


SURF: Signature-Retained Fast Video Generation

Kaixin Ding^{1†}, Xi Chen¹, Sihui Ji¹, Yuan Gao², Liang Hou²,
Xin Tao², Hengshuang Zhao¹ 

¹The University of Hong Kong, ²Kling Team, KuaishouTechnology

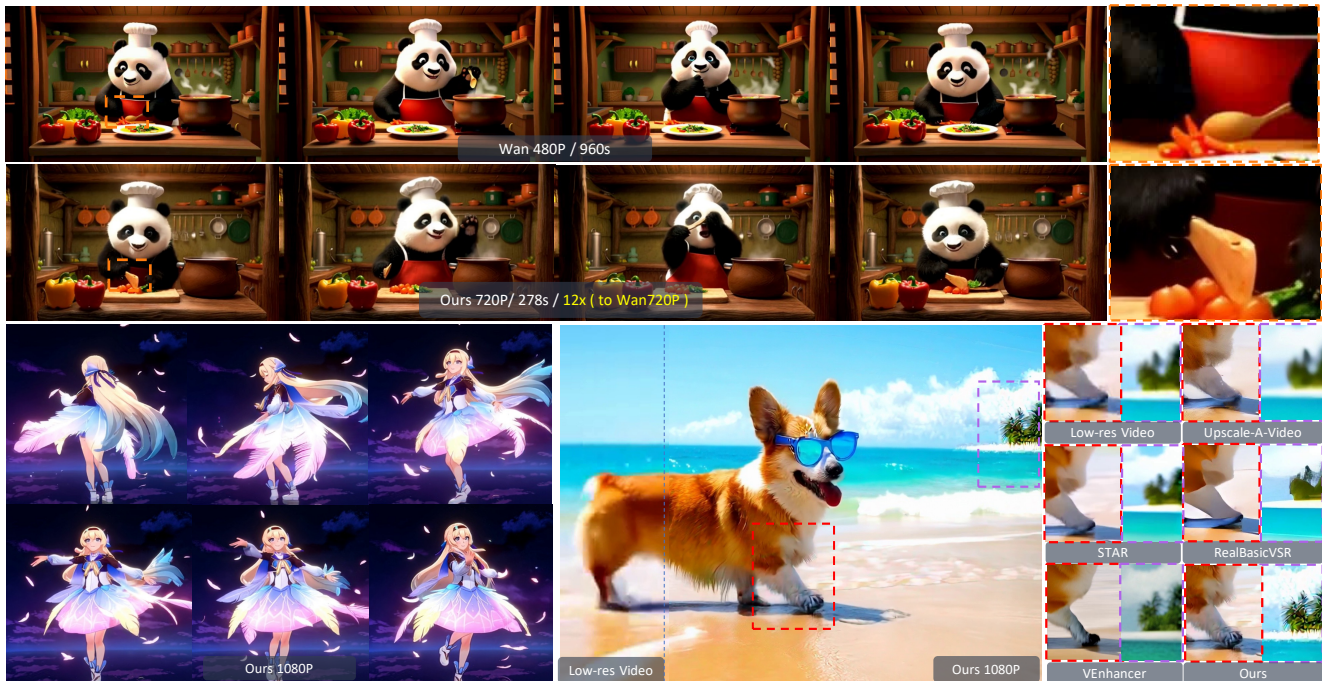


Figure 1. SURF generates high-resolution and high-quality videos efficiently. It acts as a simple plug-in and retains the signatures (*i.e.*, layout, semantic, motion, *etc.*) of the baseline model with more than 12 \times speed-up. More results are available on [Project Page](#).

Abstract

The demand for high-resolution video generation is growing rapidly. However, the generation resolution is severely constrained by slow inference speeds. For instance, Wan 2.1 requires over 50 minutes to generate a single 720p video. While previous works explore accelerating video generation from various aspects, most of them compromise the distinctive signatures (*e.g.*, layout, semantic, motion) of the original model. In this work, we propose SURF, an efficient framework for generating high-resolution videos, while maximally keeping the signatures. Specifically, SURF divides video generation into two stages: First, we leverage the pretrained model to infer at optimal resolution and downsample latent to generate low-resolution previews in fast speed; then we design a Refiner to upscale the preview. In the preview stage, we identify that directly

inferring a model (trained with higher resolution) on lower resolution causes severe losses in signatures. So we introduce noise reshifting, a training-free technique that mitigates this issue by conducting initial denoising steps on the original resolution and switching to low resolution in later steps. In the refine stage, we establish a mapping relationship between the preview and the high-resolution target, which significantly reduces the denoising steps. We further integrate shifting windows and carefully design the training paradigm to get a powerful and efficient Refiner. In this way, SURF enables generating high-resolution videos efficiently while maximally closer to the signatures of the given pretrained model. SURF is conceptually simple and could serve as a plug-in that is compatible with various base model and acceleration methods. For example, it achieves 12.5 \times speedup for generating 5-second, 16fps, 720p Wan 2.1 videos and 8.7 \times speedup for generating 5-second, 24fps, 720p HunyuanVideo.

[†] Work done at Kling Team, Kuaishou Technology.


 Corresponding Author



Figure 2. **Demonstration for signature retention.** Distilled model loses the signatures of the base model, which could lead to misaligned body parts and weak semantic consistency. Our method maintains the layout, semantics, and motion of base model with huge speed up.

1. Introduction

Video generation [19, 34, 45] has witnessed remarkable advancements in recent years, with sophisticated models continually pushing the boundaries of quality and capability. However, the computational cost remains a critical bottleneck, significantly hindering further advances towards higher resolution and richer details. For example, it costs roughly 50 min to generate a 5s 720p video for the current SoTA video generation models [18, 34] and still 6 min for distillation model [47].

Facing this challenge, existing researches explored various strategies to make the generation process more efficient. Specifically, methods [22, 30] leverage step distillation to reduce the total denoising steps. Studies like [10, 48] primarily focus on attention sparsity to improve efficiency. Furthermore, cascade multi-scale generation techniques [6, 25] have been proposed to enhance the efficiency of high-resolution image generation.

Although these methods could improve the generation efficiency, most of them inevitably compromise the intrinsic signatures of the original model as shown in Fig. 2. The model’s characteristics are reflected in its preferences for aesthetic style, semantically aligned layout and motion dynamics, *etc.* Preserving these characteristics are significant when accelerating a given model, as they act as a prior of the model and could directly reflect the generation quality.

To achieve this goal, we propose an efficient framework for high-resolution video generation, called SURF. This framework divides the whole process into two stages based on their natural features. In **Preview stage**, we leverage a powerful pretrained model to infer at optimal resolution and latent downsample to generate low-resolution previews. The pre-k steps focus on high noise period which determines the main content of the entire video and the post-k steps uses the low resolution latent to gain speed. We analyze that each pretrained model has its own optimal resolution (usually the training resolution). While distilling

the model could increase speed in a large, it often lacks access to the large-scale pretraining data and thus fails to preserve crucial signatures. Inspired by the observation that the early denoising steps determine the overall content, and the later steps refine the details, we introduce noise reshift method to address this problem. Specifically, we start from the pretrained model’s optimal resolution in the early denoising steps, then switch to a lower resolution for the remaining steps. In this way, the low-resolution preview could be efficiently generated and preserve the signatures of the pretrained model. In **Refine stage**, we train a powerful and efficient Refiner through establishing a mapping relationship between the low-resolution preview and the high-resolution target. It significantly reduces the NFE (number-of-evaluations) to 10 while enriching details and correcting unreasonable artifacts. In addition, we also integrate the shift window method into the upscaling model to further reduce the computation.

As shown in Fig. 1, our method generates high-detailed and high-quality videos with signatures (layout, semantic, motion, *etc.*) closer to the base model (Wan 2.1) and achieves a huge speedup. Besides, our SURF paradigm allows users to generate multiple previews efficiently at the same time and pick the satisfied content for further refinement to arbitrary size. Experiments show that SURF achieves 12.5× speedup compared with Wan 2.1 14B for generating 5-second, 16fps, 720p videos and 8.7× speedup compared with HunyuanVideo for generating 5-second, 24fps, 720p videos.

2. Related Work

High-resolution image and video generation. Existing approaches can be broadly categorized into training-based and training-free paradigms. Training-based methods combine architectural innovations and model fine-tuning strategies. Turbo2K [27] accelerates 2K video synthesis by leveraging a compressed latent space and knowledge distillation

within a hierarchical two-stage framework, ensuring structural coherence. UltraPixel [26] generates 4K-resolution images using cascade diffusion models that feature shared parameters and scale-aware layers, minimizing additional parameters for high-resolution outputs. PixArt- Σ [5] and ResAdapter [7] enhance base models through fine-tuning but remain resolution-constrained. Other approaches such as ResMaster [29] and HiPrompt [21] introduce multi-modal prompting mechanisms at the expense of computational efficiency. Training-free methods adapt inference strategies or architectures without retraining. MultiDiffusion [2] and its variants (*e.g.*, SyncDiffusion [20], Demofusion [12]) employ sliding-window denoising but suffer from repetition or computational redundancy. ScaleCrafter [14], recitifiedHr [44], FouriScale [15], and HiDiffusion [50] modify network structures but risk suboptimal performance. SVG [39] explores sparsity within the attention module but achieves only limited acceleration. Jenga [52] employs dynamic token carving to reduce redundancy in attention computation, but its generated content deviates from the original model.

Efficient video generation. The main challenge of efficient video generation lies in the quadratic complexity of attention mechanisms, especially under high-resolution settings. Several works [3, 8, 17, 36, 41, 46] transform attention into linear operations to reduce computational cost, while others adopt hybrid strategies combining local and global attention to focus on informative token pairs [39, 40, 48, 49]. FlashAttention [9] introduces a patch-divided acceleration method, and more recently, SEED-VR [35] proposes variable-sized windows near spatiotemporal boundaries to better support long video sequences. For faster sampling, rectified flow models with straight ODE trajectories [13] have been proposed for text-to-image (T2I) tasks. However, applications in text-to-video (T2V) remain limited due to the added complexity of temporal dynamics. While some methods [11, 51] reduce denoising steps, they are still constrained to long-frame, high-resolution settings and often overlook resolution-domain inconsistency. Tian et al. [33], Yang et al. [43] further explored stage-wise denoising, but their framework is restricted to a single model and only supports two-resolution transfer, without investigating how different step divisions affect generation quality, thereby offering limited acceleration benefits. These observations motivate our design of a resolution-dynamic T2V generation framework.

3. Method

In this work, we propose SURF, a framework that significantly improves the efficiency of pretrained video generation models while maximally preserving their generative signatures.

3.1. Overall Pipeline.

The overall framework of SURF is shown in Fig. 3. Current video acceleration methods often operate at a fixed latent scale. While strategies [39, 42, 48] that exploit the sparsity of video tokens within attention modules can reduce computational costs, their acceleration potential remains limited. Moreover, adopting an aggressive token dropping policy—even retaining specifically selected important tokens—inevitably degrades the learned generative signatures. Therefore, we analyze that the key factors that influence the generation speed are the **resolution** and the **number of denoising steps**.

We design the pipeline under this principle: we infer at optimal resolution and downsample the latent to generate low-resolution previews; then we add more details to the preview with fewer denoising steps at target high-resolution. SURF introduces a dynamic scaling mechanism that allocates a variable number of tokens according to the denoising timestep. Instead of permanently discarding tokens, we resize the latent scale to modulate the token count, thereby ensuring that the global information from the entire token set is preserved.

Specifically, in the preview stage, we retain the full capacity of a powerful pretrained model to establish the global structure with content and accelerate the whole process by dynamic scaling mechanism; In the refine stage, we switch to a lightweight model to both accelerate the process and enrich the details.

3.2. Preview stage

In this stage, we aim to leverage the strong generative capabilities of a given pretrained model for generating a low-resolution preview. We expect the low-resolution preview to preserve the signatures (layout, semantic, motion) of the given pretrained model. A straightforward solution is to let the pretrained model infer on the lower resolution. However, we find that each model has its own optimal resolution; inferring on mismatched resolution causes severe degradation. Inspired by the fact that the overall structure is determined by the initial denoising steps, we introduce a progressively downsampling method.

Noise reshifting. During the denoising phase, we begin with the initial gaussian noise $z_1 \in R^{b \times c \times f \times h \times w}$ and progressively downsample the clean latent representation z_0 with reduced spatial dimensions $h' \times w'$. Here b , c , f and h, w denote the batch size, number of channels, number of frames, and resolution of the model’s optimal generation, respectively. Our strategy is fundamentally structured around a turning point step k along the denoising trajectory and divides into pre- k steps and post- k steps.

Before reaching k (pre- k steps), the latent representation is denoised through an ODE-based flow matching approach,

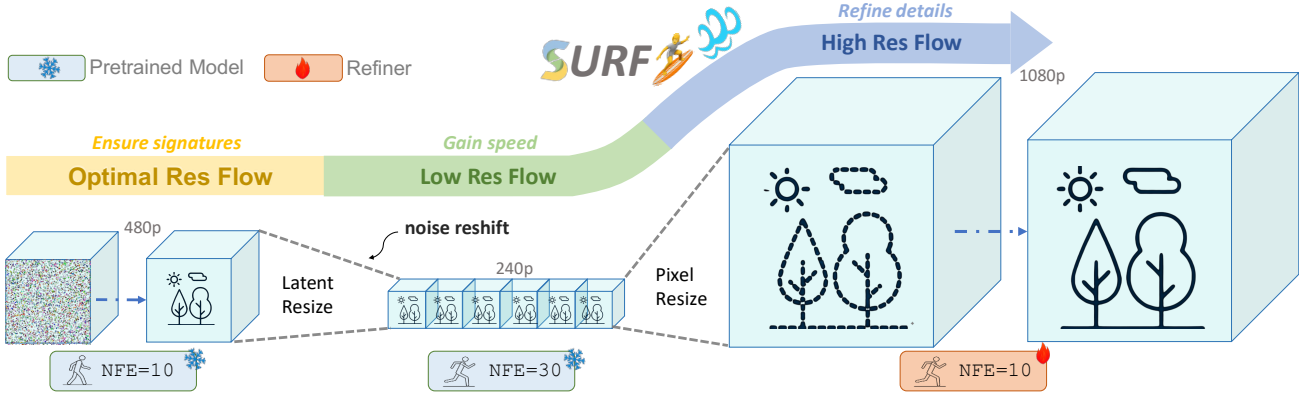


Figure 3. **Overview of the SURF.** Our framework employs a powerful pretrained model (e.g., Wan 2.1) and a lightweight Refiner network. Together, they execute the denoising process through an OptimRes-LowRes-HighRes flow (SURF), producing outputs with signature (e.g., layout, semantic, motion) closely match those of the base model (Wan 2.1), while achieving huge speed up.

described by the following iterative update:

$$\mathbf{z}_0 = \mathbf{z}_1 + \int_1^0 \mathbf{u}_\theta(\mathbf{z}_t, t) dt \quad (1)$$

where $\mathbf{u}_\theta(\mathbf{z}, i)$ represents the model-predicted direction function. Upon reaching step k , we estimate the clean latent representation as $\hat{\mathbf{z}}_0 = \mathbf{z}_k - \sigma_k \cdot \mathbf{u}_\theta(\mathbf{z}_k, k)$, where σ_k denotes the noise standard deviation at step k . Subsequently, we apply a spatial downscaling operation to the estimated latent, $\hat{\mathbf{z}}_0^\downarrow = \text{Downscale}(\hat{\mathbf{z}}_0)$. where the simple linear downscale operation is implemented in the latent space. This choice is crucial as it effectively reduces spatial resolution while vigilantly preserving essential spatiotemporal coherence. Afterwards, we reinject noise which is shifted to timestep k , into the reduced-resolution latent space, allowing the stochastic denoising process to seamlessly resume. The reshifting process is represented by:

$$\mathbf{z}_{k-1} = \hat{\mathbf{z}}_0^\downarrow + \sigma_k \cdot \tilde{\epsilon}, \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (2)$$

In post- k steps, latents denoises in lower-resolution to gain speed. This multi-scale denoising framework enables the model to first gain global semantics at coarser scales, and then get the preview at finer scales.

3.3. Refine stage

To alleviate the computational overhead, we employ a lightweight model with 1B parameters, which reduces the time cost per step (architectures and training details can be found in *Supplementary Material*).

Flow mapping. We start from the preview and linear upsample it to form blurred low-resolution latents. Then we establish a mapping between low-resolution latents z_{lr} and high-resolution latents z_{hr} by modifying the flow-matching (1), substituting z_1 with z_{lr} and z_0 with z_{hr} . Our light-weight Refiner learns z_{lr} 's directional information, facilitating high-quality video generation in a few denoising steps.

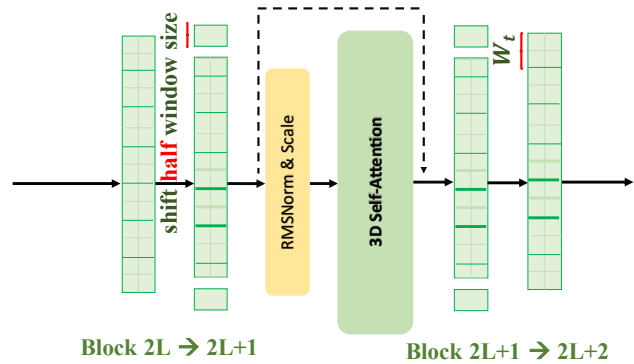


Figure 4. **Shift window** across adjacent blocks.

Training paradigm. We first obtain large amounts of high-quality videos and apply both pixel- and latent-level degradations to simulate low-quality video values. Pixel-level degradation simulates blur, while latent-level degradation prevents the task from becoming a trivial super-resolution problem, encouraging the model to exploit its generative ability. These low-quality latents are then paired with the original high-quality latents to form the training data. However, with large resolution and long frame numbers, it presents a computational challenge and results in considerable time consumption per denoising step.

Model structure. Given a video feature $\mathbf{X} \in \mathbb{R}^{T \times H \times W}$, our transformer design addresses this challenge by balancing efficiency and temporal connectivity. The first transformer block applies *regular window attention* with a $t \times H \times W$ window. Specifically, \mathbf{X} is divided into $(\frac{T}{t} + 1) \times H \times W$ windows, where each window spans t consecutive frames along the temporal dimension while fully covering the spatial dimensions. This windowed partition allows local temporal interactions to be captured before expanding to broader temporal receptive fields in deeper blocks. Our solution further integrates a cyclic shift-

window strategy into the 3D self-attention modules shown in Fig. 4. This strategy, embedded within Transformer blocks, establishes full temporal connectivity through a two-phase cycle across layers. Each consecutive layer pair collaboratively connects all frames while maintaining computational efficiency.

In detail, Transformer block $2L$ applies 3D self-attention to non-overlapping temporal windows of size W_t , adjusting position embeddings for local awareness. The subsequent Transformer block $2L + 1$ applies a temporal shift of $S_t = \frac{W_t}{2}$ to the input feature, then partitions it into windows of size W_t for attention computation. The cyclic-shifted attention mechanism can be expressed as:

$$\begin{aligned} \mathbf{X}^{(2L)} &= \text{Attention3D}(\text{Partition}(\mathbf{X}, W_t)), \\ \mathbf{X}_{\text{shifted}} &= \text{CyclicShift}\left(\mathbf{X}^{(2L)}, \frac{W_t}{2}\right), \\ \mathbf{X}^{(2L+1)} &= \text{Attention3D}(\text{Partition}(\mathbf{X}_{\text{shifted}}, W_t), \text{Mask}). \end{aligned} \quad (3)$$

When we shift the window by half the window size forward, one boundary window (e.g., the first window) contains temporally unrelated halves. Therefore, we apply an attention mask to separate them. We also employ position-frequency embeddings with 3D RoPE [31] within each window to avoid the resolution bias introduced by fixed positional embeddings. This two-block cycle (unshifted/shifted layers) guarantees global temporal connectivity, significantly reducing memory and accelerating attention calculation for large latent tensors.

4. Experiments

4.1. Implementation Details

Settings. The Refiner of SURF is trained on 24 NVIDIA A800 GPUs (80 GB each) with a total batch size of 24. We finetune the transformer using the AdamW optimizer with a learning rate of $5e-5$. We create a synthetic dataset of 100k LR-HR video frame pairs following the methodology in Wang et al. [37]. To optimize training efficiency and stabilize convergence, we employ a progressive training strategy where the model is trained at incrementally increasing frame num. We first train SURF on 21-frame inputs for 1k iterations, and then extend the input length to 81 frames and finetune the model for 4k iterations.

Evaluation and metrics. For speed assessment, we report the DiT forward pass time on NVIDIA A800, as the VAE decoding component remains constant across all configurations. We also report FLOPs to provide an intuitive comparison of computational complexity. For qualitative evaluation, we construct a video dataset of 381 low-resolution videos with prompts from VBench [16], VideoPhy [1] and PhyGenBench [23]; We evaluate each prompt with a fixed seed to ensure reproducibility. Additionally, we conducted a user study to assess human preference rates

between SURF and various efficient generation methods. As for 1080p video generation, most existing acceleration methods are not capable of handling this setting. Therefore, for a fair comparison, we benchmark our approach against super-resolution(SR) methods and adopt several widely used video SR assessment metrics. Specifically, we employ DINO [4] and CLIP [24] to evaluate frame quality via feature similarity across frames; LAION aesthetic predictor [28] to assess artistic and beauty value perceived by humans towards each video frame, and DOVER [38] to measure overall video quality.

4.2. Comprehensive Evaluations

We evaluate our method from both efficiency and quality perspectives. Specifically, We compare it with the Wan 2.1 baseline, which employs the FlowMatch scheduler with 50 NFE (number-of-evaluations), as well as accelerated Wan 2.1 variants utilizing 30% and 50% of the original steps. Additionally, we compare our approach with two commonly adopted acceleration techniques: the sparse attention mechanism used in SVG [39] and the bidirectional video distribution matching distillation (DMD) method (implemented following Team [32]).

Efficiency analysis. The quadratic cost of attention makes token count the key bottleneck for efficiency. SURF addresses this by adapting latent scales across denoising stages. The first stage processes inputs at a resolution equivalent to 480p, the second at 240p, and the third utilizes a smaller model with approximately $5\times$ fewer hidden dimensions and $2.5\times$ fewer attention heads, resulting in negligible FLOPs compared to the original model. As shown in Tab. 1, our method achieves performance comparable to Wan 2.1 at 720p resolution while reducing inference time by $12\times$. Furthermore, for 1080p video generation, our approach achieves a remarkable $43\times$ acceleration compared to a direct application of Wan 2.1 at the same resolution.

In contrast, other methods such as SVG operate on latents of the same scale, achieving acceleration primarily through hardware-efficient tensor layouts, which offer limited reduction in redundant tokens. DMD, on the other hand, focuses on reducing the number of inference steps. SURF distinguishes itself by leveraging the intrinsic properties of the denoising process, adapting latent representations at different scales to different stages of denoising. This strategy achieves substantial acceleration while preserving crucial signatures information. Although the overall runtime of distillation-based methods is comparable to ours, they fail to preserve signatures of original models and introduce quality degradation (e.g., misaligned limbs) in generated videos.

Quality analysis. We evaluated our model’s generative performance from two perspectives: general video metrics and physics-focused assessments (as original Wan 2.1 achieves

Table 1. **Quantitative comparison on Wan 2.1.** We report evaluations of the baseline (row 1), step and attention optimization methods (row 2-5) and SURF (row 6). NFE=50. The highest score is in **bold** and the second highest is underlined. Abbreviations: QS (Quality Score), AQ (Aesthetic Quality), DD (Dynamic Degree), MS (Motion Smoothness), OC (Overall Consistency), SA (Semantic Adherence), PC (Physics Commonsense).

Method	General Scene					Physical Scene		Computation Loads		
	QS \uparrow	AQ \uparrow	DD \uparrow	MS \uparrow	OC \uparrow	SA \uparrow	PC \uparrow	Time \downarrow	Speed \uparrow	PFLOPs \downarrow
Wan 2.1 [34]	<u>83.31</u>	66.9	63.89	97.65	27.08	41.82	45.45	3497 (58min)	1 \times	658.5
30%step [34]	77.92	58.43	56.94	96.95	24.56	18.18	16.36	1049	3.34 \times	197.5
50%step [34]	81.51	63.52	66.67	96.99	25.90	25.45	23.64	1748	2.00 \times	329.2
SVG [39]	83.36	65.6	<u>68.06</u>	97.69	<u>27.32</u>	25.45	20.00	2712	1.29 \times	429.9
DMD [32]	<u>83.31</u>	66.11	52.78	98.96	26.77	<u>34.55</u>	30.91	<u>282</u>	<u>12.40</u> \times	<u>39.5</u>
Ours	83.26	<u>66.86</u>	72.22	<u>97.95</u>	27.38	41.82	<u>38.18</u>	278	12.58 \times	34.3

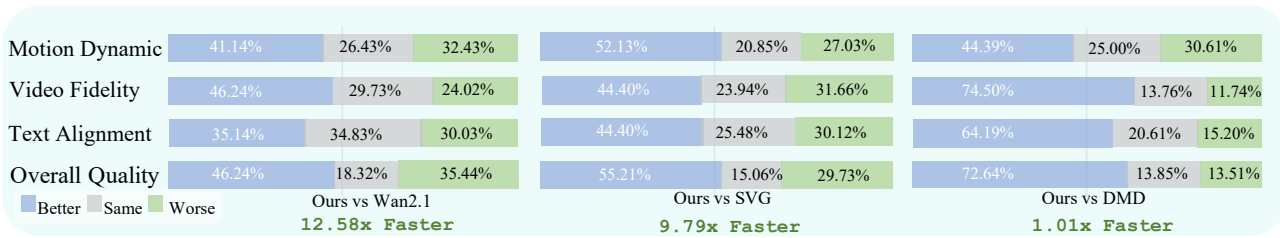


Figure 5. **User study.** We report pair-wise preference rates. SURF achieves comparable quality to Wan 2.1 with huge speed up.

Table 2. **Plug-in integration** with other acceleration methods and different model architectures.

Method	SA \uparrow	PC \uparrow	NFE/Time \downarrow	Speed \uparrow
HunyuanVideo [18]	29.09	27.27	50/3081	1 \times
+SparseAttn [52]	30.91	23.64	50/2775	1.1 \times
+ SURF	43.64	45.45	50/356	8.7 \times
AccVideo [47]	32.73	23.64	5/340	1 \times
+ SURF	36.36	38.18	5/265	1.3 \times

best in physical plausibility so we want to prove SURF has preserved this characteristics).

As illustrated in Tab. 1, our generated video exhibits visual quality comparable to Wan 2.1, while slightly surpassing SVG and outperforming the DMD method. For further validation, we provide qualitative comparisons in Fig. 6. We also conducted a perceptual evaluation using a standard win-rate methodology. We have designed questionnaire with 24 randomly selected videos from the aforementioned test datasets. A total of 37 researchers in the field of video generation were asked to evaluate the results along four dimensions: Motion, Fidelity, Semantics, and Overall Quality. The outcomes, summarized in Fig. 5, demonstrate that our method achieves performance comparable to the original model while surpassing many existing acceleration approaches from a human perspective. From extensive examples, we observe that SVG suffers from limited robustness where main subjects often remain static and adds unreasonable details; In contrast, DMD tends to introduce unnatural color artifacts and produces grainy videos with reduced fidelity.

Moreover, our method can be naturally extended to support 1080p video generation. Since most baseline models do not natively support this resolution, we evaluate 1080p results from the perspective of video super-resolution (SR). Specifically, we randomly select 100 samples from the aforementioned test datasets and conduct comparisons with existing SR approaches, aiming to assess the effectiveness of our method in generating high-resolution videos with enhanced detail and fidelity. The detailed results, as shown in Tab. 3, indicate our method’s strong performance. Notably, although GAN-based RealBasicVSR achieves competitive scores on some metrics, its outputs frequently exhibit excessive smoothing, which does not align with human perceptual preferences; Diffusion-based VEnhancer also demonstrates strong generative capabilities, however, its outputs often undergo significant deviations from the input, contradicting the principle of enhancing visual quality while preserving fidelity.

Compatibility analysis. Furthermore, SURF can function as a plug-in compatible with various diffusion model architectures and acceleration techniques. When combined with sparse attention [52], it achieves a 8.7 \times speedup on HunyuanVideo [18]. Additionally, it can be adapted to step-distillation models AccVideo [47] and achieves a 1.3 \times speedup. Detailed results are presented in Tab. 2.

4.3. Stage-wise Visualization

This section presents visualizations of SURF, providing an intuitive comparison between the preview and the final

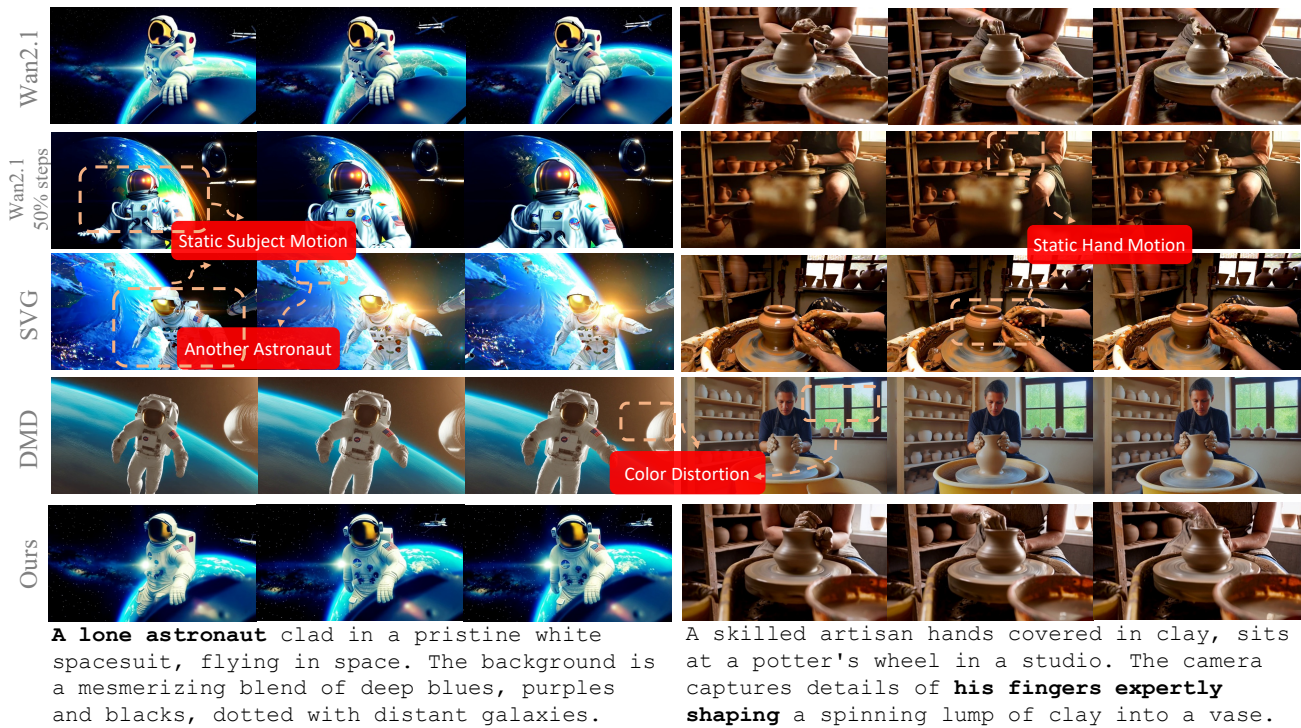


Figure 6. **Qualitative comparisons.** Our method achieves up to 12× speedup while maintaining signatures of base model. Unreasonable contents are marked in orange. Rather than aimless camera panning, SURF generates high fidelity videos with semantically aligned motion.

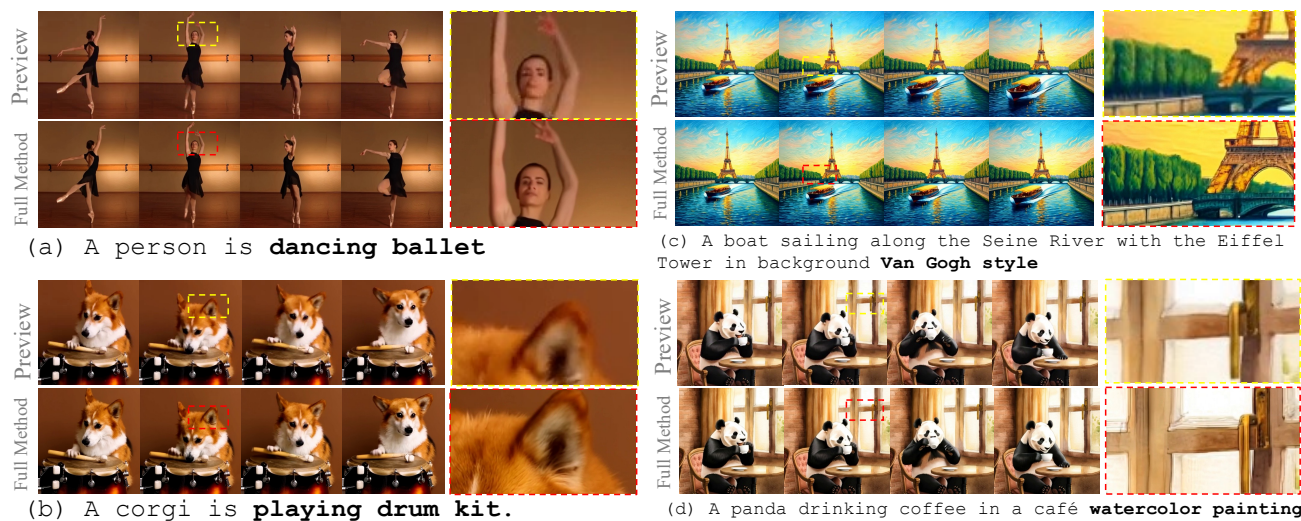


Figure 7. **Generation results for each stage.** We mark regions with artifacts and lacking detail in the preview videos using yellow boxes, while improvements from the Refiner are highlighted in red. Zoom in for a better view.

results in overall layout similarity and refinement capability. As shown in top row of Fig. 7, our design first prioritizes an efficient retention of the inherent signatures of the powerful model to ensure robust structural layout, semantic alignment, and motion dynamics. For instance, the ballet dancer’s movements are demonstrably smooth and natural. Furthermore, our method is adept at style-aware video synthesis, exemplified by results such as a

Van Gogh-style tower and a watercolor panda. Then the well-trained Refiner refines intricate details and mitigates generation artifacts. In second row, SURF further enhances visual fidelity of existed videos, including clearer facial and hand structures in (a), refined fur on the corgi in (b), enhanced texture and structural details in (c), and more semantically consistent structural and color corrections in (d). More examples in Fig. 1 further demonstrate the visual

Table 3. **Quantitative comparison** with video super-resolution method (1080p).

Method	DINO \uparrow	CLIP \uparrow	LAION \uparrow	DOVER \uparrow	NFE/Time \downarrow
RealBasicVSR	93.40	94.83	61.07	<u>80.25</u>	1/162.1s
Upscale-a-Video	93.47	95.71	60.93	71.40	30/2517.7s
VEnhancer	93.55	96.02	<u>63.46</u>	79.78	15/2467.6s
STAR	93.68	96.59	60.81	63.64	14/912.7s
Ours	93.75	<u>96.30</u>	63.50	81.20	10/76.5s

Table 5. **Ablation study** of denoising steps in refine stage. We report SURF in 720p (left block) and 1080p (right block). The recommended setting is in gray.

Step	720p						1080p			
	QS \uparrow	AQ \uparrow	DD \uparrow	MS \uparrow	OC \uparrow	Time \downarrow	DINO \uparrow	CLIP \uparrow	LAION \uparrow	DOVER \uparrow
8	83.24	66.90	72.22	97.94	27.34	244.5	93.70	96.28	63.48	80.52
9	83.17	66.90	70.83	97.95	27.38	247.0	93.75	96.31	63.48	80.89
10	83.26	66.86	72.22	97.95	27.38	249.0	93.75	96.30	63.54	81.20
11	83.22	67.03	70.83	97.95	27.38	251.8	93.71	96.32	63.60	81.57
12	83.31	66.69	72.22	97.97	27.37	254.2	93.73	96.27	63.49	81.43

aesthetic and SURF-retained of the generated videos.

4.4. Ablation

In this section, we present ablation studies to investigate the impact of each key component on the trade-off between computational efficiency and generated video quality.

Step division. We look into how to balance efficiency and quality of generating a useful preview. We have found that the division of step range (i.e., the choice of k) is very important. Since the efficiency of our method is directly proportional to the number of denoising steps performed at the initial resolution, identifying the optimal step range for the transition is crucial for balancing quality and speed. With the results presented in Tab. 4, it reveals a distinct trade-off. Employing an early transition (e.g., after step 5) accelerates the process but leads to a degradation in both layout integrity and motion quality. Conversely, delaying the transition (e.g., after step 35) not only increases the inference time but can also compromise the final layout, as the late-stage resolution shift may disrupt an already well-defined structure. Besides, We observe the denoising process of video generation that the overall compositional layout of the generated content stabilizes after approximately step 10. Based on this analysis, we identify the 10-30 step range as the optimal configuration. This range effectively preserves the structural and motion quality of the generated video.

Shift-window attention. We further investigate the impact of incorporating shift-window attention in Refiner on video generation quality. As shown in Fig. 8, we observe that previously unclear and distorted hands now exhibit clear fingernail contours and appropriate wrinkles, regardless of whether the shift-window mechanism is applied. Additionally, the clarity of distant trees is noticeably enhanced. our findings suggest that a full receptive field from global

Table 4. **Ablation study** of step division in preview stage. The recommended setting is in gray.

Setting	QS \uparrow	AQ \uparrow	DD \uparrow	MS \uparrow	OC \uparrow	Time \downarrow
5-35	82.21	<u>63.45</u>	66.67	98.23	27.12	201s
10-30	82.01	62.87	70.83	98.16	27.51	252s
20-20	81.19	62.54	<u>69.44</u>	98.05	27.57	369s
30-10	80.78	61.37	70.83	98.05	<u>27.52</u>	481s
40-0	82.89	66.57	68.06	97.70	27.35	610s

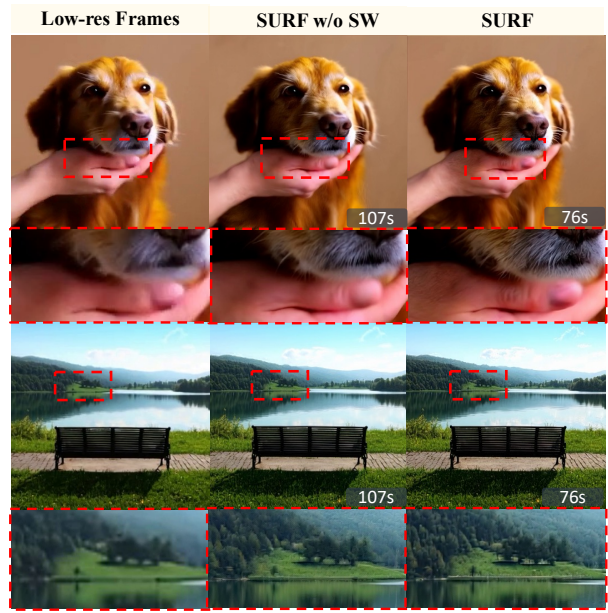


Figure 8. **Qualitative results for ablation.** SW denotes shift-window attention. The Refiner runtime on 1080p video is shown in the bottom-right corner. Our method achieves shorter runtime while preserving finer details.

attention is not critical for the refinement stage. The visual differences are negligible, indicating that local context modeling is sufficient for enhancing details at this stage.

Inference hyperparameters. We further investigate the influence of common inference-time hyperparameters on video generation performance. The performance trends across different steps in Refiner are visualized, with corresponding quantitative results summarized in Tab. 5. We analyze the effects of varying the number of diffusion steps and highlight the best-performing configurations in gray.

5. Conclusion

We present SURF, a simple yet effective framework for efficiently generating high-resolution videos while retaining signatures (e.g., layout, semantics, and motion) of a given pretrained model. SURF delivers substantial improvements in both human perceptual preference and quantitative performance metrics. It can also act as plug-in with other SOTA models. In particular, it attains a 12.5 \times speedup for generating 5-second, 16-fps, 720p videos with Wan 2.1 and 8.7 \times speedup for HunyuanVideo.

Acknowledgements. This work is supported by the National Natural Science Foundation of China (No. 62422606, 624B2124) and Hong Kong Research Grant Council General Research Fund (No. 17213925).

References

- [1] Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. Videophy: Evaluating physical commonsense for video generation. *arXiv preprint arXiv:2406.03520*, 2024. 5
- [2] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation, 2023. 3
- [3] Han Cai, Junyan Li, Muyan Hu, Chuang Gan, and Song Han. Efficientvit: Lightweight multi-scale attention for high-resolution dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 17302–17313, 2023. 3
- [4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021. 5
- [5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. 3
- [6] Shoufa Chen, Chongjian Ge, Shilong Zhang, Peize Sun, and Ping Luo. Pixelflow: Pixel-space generative models with flow. *arXiv preprint arXiv:2504.07963*, 2025. 2
- [7] Jiaxiang Cheng, Pan Xie, Xin Xia, Jiashi Li, Jie Wu, Yuxi Ren, Huixia Li, Xuefeng Xiao, Shilei Wen, and Lean Fu. Resadapter: Domain consistent resolution adapter for diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 2438–2446, 2025. 3
- [8] Krzysztof Choromanski, Valerii Likhoshesterov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarnlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. *arXiv preprint arXiv:2009.14794*, 2020. 3
- [9] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3
- [10] Hangliang Ding, Dacheng Li, Runlong Su, Peiyuan Zhang, Zhijie Deng, Ion Stoica, and Hao Zhang. Efficient-vdit: Efficient video diffusion transformers with attention tile, 2025. 2
- [11] Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. *arXiv preprint arXiv:2412.15689*, 2024. 3
- [12] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no \$\$\$\$. In *CVPR*, 2024. 3
- [13] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024. 3
- [14] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024. 3
- [15] Linjiang Huang, Rongyao Fang, Aiping Zhang, Guanglu Song, Si Liu, Yu Liu, and Hongsheng Li. Fouriscale: A frequency perspective on training-free high-resolution image synthesis, 2024. 3
- [16] Ziqi Huang, Yanan 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. 5
- [17] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pp. 5156–5165. PMLR, 2020. 3
- [18] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuo Zhuo 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. 2, 6
- [19] Kuaishou. Kling ai. <https://klingai.com/>, 2024. 2
- [20] Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

- [21] Xinyu Liu, Yingqing He, Lanqing Guo, Xiang Li, Bu Jin, Peng Li, Yan Li, Chi-Min Chan, Qifeng Chen, Wei Xue, Wenhan Luo, Qifeng Liu, and Yike Guo. Hiprompt: Tuning-free higher-resolution generation with hierarchical mllm prompts, 2024. 3
- [22] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2
- [23] Fanqing Meng, Jiaqi Liao, Xinyu Tan, Wenqi Shao, Quanfeng Lu, Kaipeng Zhang, Yu Cheng, Dianqi Li, Yu Qiao, and Ping Luo. Towards world simulator: Crafting physical commonsense-based benchmark for video generation. *arXiv preprint arXiv:2410.05363*, 2024. 5
- [24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021. 5
- [25] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024. 2
- [26] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks, 2024. 3
- [27] Jingjing Ren, Wenbo Li, Zhongdao Wang, Haoze Sun, Bangzhen Liu, Haoyu Chen, Jiaqi Xu, Aoxue Li, Shifeng Zhang, Bin Shao, Yong Guo, and Lei Zhu. Turbo2k: Towards ultra-efficient and high-quality 2k video synthesis, 2025. 2
- [28] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 5
- [29] Shuwei Shi, Wenbo Li, Yuechen Zhang, Jingwen He, Biao Gong, and Yinqiang Zheng. Resmaster: Mastering high-resolution image generation via structural and fine-grained guidance, 2024. 3
- [30] Nikita Starodubcev, Denis Kuznedelev, Artem Babenko, and Dmitry Baranchuk. Scale-wise distillation of diffusion models. *arXiv preprint arXiv:2503.16397*, 2025. 2
- [31] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neuro-computing*, 568:127063, 2024. 5
- [32] The FastVideo Team. Fastvideo: A unified framework for accelerated video generation, April 2024. URL <https://github.com/hao-ai-lab/FastVideo>. 5, 6
- [33] Ye Tian, Xin Xia, Yuxi Ren, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Yunhai Tong, Ling Yang, and Bin Cui. Training-free diffusion acceleration with bottleneck sampling. *arXiv preprint arXiv:2503.18940*, 2025. 3
- [34] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chenwei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wenten Wang, Wenting Shen, Wenyan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhifan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2, 6
- [35] Jianyi Wang, Zhijie Lin, Meng Wei, Yang Zhao, Ceyuan Yang, Fei Xiao, Chen Change Loy, and Lu Jiang. Seedvr: Seeding infinity in diffusion transformer towards generic video restoration. *arXiv preprint arXiv:2501.01320*, 2025. 3
- [36] Sinong Wang, Belinda Z Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020. 3
- [37] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *International Conference on Computer Vision Workshops (ICCVW)*. 5
- [38] Haoning Wu, Erli Zhang, Liang Liao, Chaofeng Chen, Jingwen Hou, Annan Wang, Wenxiu Sun, Qiong Yan, and Weisi Lin. Exploring video quality assessment on user generated contents from aesthetic and technical perspectives. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20144–20154, 2023. 5
- [39] Haocheng Xi, Shuo Yang, Yilong Zhao, Chenfeng Xu, Muyang Li, Xiuyu Li, Yujun Lin, Han Cai, Jintao Zhang, Dacheng Li, et al. Sparse videogen: Accelerat-

- ing video diffusion transformers with spatial-temporal sparsity. *arXiv preprint arXiv:2502.01776*, 2025. 3, 5, 6
- [40] Yifei Xia, Suhan Ling, Fangcheng Fu, Yujie Wang, Huixia Li, Xuefeng Xiao, and Bin Cui. Training-free and adaptive sparse attention for efficient long video generation. *arXiv preprint arXiv:2502.21079*, 2025. 3
- [41] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024. 3
- [42] Shuo Yang, Haocheng Xi, Yilong Zhao, Muyang Li, Jintao Zhang, Han Cai, Yujun Lin, Xiuyu Li, Chenfeng Xu, Kelly Peng, et al. Sparse videogen2: Accelerate video generation with sparse attention via semantic-aware permutation. *arXiv preprint arXiv:2505.18875*, 2025. 3
- [43] Zhen Yang, Guibao Shen, Liang Hou, Mushui Liu, Luozhou Wang, Xin Tao, Pengfei Wan, Di Zhang, and Ying-Cong Chen. Rectifiedhr: Enable efficient high-resolution image generation via energy rectification. *arXiv preprint arXiv:2503.02537*, 2025. 3
- [44] Zhen Yang, Guibao Shen, Mingyang Li, Liang Hou, Mushui Liu, Luozhou Wang, Xin Tao, Pengfei Wan, Di Zhang, and Ying-Cong Chen. Efficient training-free high-resolution synthesis with energy rectification in diffusion models. *arXiv preprint arXiv:2503.02537*, 2025. 3
- [45] 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. 2
- [46] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10819–10829, 2022. 3
- [47] Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. Accvideo: Accelerating video diffusion model with synthetic dataset. *arXiv preprint arXiv:2503.19462*, 2025. 2, 6
- [48] Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattn: Accurate sparse attention accelerating any model inference. *arXiv preprint arXiv:2502.18137*, 2025. 2, 3
- [49] Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhenghong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025. 3
- [50] Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhao Chen, Yao Tang, and Jiajun Liang. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, pp. 145–161. Springer, 2025. 3
- [51] Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yida Zhang, Yi Jiang, Zehuan Yuan, Binyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. *arXiv preprint arXiv:2502.05179*, 2025. 3
- [52] Yuechen Zhang, Jinbo Xing, Bin Xia, Shaoteng Liu, Bohao Peng, Xin Tao, Pengfei Wan, Eric Lo, and Jiaya Jia. Training-free efficient video generation via dynamic token carving. *arXiv preprint arXiv:2505.16864*, 2025. 3, 6