FlashVideo: Flowing Fidelity to Detail for Efficient High-Resolution Video Generation

Shilong Zhang[1*], Wenbo Li[2*], Shoufa Chen[1], Chongjian GE[1], Peize Sun[1], Yifu Zhang[3], Yi Jiang[3†], Zehuan Yuan[3], Bingyue Peng[3], Ping Luo[1 ‡]

[1]The University of Hong Kong
[2]The Chinese University of Hong Kong
[3]ByteDance

Abstract

DiT models have achieved great success in text-to-video generation, leveraging their scalability in model capacity and data scale. High content and motion fidelity aligned with text prompts, however, often require large model parameters and a substantial number of function evaluations (NFEs). Realistic and visually appealing details are typically reflected in high-resolution outputs, further amplifying computational demands—especially for single-stage DiT models. To address these challenges, we propose a novel two-stage framework, FlashVideo, which strategically allocates model capacity and NFEs across stages to balance generation fidelity and quality. In the first stage, prompt fidelity is prioritized through a low-resolution generation process utilizing large parameters and sufficient NFEs to enhance computational efficiency. The second stage achieves a nearly straight ODE trajectory between low and high resolutions via flow matching, effectively generating fine details and fixing artifacts with minimal NFEs. To ensure a seamless connection between the two independently trained stages during inference, we carefully design degradation strategies during the second-stage training. Quantitative and visual results demonstrate that FlashVideo achieves state-of-the-art high-resolution video generation with superior computational efficiency. Additionally, the two-stage design enables users to preview the initial output and accordingly adjust the prompt before committing to full-resolution generation, thereby significantly reducing computational costs and wait times as well as enhancing commercial viability.

Code — https://github.com/FoundationVision/FlashVideo

1 Introduction

In recent years, text-to-video (T2V) generation has achieved remarkable progress, driven by advances in diffusion probabilistic modeling (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Liu, Gong, and Liu 2022; Lipman et al. 2022), cutting-edge architectures (Ronneberger, Fischer, and Brox 2015; Peebles and Xie 2022), and the integration of extensive model parameters and large-scale datasets (He et al. 2022; Hong et al. 2022; Chen et al. 2023, 2024; Kondratyuk et al. 2024; Zheng et al. 2024; Yang et al. 2024b). Among these, DiT-based models (Peebles and Xie 2022) stand out for their scalability in accommodating larger model capacities and datasets.

In video DiTs, the key operator is the 3D full attention across time (T), height (H), and width (W), which effectively models visual relations in scenarios with large object motions and 3D consistency. The computational complexity scales as $\mathcal{O}(T^2H^2W^2 \cdot C \cdot N)$, where C represents the feature dimension (linked to model size) and N is the number of denoising steps (function evaluation). State-of-the-art methods (Kong et al. 2024; Yang et al. 2024b; Team 2025) require large model capacities, high-resolution modeling, and up to 50 denoising steps for high-quality outputs.

These requirements arise from the need to tackle key challenges in video generation, particularly ensuring high prompt fidelity and visual quality. First, achieving fidelity in both content and motion demands the model to encode extensive world knowledge. Research has shown significant improvements when increasing model parameters (C) from 2 billion to 12 billion (Yang et al. 2024b; Kong et al. 2024). Additionally, an adequate number of denoising steps (N) (Kong et al. 2024; Yang et al. 2024b) is essential for generating high-quality videos. While some efforts to reduce the number of steps have shown promising progress (Ding et al. 2024), they are limited to lower resolutions and simpler motions. Moreover, visual quality has been proven to be tightly tied to resolution in text-to-image generation ($H \times W$) (Blattmann et al. 2023b; Chen et al. 2025; Ren et al. 2024), and for T2V tasks, the integrity of motion (T) must also be maintained. However, the combination of these challenges—large parameters, sufficient denoising steps, and high resolution—significantly increases the computational cost. For instance, a 5-billion-parameter model takes 2150s to generate 1080p videos, up from just 30s at the 270p resolution (Figure 1 (d)).

To overcome these challenges, we introduce FlashVideo, a two-stage framework designed to separately optimize prompt fidelity and visual quality, as illustrated in Fig-

[*]These authors contributed equally.
[†]Project leader.
[‡]Corresponding author.

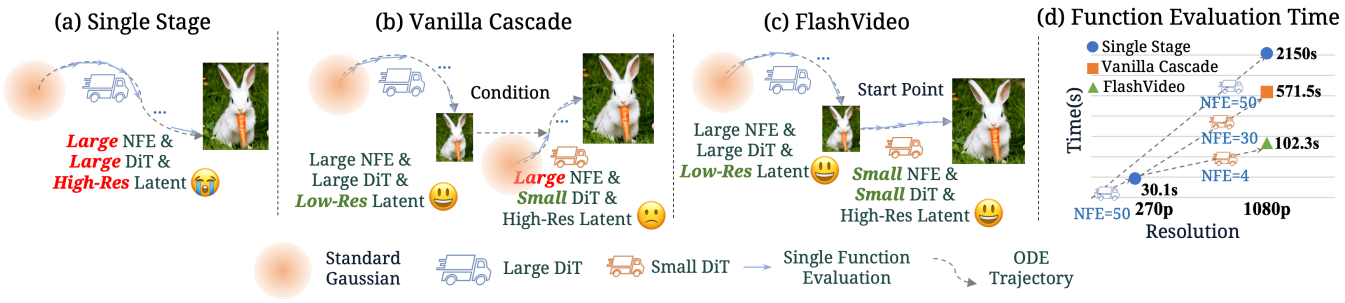

Figure 1: **Comparison between FlashVideo and other paradigms.** (a) Single Stage DiT suffers from an explosive increase in computation cost when generating at large resolutions, rising from 30s to 2150s (blue circle in (d)) when increasing the resolution from 270p to 1080p. (b) Though the vanilla cascade can reduce the model size in high resolution, its second stage still samples from Gaussian noise and only uses the first-stage results as a condition. This approach cannot effectively reduce the number of function evaluations at high resolution and still costs 571.5s (earthyellow square in (d)) to generate a 1080p video. (c) In contrast, FlashVideo not only decreases the model size in the second stage but also starts sampling from the first-stage results, requiring only 4 function evaluations at high resolution while integrating a wealth of visually pleasant details, which can generate 1080P video with only 102.3s (green triangle in (d)). Details can be found in extended version .

ure 1 (c). In the first stage, we focus on generating video content and motion that closely aligns with the user prompt. By operating at a lower resolution (*e.g.*, 270p), even though we utilize a large model with 5 billion parameters with 50 evaluation steps, the model still remains efficient, requiring only 30 seconds of function evaluation times (as shown in Figure 1 (d)). And as demonstrated in our experiments (Section 4.2), this approach preserves semantic fidelity and motion smoothness. In the second stage, we enhance the generated video at 1080p, focusing on fine-grained detail enhancement while minimizing computational overhead. This is achieved using a lighter 2-billion-parameter model and an efficient flow-matching process with fewer evaluation steps. The two-stage framework effectively balances computational efficiency with high-quality results.

While previous two-stage frameworks (Zhou et al. 2024; Wang et al. 2023; He et al. 2024) treat the first-stage low-resolution output as a condition and begin the second stage from Gaussian noise (Figure 1 (c)), this design requires 30–50 evaluation steps and still incurs significant computational cost (*e.g.*, 571 seconds for 1080p generation). In contrast, FlashVideo uses flow matching to directly traverse ODE trajectories from first stage low-quality video to the final high-quality videos, eliminating the need to start from Gaussian noise. The flow matching target also tries to constrain the ODE trajectories to be straight. This design efficiently reduces the number of function evaluations to just 4 steps. As a result, FlashVideo reduces the function evaluation time for 1080p videos to just 102s, nearly $1/20$ of the time required by a single-stage model (Figure 1 (a)), and 5 times faster than vanilla cascade frameworks (Figure 1 (b)).

Despite this speedup, crafting an appropriate degradation simulation remains essential: naive pixel-space operations (*e.g.*, resizing and blurring (Chan et al. 2022; Yang et al. 2024a)) fail to reproduce the fine-scale artifacts introduced by model synthesis. To this end, we inject noise in the latent representation to eliminate all small structures in real videos, forcing their faithful reconstruction in the second stage. Moreover, we demonstrate that—within 3D full-attention architectures, which are critical for spatio-temporal coherence—the degradation intensity must be carefully tuned to prevent the model from "cheating" by merely copying details across frames, a pitfall neglected in prior work (Pernias et al. 2023; Yu et al. 2024; Wang et al. 2023; He et al. 2024).

FlashVideo achieves top-tier performance on VBench-Long (83.29 score) while achieving impressive function evaluation time. The two-stage design allows users to preview initial output and accordingly adjust the prompt before committing to full-resolution generation, thereby significantly reducing computational costs and wait times for creating satisfactory videos.

2 Related Work

Video generation models. Recent advancements in text-to-video (T2V) generation have been remarkable (Yan et al. 2021; Hong et al. 2022; Kondratyuk et al. 2024; Ho et al. 2022b; Blattmann et al. 2023b,a; Jin et al. 2024; Team 2025). Key breakthroughs have been driven by the introduction of video diffusion and flow-matching algorithms (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Liu, Gong, and Liu 2022; Lipman et al. 2022), alongside scaled text-video datasets and DiT parameters (Peebles and Xie 2022). Despite impressive generation quality, a major challenge remains the high computational cost for generating high-resolution videos.

Cascade diffusion models. Numerous attempts have been made to explore cascade architectures in the text-to-image and text-to-video domains (Saharia et al. 2022; Gu et al. 2023; Ho et al. 2022a; Pernias et al. 2023; Yu et al. 2024; Zhang et al. 2023; Wang et al. 2023; He et al. 2024). Researchers are motivated by the challenge that generating high-resolution images/videos in a single stage is both difficult and resource-intensive. In a cascade design, generation starts with a low-resolution sample, followed by an upsam-

pling model to enhance visual appeal at higher resolutions. However, most methods perform the second-stage upsampling from pure noise, conditioning it on the low-resolution input, which requires a large number of function evaluations. While (Teng et al. 2023; Zhang et al. 2023; Xing et al. 2024) have attempted to start from the first-stage distribution, their theories and implementations are complex, resulting in a high number of inference steps. Moreover, (Fischer et al. 2023) proposes a pure super-resolution method for T2I using flow matching, but the limited generative priors in the second-stage model hinder substantial visual improvements. In this paper, we adhere to the principle of retaining only the most effective designs, developing FlashVideo, an efficient yet simple two-stage framework that achieves high-quality, high-resolution video generation with excellent computational efficiency. More related works on diffusion model acceleration can be found in our extended version.

3 Method

3.1 Overview

In the FlashVideo framework, video pixels $x \in \mathbb{R}^{H\times W\times T}$ are first compressed into latent features $f \in \mathbb{Q}^{h\times w\times t}$ using a 3D causal VAE (Yang et al. 2024b), where $h = {}^{H}/_{8}$, $w = {}^{W}/_{8}$, and $t = {}^{(T-1)}/_{4} + 1$. The model is designed to generate 6-second videos (with 8 frames per second, so $T = 49$) at 1080p resolution. As shown in Figure 2, we then employ a two-stage, low-to-high-resolution generation pipeline, where each stage is optimized with tailored model sizes and training strategies to ensure computational efficiency.

3.2 Low-Resolution Stage I for Prompt Fidelity

In the first stage, the goal is to generate videos with well-aligned content and motion corresponding to the input prompt. To achieve this, we initialize with a large-capacity model, CogVideoX-5B (Yang et al. 2024b), which contains 5 billion parameters. For improved semantic fidelity and computational efficiency, we perform parameter-efficient fine-tuning (PEFT) to adapt the model to a lower resolution of 270p. We find that adjusting the target resolution of the MMDiT architecture (Esser et al. 2024) is straightforward, which is achieved by applying LoRA (Hu et al. 2021) with rank 128 to all attention (Vaswani 2017), FFN, and adaptive layer normalization (Perez et al. 2018) layers. Compared to full-parameter tuning, PEFT demonstrates greater robustness, especially when fine-tuned with a small batch size of 32. In contrast, full-parameter tuning with such a small batch size significantly degrades generation quality. All other configuration settings, including the denoising scheduler and prediction target, are kept consistent with CogVideoX-5B. By reducing the resolution from 480p to 270p, our Stage I model achieves a significantly higher VBench semantic score(82.03) than the original CogVideoX (77.04), along with nearly 3× speed-up, as the model no longer needs to focus on high-frequency details .

3.3 High-Resolution Stage II for Details

Model architecture. For fine-grained detail enhancement, we employ another model that adheres to the block design specified in CogVideoX-2B (Yang et al. 2024b).

But, we replace the original absolute position embedding with 3D RoPE (Su et al. 2024), as it offers better scalability for higher resolutions during training and inference (see Figure 3 and Section. 5). Unlike the previous approaches in (Wang et al. 2023; Zhang et al. 2023; He et al. 2024), which use a spatial-temporal decomposition framework, we find that utilizing full 3D attention is crucial for maintaining consistency of enhanced visual details in videos with significant motion and scale variance, as shown in Figure 6. As illustrated in Figure 2, the language embedding from the first stage is directly utilized in this stage.

Low-cost resolution transport. Applying the conventional diffusion process at the high-resolution stage—starting from Gaussian noise and conditioned on low-resolution video—demands substantial computational resources. To improve efficiency while maintaining high-quality detail generation, we adopt flow matching (Liu, Gong, and Liu 2022; Lipman et al. 2022) to map the low-resolution latent representation, $\mathbf{Z}_{LR}$, to the high-resolution latent representation, $\mathbf{Z}_{HR}$. Intermediate points are computed through linear interpolation between $\mathbf{Z}_{LR}$ and $\mathbf{Z}_{HR}$, as outlined in Algorithm 1.

Algorithm 1: Training Stage

Input: High quality video dataset D_{HR}, model F_θ, VAE encoder $\mathcal{E}$
Output: Model F_θ

repeat
  Sample $\mathbf{X}_{HR} \sim D_{HR}$;
  $\mathbf{Z}_{HR} \leftarrow \mathcal{E}(\mathbf{X}_{HR})$;
  $\mathbf{Z}_{LR} \leftarrow DEG_{latent}(\mathcal{E}(DEG_{pixel}(\mathbf{X}_{HR})))$;
  $Target \leftarrow \mathbf{Z}_{HR} - \mathbf{Z}_{LR}$;
  $t \sim Uniform([0, 1])$;
  $\mathbf{Z}_t \leftarrow (1-t)\cdot \mathbf{Z}_{LR} + t \cdot \mathbf{Z}_{HR}$;
  Take gradient descent step on $\nabla_\theta \left\| Target - F_\theta(\mathbf{Z}_t, t)\right\|^2$;
until *converged*;

This approach eliminates redundant sampling steps at the initialization phase and avoids reliance on additional control parameters, such as those proposed in (Zhang, Rao, and Agrawala 2023; Yu et al. 2024; He et al. 2024). Furthermore, the t-independent target $\mathbf{Z}_{HR} - \mathbf{Z}_{LR}$ results in straighter ODE trajectories, enabling few-step generation. During training, $\mathbf{Z}_{LR}$ is simulated, as discussed later. In the testing phase, noise-augmented videos generated in the first stage serve as the starting point, and a commonly used Euler solver with $S = 4$ steps, as outlined in Algorithm 2, is employed. Other higher-order solvers can also be used for practical applications.

Low quality video simulation. To train the second-stage model, we establish paired low-resolution and high-resolution latent representations, $\mathbf{Z}_{LR}$ and $\mathbf{Z}_{HR}$. Starting

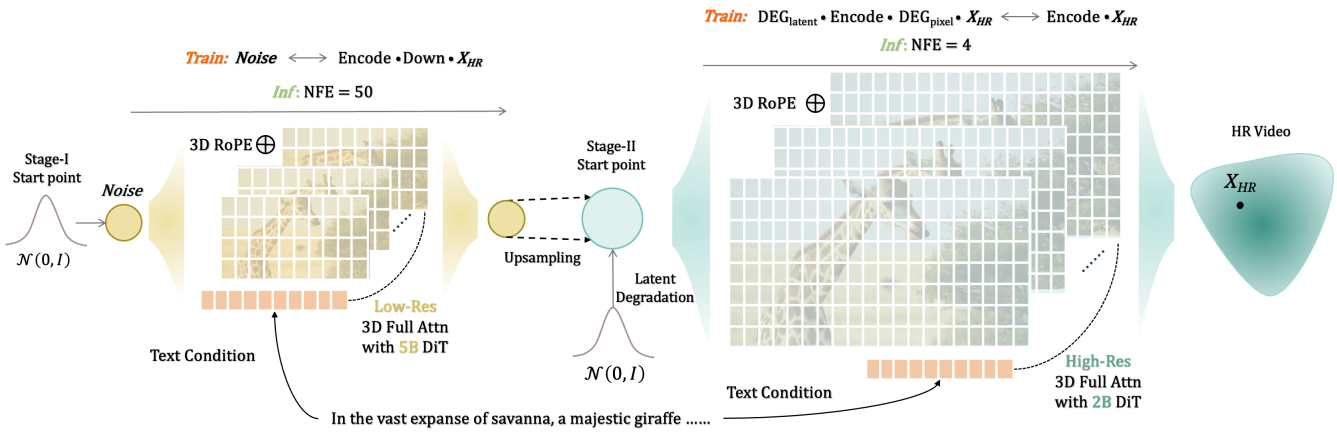

Figure 2: **The overall pipeline of FlashVideo**. FlashVideo adopts a cascade paradigm comprised of a 5-billion-parameter DiT at the low resolution (*i.e.*, Stage I) and a 2-billion-parameter DiT at a higher resolution (*i.e.*, Stage II). The 3D RoPE is employed at both stages to model the global and relative spatiotemporal distances efficiently. We construct training data pairs for Stage I by randomly sampling Gaussian noise and low-resolution video latent. For Stage II, we apply both pixel and latent degradation to high-quality videos to obtain low-quality latent values. These are then paired with high-quality latents to serve as training data. During inference, we retain a sufficient $NFE = 50$ at a low resolution of 270p for Stage I. The generated videos retain high fidelity and seamless motion, albeit with detail loss. These videos are then upscaled to a higher resolution of 1080p and processed by latent degradation. With only 4 steps, our Stage II regenerates accurate structures and rich high-frequency details.

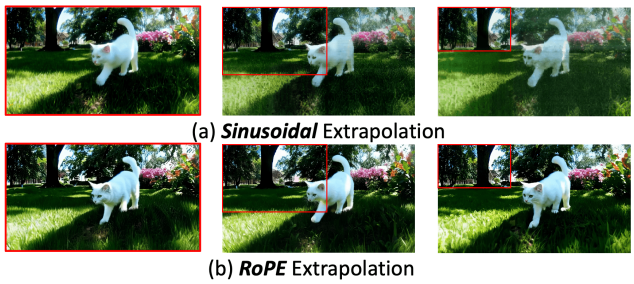

Figure 3: **Resolution extrapolation with different position embeddings.** RoPE maintains details at higher resolutions; absolute embeddings cause artifacts beyond training resolution.

Algorithm 2: Inference Stage

Input: Low-quality video $\mathbf{X}_{LR}$, model F_θ, VAE encoder $\mathcal{E}$, decoder $\mathcal{D}$, step number S
Output: High quality video $\mathbf{X}_{HR}$

$\mathbf{Z}_{LR} \leftarrow DEG_{latent}(\mathcal{E}(\mathbf{X}_{LR}))$;
$\Delta_t \leftarrow 1/S$;
$Z \leftarrow \mathbf{Z}_{LR}$;
$t \leftarrow 0$;
for $step = 0$ **to** $S - 1$ **do**
  $\Delta_z \leftarrow F_\theta(Z, t) \cdot \Delta_t$;
  $Z \leftarrow Z + \Delta_z$;
  $t \leftarrow t + \Delta_t$;
$\mathbf{Z}_{HR} \leftarrow Z$;
$\mathbf{X}_{HR} \leftarrow \mathcal{D}(\mathbf{Z}_{HR})$;
return $\mathbf{X}_{HR}$

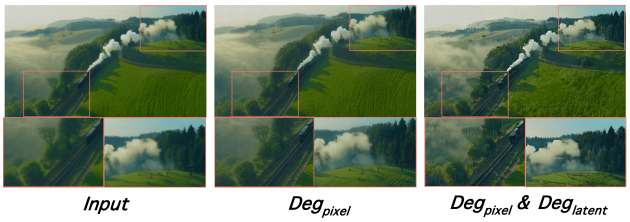

Figure 4: **Visualization of DEG_{pixel} and DEG_{latent} impact on quality enhancement.** The leftmost image shows the *input*, generated by the first-stage model. The term DEG_{pixel} stands for the improved result yielded from the model trained only with pixel-space degradation, which adds high-frequency details to the *input*. Further, DEG_{pixel} & DEG_{latent} refers to the enhanced result with the model trained under both types of degradation, which further improves small structures and fixes artifacts.

from a high-quality video $\mathbf{X}_{HR}$, we apply a sequence of blur and resize operations with randomized strengths in the pixel space (details provided in the *Suppl. Materials*, Section 2), yielding the low-resolution video. This process, denoted as DEG_{pixel}, is outlined in Algorithm 1. Training on this simulated data enables the model to enhance images with high-frequency details, improving overall clarity, as demonstrated in Figure 4.

However, simulating low-resolution data solely through DEG_{pixel} retains strong structure fidelity between low- and high-resolution videos, which limits the model's ability to regenerate accurate structures for small objects at high resolutions—especially when artifacts are present in the first-stage output. This limitation often manifests when there are

poor structural representations for small objects, such as blurry tree branches in Figure 4 or distorted face and finger features in Figure 5. To address this issue, we introduce latent degradation, DEG_{latent}, which perturbs the latent representation with Gaussian noise. This approach allows the model to diverge from the input and generate more reasonable structures for small objects. As shown in Figure 4, compared to DEG_{pixel}, the combination of DEG_{latent} enables the model to produce sharper and more detailed tree branches and tiny background objects, significantly enhancing visual quality.

The overall simulation process during training can be described as follows: First, pixel-space degradation is applied to the high-quality video, yielding a degraded version. This is then encoded into the latent space, represented as: $Z = \mathcal{E}\left(DEG_{pixel}\left(\mathbf{X}_{HR}\right)\right)$. Next, the latent representation is blended with Gaussian noise $n \sim N(0,1)$ to simulate low-quality, defined as:

$$Z_{LR} = DEG_{latent}(Z) = \alpha_{step} \cdot Z + \beta_{step} \cdot n\,. \qquad (1)$$

The equation satisfies $\alpha_{\text{step}}^2 + \beta_{\text{step}}^2 = 1$. The parameter $step$ determines the strength of noise augmentation. The range of $step$ is 0–1000, where higher values indicate greater degradation. To ensure the model can perceive the noise strength, we introduce a noise strength embedding, which is added to the time embedding. At the inference stage, only DEG_{latent} is applied to the first-stage output. However, setting this degradation strength requires careful design in our full attention framework, as discussed in the subsequent paragraph.

Latent degradation strength search. Although the Stage II full-attention architecture enhances spatio-temporal coherence (see Figure 6), it may exhibit a "cheating" behavior: under weak latent degradation (*e.g.*, noise steps < 400), the model simply copies details across frames, leaving artifacts uncorrected. Conversely, overly aggressive degradation (*e.g.*, noise steps > 800) overwhelms the generator, degrading overall fidelity. Hence, precise tuning of the latent degradation magnitude (DEG_{latent}) is essential to erase undesirable artifacts while retaining structures produced in Stage I, which is largely unexplored. Traditional super-resolution approaches (Chan et al. 2022; Yang et al. 2024a) do not target model-introduced artifacts (Figure 5), while recent diffusion-based video enhancement techniques without global attention (Wang et al. 2023; Zhou et al. 2024; He et al. 2024) avoid this cheating at the expense of temporal consistency (Figure 6). To identify an optimal DEG_{latent} range, we begin training with a broad noise-step interval (600–900), then refine it to 650–750 based on empirical performance (Table 4).

4 Experiments

4.1 Data Collection and Implementation Details

Our training dataset consists of 2 million 1080p videos and 1.5 million high-resolution images. The image resolution is approximately 2048×2048. Each video and image is annotated with captions. We also curate a high-quality subset of 50,000 videos for human preference alignment. These videos are manually selected based on aesthetic quality, visual richness, and motion diversity. Details on data sources and processing are in the extended version.

For the first-stage model, only video data is used during training. All videos are resized to a resolution of 270p. The second-stage model is trained progressively from low to high resolution. This approach benefits from the resolution scalability introduced by 3D RoPE (Su et al. 2024; Yang et al. 2024b) in such a video enhancement task, as detailed in Section 5. The training starts with large-scale pretraining on low-resolution images and videos. The model is then adapted to high-resolution data. Finally, a quality-tuning stage is performed on the curated subset of 50,000 videos. For latent degradation strength search, we use a wide range (600–900) during pretraining and early high-resolution training, then narrow it to 650–750 according to the evaluation results. Additional details about training (e.g., iterations for each stage, optimizer, degradation strength) and inference (e.g., classifier guidance, degradation strength) are provided in the extended version.

Method	*Total*	*Quality*	*Semantic*	*Subj. Cons.*	*Motion Smooth.*	*Aesth. Qual.*	*Img. Qual.*	*Obj. Class*	*Overall Cons.*
Wan2.1†	84.70	85.64	80.95	95.92	96.92	61.53	67.28	94.24	27.44
HunyuanVideo†	83.24	85.09	75.82	97.37	98.99	60.36	67.56	86.10	26.44
Vchitect(VEnhancer)	82.24	83.54	77.06	96.83	98.98	60.41	65.35	86.61	27.57
CogVideoX-1.5	82.17	82.78	79.76	96.87	98.31	62.79	65.02	87.47	27.30
CogVideoX-5B	81.61	82.75	77.04	96.23	96.92	61.98	62.90	85.23	27.59
CogVideoX-2B	80.91	82.18	75.83	96.78	99.02	60.82	61.68	83.37	26.66
Mochi-1	80.13	82.64	70.08	96.99	99.02	56.94	60.64	86.51	25.15
LTX-Video	80.00	82.30	70.79	96.56	98.96	59.81	60.28	83.45	25.19
OpenSora-1.2	79.76	81.35	73.39	96.75	98.50	56.85	63.34	82.22	26.85
OpenSoraPlan-V1.1	78.00	80.91	66.38	95.73	98.28	56.85	62.28	76.30	26.52
FlashVideo-8fps	82.80	82.99	82.03	96.91	96.84	62.55	66.96	90.02	27.65
FlashVideo-24fps	83.29	83.72	81.60	97.14	98.83	62.41	66.12	88.45	27.60

Table 1: **Comparison with state-of-the-art open-source models on VBench-Long benchmark (Huang et al. 2024).** This includes the Wan2.1 (Team 2025), HunyuanVideo (Kong et al. 2024), Vchitect-2.0 (Team 2024), VEnhancer (He et al. 2024), CogVideoX (Yang et al. 2024b), Mochi-1, LTX-Video (HaCohen et al. 2024), OpenSora (Zheng et al. 2024) and OpenSoraPlan (Lin et al. 2024). FlashVideo employs a cascade paradigm to deliver top-tier semantic fidelity and quality. † Concurrent work. More detailed results can be found in our extended version.

4.2 Quantitative Results

We first evaluate our model on the VBench-Long (Huang et al. 2024) benchmark using its long prompt, reporting the total score and selected representative sub-scores in Table 1. For video enhancement method comparison and further ablation study, we construct a curated test set of 100 text prompts (in the extended version) with detailed descriptions and generate the corresponding low-resolution 6-

Figure 5: Visual comparison with various video enhancement methods.

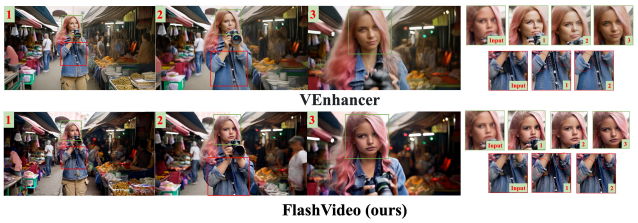

Figure 6: **Long-range detail consistency.** VEhancer (He et al. 2024) distorts identity under large scale changes, while FlashVideo preserves consistent face and texture.

second 49-frame videos using Stage I, incorporating diverse visual elements such as characters, animals, fabrics, and landscapes. We refer to this test set as Texture100. We utilize widely recognized image quality assessment metrics, including MUSIQ (↑) (Ke et al. 2021), MANIQA (↑) (Yang et al. 2022), CLIPIQA (↑) (Wang, Chan, and Loy 2023), and NIQE (↓) (Mittal, Soundararajan, and Bovik 2012), along with the video metric DOVER (Wu et al. 2023), to assess the perception of distortions (Tech ↑) and content preference and recommendation (Aesth ↑).

VBench-Long benchmark. Noting that VBench metrics tend to favor higher frame rates, we apply a real-time video frame interpolation method (Huang et al. 2022) to upscale the frame rate from 8 fps to 24 fps. This interpolation incurs negligible post-processing time (within 4 seconds), ensuring fair comparisons with high-frame-rate methods. (Discussion on VBench's frame rate preference is provided in the extended version.)

As shown in Table 1, both our 8fps and 24fps models achieve high semantic scores (82.03 and 81.60, respectively). However, relying solely on the first-stage model results in aesthetic and imaging quality scores below top-tier methods, with 60.74 and 61.87 for 270p. After applying the second stage, both quality scores improve significantly, reaching state-of-the-art levels of approximately 62.55 and 66.96, respectively, as reported in Table 1. These results validate our approach of initially reducing the resolution in Stage I to ensure high prompt fidelity at a lower computational cost, followed by quality enhancement in Stage II. On the other hand, our entire functional evaluation only takes about 2 minutes, significantly outperforming other methods in terms of efficiency. For example, a concurrent work, Hunyuan Video (Kong et al. 2024), which achieves a total score of 83.24 using a larger 13B single-stage model, requires 1742 seconds for function evaluation to generate 720p (720×1280) results.

4.3 Video Enhancement Comparison

To comprehensively evaluate the effectiveness of our tailored Stage II, we compare it against several state-of-the-art video enhancement methods, including VEnhancer (He et al. 2024), Upscale-a-Video (Zhou et al. 2024), and RealBasicVSR (Chan et al. 2022). As shown in Table 2, we summarize the key attributes for designing Stage II: maintaining high fidelity for a reliable preview, correcting artifacts from Stage I, and ensuring detail consistency across both adjacent frames (avoid flickering) and long-range frames.

Table 2 demonstrates that FlashVideo not only outperforms existing approaches across key metrics but also achieves substantially higher efficiency. Although RealBasicVSR obtains competitive scores on certain metrics, its outputs tend to be overly smooth, underscoring a disparity between quantitative metrics and human perceptual quality. Therefore, we advise using numerical evaluations as a complementary guide, with primary emphasis placed on qualitative inspection. In Figure 5, FlashVideo corrects facial and finger artifacts (a) and faithfully restores high-frequency details (b), while VEnhancer distorts key elements, such as "standing water" patterns and the overall dim tone in (a), and suffers from high computational cost. By contrast, FlashVideo runs nearly seven times faster than VEnhancer. Besides, Upscale-a-Video and RealBasicVSR fail to correct these artifacts and instead yield overly smoothed frames, and Upscale-a-Video exhibits noticeable detail flickering between adjacent frames.

Regarding long-range consistency, Figure 6 shows that FlashVideo's 3D full-attention backbone maintains both identity and detail consistency under large motions and scale variations far better than VEnhancer, which separates spatial and temporal processing. VEnhancer struggles to preserve facial identity across the key frames and introduces inconsistencies in fine details.

5 Ablation

Different Ways to Achieve Stage I At 270p resolution, our Stage I model with LoRA outperforms full fine-tuning across all evaluation metrics—MUSIQ (23.93 vs. 20.53), CLIPIQA (0.286 vs. 0.273), Tech (8.57 vs. 8.53), Aesthetics (97.87 vs. 97.64), and VBench semantic score (82.03 vs. 78.54)—demonstrating superior artifact reduction and improved preservation of both visual quality and semantic alignment.

Position Embedding in Stage II To improve training efficiency, we train Stage II at low resolution and fine-tune at higher resolutions. Comparing RoPE (Su et al. 2024) with default absolute position embeddings (Vaswani 2017) in the 2B DiT (Yang et al. 2024b), RoPE shows better detail and fewer artifacts when evaluated up to 1440×2560(Figure 3). Extrapolating RoPE-trained models from 1080p to 2K further improves all metrics—MUSIQ↑ (62.40 vs. 58.69),

		Frame Quality				Video Quality		Attributes			
	#NFE / Time	MUSIQ(↑)	MANIQA(↑)	CLIPIQA(↑)	NIQE(↓)	Tech(↑)	Aesth(↑)	First Stage as Preview	Artifacts Correction	Long-Range Detail Consistency	Avoids Detail Flickering
RealbasicVSR	1 / 71.5s	<u>54.26</u>	0.272	<u>0.418</u>	<u>5.281</u>	10.71	**99.42**	✓	×	**NA**	✓
Upscale-A-Video	30 / 376.6s	23.67	0.201	0.285	12.02	7.690	97.61	✓	×	**NA**	×
VEnhancer	30 / 549.2s	51.69	<u>0.280</u>	0.385	5.330	<u>11.63</u>	98.39	×	✓	×	✓
FlashVideo (Ours)	4 / 72.2s	**58.69**	**0.296**	**0.439**	**4.501**	**11.86**	<u>98.92</u>	✓	✓	✓	✓

Table 2: **Comparison with video enhancement methods.** Best results are **bold**, second-best <u>underlined</u>. We highlight key design attributes for Stage II; only our tailored method meets all criteria. Visual examples are available on the project page. "NA" indicates limited new details, preventing evaluation of long-range consistency.

MANIQA↑ (0.354 vs. 0.296), CLIPIQA↑ (0.497 vs. 0.439), NIQE↓ (4.463 vs. 4.501), Tech↑ (12.25 vs. 11.86), and Aesth↑ (99.20 vs. 98.92)—demonstrating RoPE's effectiveness for resolution extrapolation.

Low-Quality Video Simulation in Stage II As shown in Figure 4, latent and pixel degradation are key for simulating low-quality videos. We now provide a detailed quantitative evaluation. For computational efficiency, we conduct the experiment using 5-frame 1080p video inputs. We train two models for 10,000 iterations: one with only pixel degradation applied, and the other with both pixel and latent degradation. As shown in Table 3, the baseline represents the results from Stage I. When the Stage II model is applied with pixel degradation (DEG_{pixel}), the first-stage output is significantly improved, with high-frequency textures being added and overall visual quality boosted. Furthermore, incorporating latent degradation (DEG_{latent}) leads to even further enhancement, producing clearer and more realistic structures for small objects and fixing the artifacts.

Deg		Frame Quality				Video Quality	
D_p	D_l	MUSIQ(↑)	MANIQA(↑)	CLIPIQA(↑)	NIQE(↓)	Tech(↑)	Aesth(↑)
		23.61	0.200	0.286	12.02	6.43	97.32
✓		49.12	0.253	0.364	4.95	7.12	99.02
✓	✓	**55.45**	**0.273**	**0.409**	**4.69**	**9.09**	**98.96**

Table 3: Comparison of frame quality and video quality when applying different degradations. Best results are in **bold**.

Latent Degradation Strength. We find that latent degradations below 500 noise steps lead the model to "cheat" and fail to correct first-stage artifacts. To identify an effective strength, we conduct a noise-step search over the 600–900 interval (upper half of Table 4) and observe that the 650–750 window consistently delivers the best artifact removal and visual content preservation. Accordingly, we adopt the 650–750 noise-step range for our final training regime (results shown in the lower half of Table 4).

Number of Function Evaluations. As shown in Table 5, both frame and video quality improve significantly as NFE increases from 1 to 4. Beyond 4, further increasing NFE offers minimal gains. Therefore, we recommend setting NFE between 4 and 6 in practice.

		Frame Quality				Video Quality	
Train	Inf	MUSIQ(↑)	MANIQA(↑)	CLIPIQA(↑)	NIQE(↓)	Tech(↑)	Aesth(↑)
600-900	600	53.62	0.269	0.403	4.911	11.85	99.03
	650	53.98	0.269	0.399	4.832	11.77	99.06
	700	53.82	0.274	0.399	4.763	11.93	99.02
	750	54.06	0.279	0.400	4.785	11.96	98.92
	800	53.50	0.276	0.403	4.663	11.72	98.91
	850	51.39	0.279	0.391	4.787	11.26	98.72
650-750	650	58.49	0.294	0.431	4.583	11.96	98.84
	700	57.80	0.290	0.418	4.531	12.01	98.78
	750	57.62	0.294	0.422	4.437	12.10	98.72

Table 4: Results under different latent degradation strengths. Top: strength search phase. Bottom: final training with selected range.

	Frame Quality				Video Quality	
NFE	MUSIQ(↑)	MANIQA(↑)	CLIPIQA(↑)	NIQE(↓)	Tech(↑)	Aesth(↑)
1	48.60	0.253	0.307	5.148	8.643	98.03
3	57.59	0.290	0.418	4.543	11.39	98.62
4	58.69	0.296	0.439	4.501	11.86	98.92
6	59.17	0.295	0.440	4.521	12.48	99.05
7	59.48	0.298	0.445	4.578	12.20	99.01
8	59.64	0.298	0.451	4.554	12.05	99.16

Table 5: Results under different numbers of function evaluations (NFEs).

6 Conclusions

We present FlashVideo, a two-stage framework that optimizes prompt fidelity and visual quality separately, enabling efficient use of model capacity and NFEs across resolutions. Stage I prioritizes fidelity at low resolution with large parameters and sufficient NFEs, while Stage II refines high-resolution details with fewer NFEs and small parameters. We also design degradation strategies during the second-stage training to ensure a seamless connection between the two independently trained stages. Experiments show the effectiveness of this approach. Moreover, FlashVideo delivers preliminary results at a very low cost, enabling users to decide whether to proceed to the enhancement stage.

Acknowledgments

This paper is partially supported by the National Key R&D Program of China No.2022ZD0161000 and the General Research Fund of Hong Kong No.17208825 and 17209324.

References

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

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

Chan, K. C.; Zhou, S.; Xu, X.; and Loy, C. C. 2022. Investigating Tradeoffs in Real-World Video Super-Resolution. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, H.; Xia, M.; He, Y.; Zhang, Y.; Cun, X.; Yang, S.; Xing, J.; Liu, Y.; Chen, Q.; Wang, X.; Weng, C.; and Shan, Y. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. arXiv:2310.19512.

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. arXiv:2401.09047.

Chen, J.; Ge, C.; Xie, E.; Wu, Y.; Yao, L.; Ren, X.; Wang, Z.; Luo, P.; Lu, H.; and Li, Z. 2025. PIXART-Sigma: Weak-to-Strong Training of Diffusion Transformer for 4K Text-to-Image Generation. In *European Conference on Computer Vision*, 74–91. Springer.

Ding, Z.; Jin, C.; Liu, D.; Zheng, H.; Singh, K. K.; Zhang, Q.; Kang, Y.; Lin, Z.; and Liu, Y. 2024. DOLLAR: Few-Step Video Generation via Distillation and Latent Reward Optimization. *arXiv preprint arXiv:2412.15689*.

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 *Forty-first International Conference on Machine Learning*.

Fischer, J. S.; Gui, M.; Ma, P.; Stracke, N.; Baumann, S. A.; Hu, V. T.; and Ommer, B. 2023. Boosting Latent Diffusion with Flow Matching. arXiv:2312.07360.

Gu, J.; Zhai, S.; Zhang, Y.; Susskind, J. M.; and Jaitly, N. 2023. Matryoshka diffusion models. In *The Twelfth International Conference on Learning Representations*.

HaCohen, Y.; Chiprut, N.; Brazowski, B.; Shalem, D.; Moshe, D.; Richardson, E.; Levin, E.; Shiran, G.; Zabari, N.; Gordon, O.; Panet, P.; Weissbuch, S.; Kulikov, V.; Bitterman, Y.; Melumian, Z.; and Bibi, O. 2024. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103*.

He, J.; Xue, T.; Liu, D.; Lin, X.; Gao, P.; Lin, D.; Qiao, Y.; Ouyang, W.; and Liu, Z. 2024. VEnhancer: Generative Space-Time Enhancement for Video Generation. *arXiv preprint arXiv:2407.07667*.

He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent Video Diffusion Models for High-Fidelity Long Video Generation.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Ho, J.; Saharia, C.; Chan, W.; Fleet, D. J.; Norouzi, M.; and Salimans, T. 2022a. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47): 1–33.

Ho, J.; Salimans, T.; Gritsenko, A.; Chan, W.; Norouzi, M.; and Fleet, D. J. 2022b. Video diffusion models. *Advances in Neural Information Processing Systems*, 35: 8633–8646.

Hong, W.; Ding, M.; Zheng, W.; Liu, X.; and Tang, J. 2022. CogVideo: Large-scale Pretraining for Text-to-Video Generation via Transformers. *arXiv preprint arXiv:2205.15868*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

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

Huang, Z.; Zhang, T.; Heng, W.; Shi, B.; and Zhou, S. 2022. Real-Time Intermediate Flow Estimation for Video Frame Interpolation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jin, Y.; Sun, Z.; Li, N.; Xu, K.; Jiang, H.; Zhuang, N.; Huang, Q.; Song, Y.; Mu, Y.; and Lin, Z. 2024. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5148–5157.

Kondratyuk, D.; Yu, L.; Gu, X.; Lezama, J.; Huang, J.; Schindler, G.; Hornung, R.; Birodkar, V.; Yan, J.; Chiu, M.-C.; Somandepalli, K.; Akbari, H.; Alon, Y.; Cheng, Y.; Dillon, J. V.; Gupta, A.; Hahn, M.; Hauth, A.; Hendon, D.; Martinez, A.; Minnen, D.; Sirotenko, M.; Sohn, K.; Yang, X.; Adam, H.; Yang, M.-H.; Essa, I.; Wang, H.; Ross, D. A.; Seybold, B.; and Jiang, L. 2024. VideoPoet: A Large Language Model for Zero-Shot Video Generation. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 25105–25124. PMLR.

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*.

Lin, B.; Ge, Y.; Cheng, X.; Li, Z.; Zhu, B.; Wang, S.; He, X.; Ye, Y.; Yuan, S.; Chen, L.; et al. 2024. Open-Sora Plan: Open-Source Large Video Generation Model. *arXiv preprint arXiv:2412.00131*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

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*.

Mittal, A.; Soundararajan, R.; and Bovik, A. C. 2012. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3): 209–212.

Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. *arXiv preprint arXiv:2212.09748*.

Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; and Courville, A. 2018. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Pernias, P.; Rampas, D.; Richter, M. L.; Pal, C. J.; and Aubreville, M. 2023. Würstchen: An efficient architecture for large-scale text-to-image diffusion models. *arXiv preprint arXiv:2306.00637*.

Ren, J.; Li, W.; Chen, H.; Pei, R.; Shao, B.; Guo, Y.; Peng, L.; Song, F.; and Zhu, L. 2024. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, 234–241. Springer.

Saharia, C.; Ho, J.; Chan, W.; Salimans, T.; Fleet, D. J.; and Norouzi, M. 2022. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 4713–4726.

Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, 2256–2265. PMLR.

Su, J.; Ahmed, M.; Lu, Y.; Pan, S.; Bo, W.; and Liu, Y. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568: 127063.

Team, V. 2024. Vchitect-2.0: Parallel Transformer for Scaling Up Video Diffusion Models. https://github.com/Vchitect/Vchitect-2.0.

Team, W. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.

Teng, J.; Zheng, W.; Ding, M.; Hong, W.; Wangni, J.; Yang, Z.; and Tang, J. 2023. Relay diffusion: Unifying diffusion process across resolutions for image synthesis. *arXiv preprint arXiv:2309.03350*.

Vaswani, A. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Wang, J.; Chan, K. C.; and Loy, C. C. 2023. Exploring CLIP for Assessing the Look and Feel of Images. In *AAAI*.

Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2023. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*.

Wu, H.; Zhang, E.; Liao, L.; Chen, C.; Hou, J. H.; Wang, A.; Sun, W. S.; Yan, Q.; and Lin, W. 2023. Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives. In *International Conference on Computer Vision (ICCV)*.

Xing, Z.; Dai, Q.; Hu, H.; Wu, Z.; and Jiang, Y.-G. 2024. Simda: Simple diffusion adapter for efficient video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7827–7839.

Yan, W.; Zhang, Y.; Abbeel, P.; and Srinivas, A. 2021. VideoGPT: Video Generation using VQ-VAE and Transformers. arXiv:2104.10157.

Yang, S.; Wu, T.; Shi, S.; Lao, S.; Gong, Y.; Cao, M.; Wang, J.; and Yang, Y. 2022. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1191–1200.

Yang, X.; He, C.; Ma, J.; and Zhang, L. 2024a. Motion-Guided Latent Diffusion for Temporally Consistent Real-world Video Super-resolution.

Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024b. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. *arXiv preprint arXiv:2408.06072*.

Yu, F.; Gu, J.; Li, Z.; Hu, J.; Kong, X.; Wang, X.; He, J.; Qiao, Y.; and Dong, C. 2024. Scaling up to excellence: Practicing model scaling for photo-realistic image restoration in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 25669–25680.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

Zhang, S.; Wang, J.; Zhang, Y.; Zhao, K.; Yuan, H.; Qin, Z.; Wang, X.; Zhao, D.; and Zhou, J. 2023. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*.

Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-Sora: Democratizing Efficient Video Production for All.

Zhou, S.; Yang, P.; Wang, J.; Luo, Y.; and Loy, C. C. 2024. Upscale-A-Video: Temporal-Consistent Diffusion Model for Real-World Video Super-Resolution. In *CVPR*.