Salehi, F.; and Weber, R. M. 2025. HiWave: Training-free high-resolution image generation via wavelet-based diffusion sampling. *arXiv preprint arXiv:2506.20452*.

Wan, T.; Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; et al. 2025. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2024a. LaVie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 1–20.

Wang, Y.; He, Y.; Li, Y.; Li, K.; Yu, J.; Ma, X.; Li, X.; Chen, G.; Chen, X.; Wang, Y.; Luo, P.; Liu, Z.; Wang, Y.; Wang, L.; and Qiao, Y. 2024b. InternVid: A large-scale video-text dataset for multimodal understanding and generation. In *International Conference on Learning Representations*.

Wang, Z.; Li, A.; Zhu, L.; Guo, Y.; Dou, Q.; and Li, Z. 2024c. Customvideo: Customizing text-to-video generation with multiple subjects. *arXiv preprint arXiv:2401.09962*.

Wu, J.; Li, X.; Zeng, Y.; Zhang, J.; Zhou, Q.; Li, Y.; Tong, Y.; and Chen, K. 2024. MotionBooth: Motion-aware customized text-to-video generation. In *Conference on Neural Information Processing Systems*.

Xing, J.; Xia, M.; Liu, Y.; Zhang, Y.; Zhang, Y.; He, Y.; Liu, H.; Chen, H.; Cun, X.; Wang, X.; et al. 2025. Make-your-video: Customized video generation using textual and structural guidance. *IEEE Transactions on Visualization and Computer Graphics*.

Zhang, P.; Chen, Y.; Su, R.; Ding, H.; Stoica, I.; Liu, Z.; and Zhang, H. 2025. Fast video generation with sliding tile attention. In *International Conference on Machine Learning*.

Zheng, Q.; Guo, Y.; Deng, J.; Han, J.; Li, Y.; Xu, S.; and Xu, H. 2024. Any-size-diffusion: Toward efficient text-driven synthesis for any-size HD images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 7571–7578.