

FreeSwim: Revisiting Sliding-Window Attention Mechanisms for Training-Free Ultra-High-Resolution Video Generation

Yunfeng Wu^{*1,2}, Jiayi Song^{*1,3}, Zhenxiong Tan⁴, Zihao He¹, and Songhua Liu^{†1}

¹ School of Artificial Intelligence, Shanghai Jiao Tong University

² Xi'an Jiaotong-Liverpool University

³ Xi'an Jiaotong University

⁴ National University of Singapore

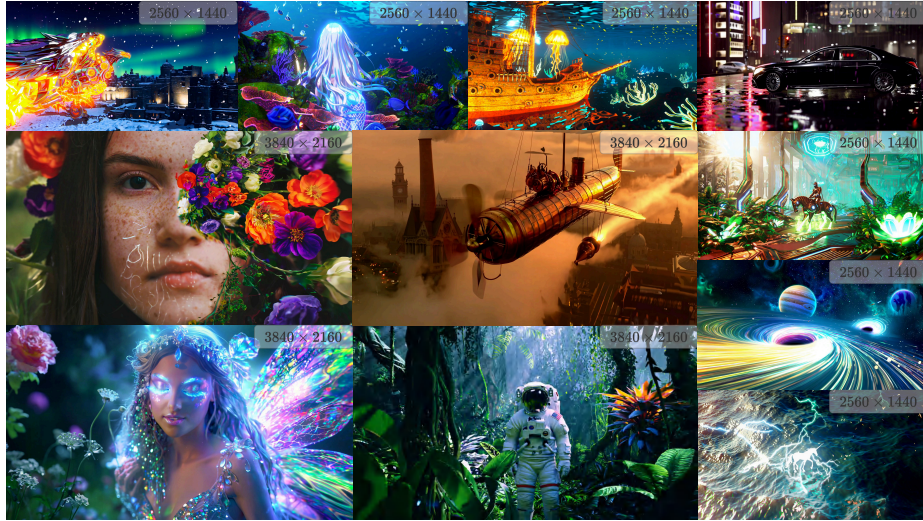


Fig. 1: Ultra-resolution results generated by our method built upon Wan2.2 [51]. Resolution is marked on the top-right corner of each result in the format of **width**×**height**. Corresponding prompts can be found in the appendix.

Abstract. The quadratic time and memory complexity of the attention mechanism in modern Transformer-based video generators makes end-to-end training for ultra-high-resolution videos prohibitively expensive. Motivated by this limitation, we introduce a training-free approach that leverages video Diffusion Transformers pre-trained at their native scale to synthesize higher-resolution videos without any additional training or adaptation. At the core of our method lies an inward sliding-window

^{*} Equal contribution.

[†] Corresponding author (liusonghua@sjtu.edu.cn).

attention mechanism, which originates from a key observation: maintaining each query token’s training-scale receptive field is crucial for preserving visual fidelity and detail. However, naive local window attention, unfortunately, often leads to repetitive content and exhibits a lack of global coherence in the generated results. To overcome this challenge, we devise a *dual-path pipeline* that backs up window attention with a novel cross-attention override strategy, enabling the semantic content produced by local attention to be guided by another branch with a full receptive field and, therefore, ensuring holistic consistency. Furthermore, to improve efficiency, we incorporate a cross-attention caching strategy for this branch to avoid the frequent computation of full 3D attention. Extensive experiments demonstrate that our method delivers ultra-high-resolution videos with fine-grained visual details and high efficiency in a training-free paradigm. Meanwhile, it achieves superior performance on VBench, even compared to training-based alternatives, with competitive or improved efficiency. Code is available [here](#).

1 Introduction

Empowered by the attention mechanism’s ability to capture complex token-wise dependencies [50], Diffusion Transformers (DiTs) [41] have shown remarkable performance in generating high-quality images [5–7, 11, 12, 17, 32, 53] and videos [13, 21, 24, 29, 36, 54]. However, the quadratic complexity of attention with respect to spatial resolution makes scaling up these models computationally prohibitive, and this issue becomes even more severe in video generation, where an additional temporal dimension further amplifies the cost. As a result, most existing diffusion models are trained at relatively low resolutions. This stands in stark contrast to the rapidly growing demand for ultra-high-definition content, such as 4K videos, in modern visual applications.

Considering the substantial challenges of training high-resolution video generation models, we pose a natural question in this work: *Can we generate ultra-high-resolution videos using existing generators pre-trained only at lower resolution, such as 480P, without any additional training or adaptation?* As illustrated in Fig. 3(a), this is a highly non-trivial problem—directly performing inference at higher resolutions often leads to unrealistic and incoherent results.

To further investigate this issue, we take inspiration from existing training-free methods for high-resolution image generation—covering aspects such as resolution-aware attention scaling [28], positional embedding adjustment [42], denoising schedule refinement [11], and low-resolution guidance strategies [3, 10]—and explore whether these techniques can be effectively extended to the video domain. Unfortunately, although these methods succeed in image generation, they fail to generalize effectively to video scenarios. As shown in Fig. 3(b), (c), and (d), they either remain stuck in inconsistent semantic layouts or suffer from poor fine-grained quality, counteracting the fundamental objective of ultra-high-resolution video synthesis.

We therefore speculate that the unsatisfactory results stem from the limited scalability of previous methods: video generation inherently requires modeling

complex 3D dependencies, making such approaches less generalizable to unseen resolution scales, even when equipped with various resolution-aware techniques. As illustrated in Fig. 2, this effect is reflected in the attention maps: those from Lumina-Next [12], a DiT-based image generator, remain highly consistent across different resolution scales, whereas those from Wan2.1 [51], a state-of-the-art video DiT, fail to maintain such consistency.

Based on this analysis, the key to resolving the issue, in fact, lies in avoiding unseen generalization—specifically, preserving for each query token the same spatial receptive field that it experienced during training. For example, in a DiT trained for 480P video generation, each query token attends only to key/value tokens within its local 480P window when synthesizing higher-resolution videos. This forms the core of our approach: an inward sliding-window attention mechanism. As illustrated in Fig. 2(g), this design substantially improves generation fidelity, producing fine-grained and visually rich details.

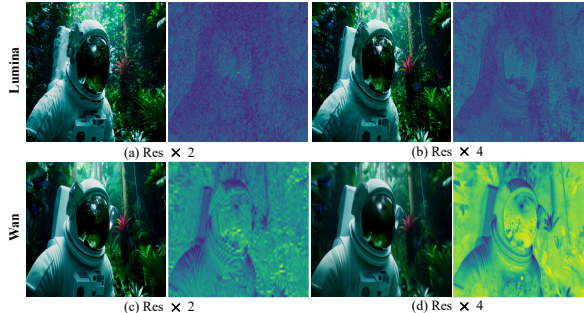


Fig. 2: Visual results and corresponding attention maps at $\times 2$ and $\times 4$ native resolutions from Lumina-Next [12] and Wan2.1 [51], based on the same DiT architecture family.

We then turn back to the full-attention results shown in Fig. 3(d), which effectively preserve the overall semantic layout despite lacking fine visual details, and complement the proposed window attention approach with an additional full-attention branch. As shown in Fig. 3(h), by interacting through an override scheme at the cross-attention layers—the direct sources of semantic information—such a dual-branch pipeline effectively brings the best of both worlds, achieving both coherent semantics and visually appealing fine-grained details.

Moreover, the full-attention branch inevitably incurs substantial computational overhead. To mitigate this, inspired by recent caching-based acceleration techniques [31, 34, 35, 58], we introduce a cross-attention feature reuse strategy that reduces the need for frequent full 3D attention computation, achieving over a $2\times$ speedup with negligible performance degradation.

We conduct extensive experiments on modern DiT-based video generation models, *e.g.*, Wan2.1 [51] and LTX-Video [13]. As shown in Fig. 1, our method can synthesize 4K-resolution videos with strong semantic coherence and visually compelling fine-grained details. Quantitatively, it establishes new state-of-the-art results on VBench [25], even surpassing previous training-based methods in



Fig. 3: Qualitative comparison of generated results under different strategies based on Wan2.1 [51]. Except for (a) and (b), all other results are produced using a coarse-to-fine scheme, where a base video (832×480) is first generated by the convention text-to-video pipeline, followed by high-resolution refinement (1920×1088) through SDEdit [38].

both generation quality and efficiency. Our contributions can be summarized as follows:

- We delve into the field of training-free ultra-high-resolution video generation and propose the first approach tailored for Wan2.1 [51], to the best of our knowledge.
- We identify the limitations of prior resolution extrapolation methods for image generation when applied to videos and propose an inward sliding-window attention mechanism complemented by a dual-path strategy to effectively overcome these issues.
- Extensive experiments and user evaluations confirm that our method achieves superior performance in training-free ultra-high-resolution video generation, outperforming even training-based counterparts.

2 Related Work

2.1 Diffusion Models for Video Generation

Diffusion models (DMs) [18, 49] have been extensively explored for video generation, attracting substantial attention. Unlike conventional UNet-based video diffusion models [2, 4, 16, 20] that rely on convolutional backbones, Diffusion Transformer (DiT) [41] adopts a Transformer architecture as its core. This architectural change enables DiT to capture richer long-range dependencies and more intricate spatiotemporal relationships within video sequences. Early works [27,

37, 54] demonstrated the effectiveness of spatiotemporal transformers with self-attention and a global receptive field. More recently, LTX-Video [13] further optimized the interaction between the Video-VAE and the denoising transformer, while Wan2.1 [51] and Hunyuan [29], both trained on large-scale video datasets, exhibited impressive performance in generating realistic videos. Although these models greatly improve generation quality, their native resolution remains limited for high-quality applications.

2.2 High-Resolution Visual Generation

In the image domain, training-free high-resolution visual generation has been extensively explored. Most existing approaches [43] are built upon U-Net [48], which significantly suffers from repetitive patterns that arise from the limited local receptive fields in high-resolution synthesis. Several works [1, 9, 15] further expand receptive fields through dilated convolutions and fuse local and global patches to effectively suppress repetition. With the advent of recent foundational diffusion models [30], DiT have become the dominant architecture, benefiting from the attention mechanism’s capacity to model complex token-wise dependencies. Training-free methods [3, 10] on DiT are capable of maintaining global layout consistency while synthesizing fine-grained visual details.

In the video domain, high-resolution synthesis of DiT-based models introduce additional challenges, including severe computational overhead, blurring, and structural distortion. Some works [55] concentrate on developing the system infrastructure required for efficient high-resolution video generation. Most current video generation methods [23, 44, 46] rely on fine-tuning with high-resolution data. In this work, we propose a fully training-free video generation framework tailored for modern DiT architectures, effectively ensuring global layout consistency while producing fine-grained details in high-resolution synthesis.

3 Methods

Generating 4K videos poses significant challenges, as it requires both fine-grained detail synthesis and high-level semantic coordination. Following prior works [3, 10], we adopt a coarse-to-fine generation strategy (Sec. 3.1): first generating a video from a text prompt at the model’s native resolution and then producing a higher-resolution version based on this initial output. Since the first stage follows the standard inference procedure of a pre-trained text-to-video DiT, our focus in this section is on the second stage. The key components include an inward sliding-window attention mechanism (Sec. 3.2), a cross-attention override scheme (Sec. 3.3), and a feature reuse strategy for the cross-attention layers (Sec. 3.4). An overview of our method is illustrated in Fig. 4.

3.1 Preliminary: Coarse-to-Fine Generation

As shown in Fig. 3(a), it is difficult for a text-to-video DiT pre-trained only at low resolution scales to generate high-resolution content. Although prior works [10,

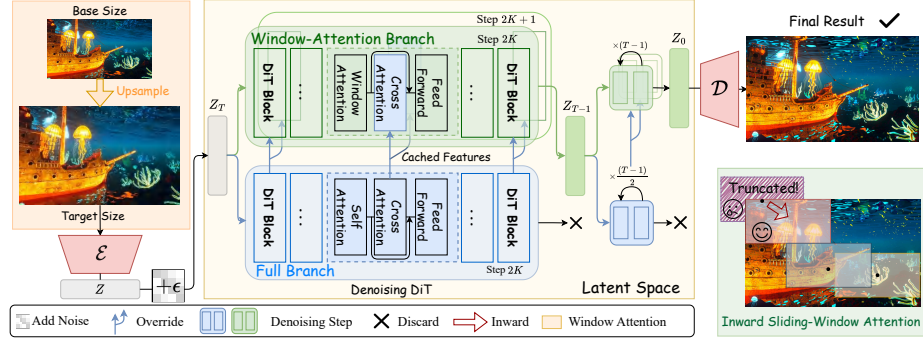


Fig. 4: FreeSwim Framework Overview. **Upper Left:** Our parallel Dual-Path pipeline combined with cross-attention override to achieve correct global semantic structure, with the Full-Branch Feature Reuse strategy in the figure representing the case where the cross attention of Full-Branch is computed and updated every two steps ($P=2$). **Bottom Right:** Our inward window attention ensures that during inference, the spatial dimension is strictly controlled at the same scale as during training.

[11, 28, 42] propose a series of insightful toolkits for resolution extrapolation, they are helpful for enhancing local details but still fail to alleviate the issue of disordered layouts in high-resolution video generation, as illustrated in Fig. 3(b). Therefore, following [3, 10], we apply a coarse-to-fine pipeline, where the overall video layouts are handled in the low-resolution stage.

Specifically, inspired by SDEdit [38], we first conduct text-to-video generation in the model’s native resolution. Then, as shown in Fig. 4, this basic result is up-sampled to the target size in the pixel space and converted to the latent space through the VAE encoder. A random Gaussian noise is added to the encoded latents with a predefined strength, which preserves the overall content structure while perturbing fine-grained details. The high-resolution inference process begins from this noisy state. As shown in Fig. 3(c) and (d), this strategy plays a crucial role in producing reasonable content layouts and yields further performance gains when combined with the aforementioned inference toolkits.

3.2 Inward Sliding-Window Attention

Nevertheless, in video generation, where a model has to capture complex 3D interactions, the quality of fine-grained details remains unsatisfactory. In other words, previous methods effective in 2D resolution extrapolation struggle to generalize to video scenarios when the token count surpasses that observed during training.

Based on this insight, a straightforward way to maintain the same token count N as during training is to reduce the number of latent frames F so that $N = H \times W \times F$, where H and W denote the sizes of the spatial dimensions. For example, using a 480P model to generate 1080P content requires reducing the

frame count by $2.25\times$ to keep the token length unchanged. However, as shown in Fig. 3(e), this strategy brings limited improvement.

We therefore hypothesize that preserving the native size of each dimension—both spatial and temporal—is essential for performance. Methodologically, this corresponds to a sliding-window attention mechanism: each query token attends only to key/value tokens within a local window whose receptive field matches the resolution scale used during training, *e.g.*, using a 480P window for a 480P model.

Unlike classic window attention, such as Swin Transformer [33] and Neighborhood Attention [14], which truncate attention windows for boundary tokens due to limited spatial context, we find it essential to maintain a consistent number of interactable key/value tokens for all queries, including those at the boundaries. To this end, we propose a novel inward sliding-window attention mechanism: when a query’s window extends beyond the video boundaries, we dynamically shift it inward based on the token’s spatial location, ensuring that its receptive field size matches the training resolution without truncation, as shown in Fig. 4(bottom right), which can be formulated as:

$$\begin{aligned}\Delta_w^{(q)} &= \max(\frac{w}{2} - x_q, \frac{w}{2} + x_q - W + 1, 0), \\ \Delta_h^{(q)} &= \max(\frac{h}{2} - y_q, \frac{h}{2} + y_q - H + 1, 0), \\ M_{qk} &= \begin{cases} 1, & \text{if } |x_q - x_k| \leq \frac{w}{2} + \Delta_w^{(q)} \\ & \text{and } |y_q - y_k| \leq \frac{h}{2} + \Delta_h^{(q)}, \\ 0, & \text{otherwise,} \end{cases} \\ O &= \text{Softmax}((QK^\top) \cdot M / \sqrt{d})V,\end{aligned}\tag{1}$$

where x_* and y_* denote the spatial coordinates of a token; W and H are the target dimensions, while w and h are the native dimensions; Q , K , V , and M represent the query, key, value, and attention mask matrices, respectively; $\cdot$ indicates element-wise multiplication; d is the feature dimension size; and O is the resulting attention output of the layer.

This inward sliding-window attention scheme is applied to all self-attention layers—the core modules for token interactions in modern DiT-based video generators like Wan2.1—in place of the original 3D full attention. As shown in Fig. 3(f) and (g), this approach significantly enhances local details, while the inward mechanism effectively eliminates boundary artifacts.

3.3 Dual-Path with Cross-Attention Override

As shown in Fig. 3(g), the inward sliding window attention mechanism mentioned in Sec. 3.2, while improving fine-grained details, inevitably produces repetitive patterns as a side effect. The underlying reason is the confined respective field of each token, analogous to that in previous U-Net-based generators [48] like SD-1.5 [47] and SD-XL [43], where convolution serves as the primary means of

feature interaction. A variety of methods have been developed to enable resolution extrapolation for them [9, 15, 45, 56].

In this paper, we provide a tailored approach for DiT-based models. Re-examining Fig. 3(d) and (g), we observe a trade-off between global layout consistency and fine-grained detail quality. In other words, although full 3D attention produces less refined local details, it yields more coherent semantics, suggesting a potential source of global guidance in a dual-path framework that integrates both full-attention and window-attention branches. For the interaction between the two paths, we leverage the latent-to-text cross-attention modules, which serve as the direct source of semantic content.

Specifically, at each denoising step, we duplicate the input latent maps, concatenate them along the batch dimension, and apply full 3D attention and the proposed inward sliding-window attention to the self-attention modules, respectively. The cross-attention results of the latter branch, denoted as O_{Cross}^{Window} , are guided by the former, denoted as O_{Cross}^{Full} , with a strength of λ :

$$O_{Cross}^{Window} \leftarrow \lambda O_{Cross}^{Full} + (1 - \lambda) O_{Cross}^{Window}. \quad (2)$$

In practice, we find that setting $\lambda = 1$ achieves the most robust performance without sacrificing fine-grained detail. This corresponds to an override scheme, where the cross-attention features from the window-attention branch are entirely replaced by those from the full branch. Please refer to Sec. 4.4 for more discussion. As shown in Fig. 3(h), the proposed inward sliding-window attention, reinforced by this dual-path design, brings the best of both worlds, enabling training-free high-resolution video generation with coherent semantic structures and appealing details.

3.4 Full-Branch Feature Cache and Reuse

By default, at each denoising step, a latent map must pass through the full-attention branch to compute cross-attention guidance for the window-attention branch, which introduces substantial computational overhead, especially at ultra-high-resolution scales like 4K. To address this, motivated by prior work on caching for diffusion models [31, 34, 35, 58], we propose a step-wise reuse strategy for the cross-attention features from the full branch.

Specifically, we update and cache the cross-attention guidance from the full-attention branch every P steps. The cached results are then reused to override the cross-attention outputs of the window-attention branch in the intervening steps, which has minimal impact on video quality when P is set to $2 \sim 8$ according to our experiments. Fig. 4 presents the case of $P = 2$.

Moreover, for models using classifier-free guidance [19], we find it unnecessary to apply the dual-path pipeline to the unconditional branch. In other words, we omit full-attention guidance for this branch and rely solely on inward sliding-window attention in the self-attention layers. Together, these two designs achieve over a more than $2\times$ inference speedup at 1080P resolution.

Models	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Overall Consistency	Overall Score
LTX-Video [13]							
CineScale (w/o LoRA) [44]	89.7%	97.3%	98.7%	<u>41.2%</u>	36.6%	16.2%	<u>63.3%</u>
Ours (w/o Cache)	<u>91.3%</u>	97.3%	<u>98.5%</u>	41.3%	50.0%	<u>18.5%</u>	66.2%
Ours (w/ Cache, P=5)	91.4%	<u>97.2%</u>	<u>98.5%</u>	41.3%	<u>49.9%</u>	18.6%	66.2%
Wan2.1 14B [51]							
CineScale (w/o LoRA) [44]	96.3%	96.5%	98.8%	<u>59.9%</u>	<u>65.1%</u>	<u>23.8%</u>	<u>73.4%</u>
Ours (w/o Cache)	<u>95.7%</u>	<u>96.4%</u>	<u>98.3%</u>	60.9%	65.8%	24.6%	73.6%
Wan2.1 1.3B [51]							
Real-ESRGAN [52]	94.7%	96.4%	97.9%	55.9%	<u>63.9%</u>	24.8%	72.3%
Upscale-A-Video [57]	91.0%	97.0%	97.2%	55.6%	62.9%	22.9%	71.1%
I-Max [10]	92.8%	96.9%	95.3%	<u>59.6%</u>	61.1%	23.0%	71.5%
HiFlow [3]	90.7%	96.3%	95.2%	54.8%	54.4%	24.7%	69.4%
CineScale (w/o LoRA) [44]	94.3%	96.9%	98.2%	45.9%	40.9%	16.8%	65.5%
CineScale (w/ LoRA) [44]	93.9%	96.8%	97.4%	56.7%	60.7%	23.5%	71.5%
Ours (w/o Cache)	94.7%	97.2%	97.7%	57.9%	63.4%	<u>25.1%</u>	72.7%
Ours (w/o CFG-Full)	97.1%	98.2%	98.2%	58.6%	61.8%	<u>25.1%</u>	<u>73.2%</u>
Ours (w/ Cache, P=2)	95.1%	<u>97.6%</u>	97.9%	61.2%	65.3%	24.9%	73.7%
Ours (w/ Cache, P=5)	94.4%	<u>97.4%</u>	97.8%	56.2%	62.6%	25.2%	72.3%
Ours (w/ Cache, P=8)	<u>95.2%</u>	<u>97.6%</u>	<u>98.0%</u>	54.3%	61.0%	23.6%	71.7%

Table 1: Video comparison with both training-based and training-free models at 1080P. Our method FreeSwim achieves the best scores across almost all metrics and ranks first overall in the Wan Series [51] and LTX Series [13] respectively. The highest value is **bold**, and the second-highest is underlined.

4 Experiments

4.1 Settings and Implementation Details

In this part, we conduct experiments using Wan2.1 [51] and LTX-Video [13], known for their state-of-the-art performance in video generation. Wan2.1 [51] includes both a 1.3B version trained solely at 832×480 (480P) resolution and a 14B version trained on a mixed dataset of 480P and 1280×720 (720P). LTX-Video [13] is trained solely on 512×768 . More results from other DiT models are presented in the Appendix. We utilize FlexAttention [8] in PyTorch [40], which provides efficient low-level optimization for sparse attention, to implement the proposed inward sliding-window attention.

We evaluate the performance of our method using VBench [25], a benchmark designed to assess both visual quality and semantic coherence. For video generation at resolutions of 1920×1088 (1080P) and 3380×1920 (3K), we randomly select 60 prompts from the standard prompt suite of VBench [25] and, to account for randomness, we randomly sample five different seeds for each prompt. Following the official evaluation settings of VBench, each method generates five videos per prompt using these five distinct random seeds. We then apply the official VBench metrics to ensure a fair comparison across all methods.

4.2 Main Comparisons

We evaluate our method on LTX-Video [13] and Wan2.1 [51]. We consider the following baselines: (1) **Low-Level Super Resolution Methods**, including Real-ESRGAN [52] and Upscale-A-Video [57], whose base videos are derived from Wan2.1 [51]; (2) **Training-Free High-Resolution Generation Methods**, including I-Max [10], HiFlow [3], and CineScale [44] without LoRA, *i.e.*, a coarse-to-fine strategy based on [38] that is also adopted in our approach; and (3) **Training-Based High-Resolution Generation Methods**, including CineScale [44] with LoRA [22], which conducts adaptation of high-resolution video data at 2K resolution.

As shown in Tabs. 1 and 2, our method demonstrates superior performance across both 1080P and 3K resolutions, particularly in terms of aesthetic quality (including layout, color richness, and harmony), imaging quality (capturing distortions such as over-exposure, noise, and blur), and overall consistency (reflecting both semantic and style alignment). These results validate the effectiveness of our method, enhancing fine-grained aesthetic details and improving the global-layout accuracy of the generated videos. Our method remains competitive to the baselines on other metrics and ranks first on the overall scores.

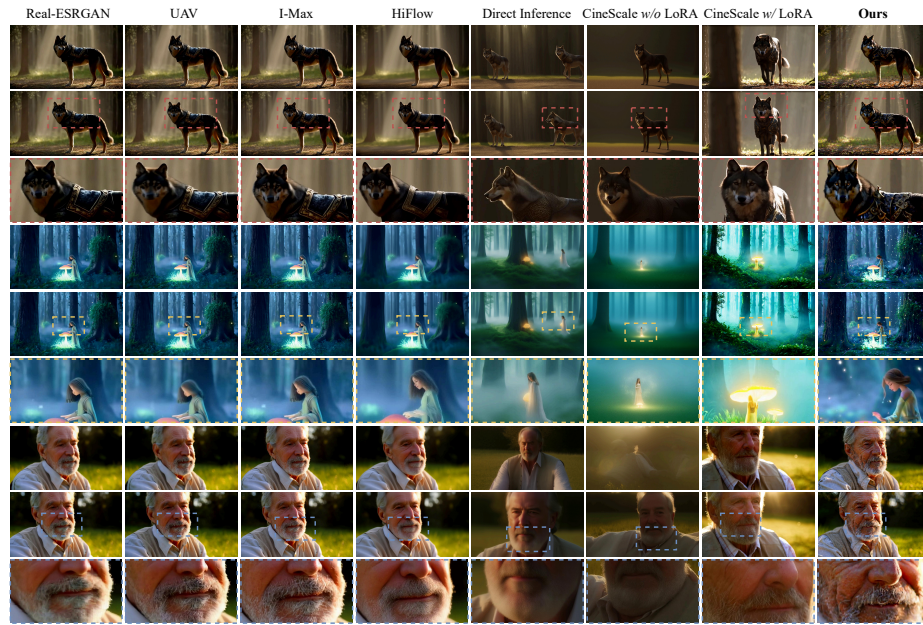


Fig. 5: Visual comparison of synthesized 1080P videos for models based on Wan. Our method FreeSwim yields high-resolution videos characterized by high-fidelity details and coherent structure. Best viewed zoomed in.

To visually demonstrate the superiority of our approach, we conduct a fair comparison on the same base videos generated by Wan2.1-1.3B [51], as shown in Fig. 5. For training-based SR baselines [52, 57], we directly feed the generated base videos into their pipelines. For training-free high-resolution generation methods [3, 10, 44], we adapt their Wan2.1-based implementations to ensure a consistent backbone. Under the same prompt, our method consistently produces the most visually appealing results, featuring the richest fine-grained details and the highest semantic consistency with the prompt. In particular, in the second example, our generated video vividly presents the prompt-specific details, such as the facial details, which are barely manifested in the results of other approaches.

Models	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Overall Consistency	Overall Score
CineScale <i>w/o</i> LoRA [44]	94.7	98.7	97.7	<u>43.2</u>	36.5	12.6	<u>63.9</u>
CineScale <i>w/</i> LoRA [44]	<u>91.8</u>	97.7	<u>99.2</u>	46.0	52.2	<u>17.5</u>	67.4
Ours (<i>w/o</i> Cache)	95.2	<u>96.7</u>	<u>99.2</u>	46.0	<u>52.2</u>	17.5	67.4
Ours (<i>w/</i> Cache, P=5)	94.7	<u>96.0</u>	<u>99.4</u>	50.1	<u>57.8</u>	22.5	70.1

Table 2: Video comparison with both training-based and training-free variants of CineScale [44] at 3K. Our method FreeSwim achieves the best scores across almost all metrics and ranks first overall in the Wan Series [51].

4.3 User Study

To further qualitatively evaluate the performance of our method, we conduct a human study to assess the subjective aesthetic perception of the generated videos. Specifically, we compare four methods, all based on Wan2.1-1.3B [51]: the training-based approach CineScale [44] (*w/* LoRA), the training-free variant CineScale [44] (*w/o* LoRA), the direct inference results of the original model, and our proposed method. All methods are evaluated using the same prompts from VBench [25].

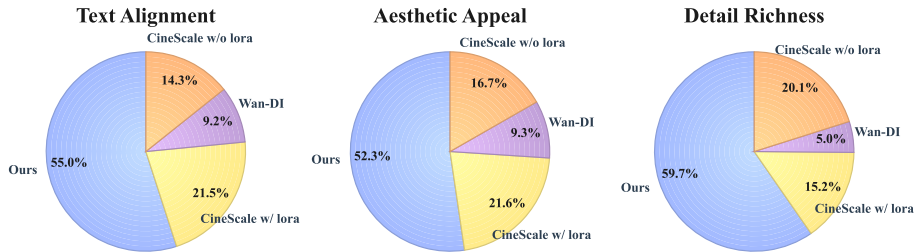


Fig. 6: Results of user study. Participants were expected to select the best method based on Text Alignment, Aesthetic Quality, and Video Quality.

Model	Window Self Attn	Cross Attn Override	w/o CFG -Full	Cache & Reuse	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Overall Consistency	Overall Score	Inference Time
Wan-Only Full	✗	✗	✗	✗	94.3%	96.9%	98.2%	45.9%	40.9%	16.8%	65.5%	66 min 46 s
Wan-Only Window(Native)	✓	✗	✗	✗	94.6%	97.1%	97.5%	57.0%	57.1%	25.1%	71.4% +5.9%	32 min 29 s
Wan-Only Window(Inward)	✓	✗	✗	✗	94.6%	97.2%	97.5%	58.2%	61.9%	25.3%	72.5% +7.0%	32 min 30 s
Ours (w/o Cache)	✓	✓	✗	✗	94.7%	97.2%	97.7%	57.9%	63.4%	25.1%	72.7% +7.2%	79 min 15 s
Ours (w/o CFG-Full)	✓	✓	✓	✗	97.1%	98.2%	98.2%	58.6%	61.8%	25.1%	73.2% +7.7%	64 min 5 s
Ours (w/ Cache, P=2)	✓	✓	✓	✓	95.1%	97.6%	97.9%	61.2%	65.3%	24.9%	73.7% +8.2%	54 min 40 s 1.2×

Table 3: Ablation study on the Wan2.1-1.3B [51] at 1080P resolution. Higher is better for all metrics except inference time.

During the study, each participant was presented with the generated videos in a randomized order and expected to select the best video based on three criteria: Aesthetic Appeal, Detail Richness, and Text Alignment. Aesthetic Appeal referred to the overall visual attractiveness and artistic quality of the video, while Detail Richness evaluated the level of fine-grained details present in the video frames. Text Alignment assessed how well the generated content adhered to the textual input, ensuring that the video accurately reflected the provided prompts. In total, 29 participants took part in the study and provided their subjective preferences. As shown in Fig. 6, our method consistently receives the highest preference across all the metrics.

4.4 Ablation Study

Contribution of Proposed Strategy. To validate the effectiveness of our proposed method, we conduct detailed ablation studies on Wan2.1-1.3B [51] as shown in Tab. 3. First, we demonstrate that incorporating *Inward Sliding-Window Attention* significantly enhances visual details compared to direct inference, resulting in substantial improvements in multiple metrics, including aesthetic quality and image quality. To address the repetition problem caused by window attention, the proposed *Dual-Path with Cross-Attention override* strategy further improves semantic consistency in the generated videos, enhancing both imaging quality and overall score. Additionally, to mitigate the computational overhead of the dual-path design, we adopt a *Full-Branch Feature Cache and Reuse* strategy, which not only significantly accelerates inference but also delivers a 1.0% improvement over the variants of *w/o* cache.

Effect of the Inward Window. To rigorously control the receptive field to match the resolution scale used during training, we propose the *Inward Sliding-Window Attention*. We provide a quantitative comparison between this method and Native Window Attention, as shown in Tab. 3 (second and third rows). The results show that the inward approach significantly improves aesthetic quality and imaging quality. And, as shown in Fig. 3(f) and (g) for a qualitative comparison, the approach effectively addresses the artifacts, confirming that these techniques contribute to the necessary quality improvements. Furthermore, by utilizing Flex Attention [8] with complied masking logic, the inward method does not introduce any computational delays, as shown in Fig. 8.

Effect of the λ Coefficient. In this section, we study the effect of the proposed *Dual-Path with Cross-Attention Override*. As a control factor of cross-

attention strength between window and full branches, λ in Eq. 2 plays a crucial role in balancing global layout consistency and fine-grained detail quality. To better understand its influence, we conduct a detailed analysis of the hyperparameter λ in the override strategy.

As shown in Tab. 4, increasing λ , which increases the proportion of the full branch, leads to improvements in aesthetic quality, image quality, and overall consistency. Specifically, when $\lambda = 1.0$, corresponding to our proposed *Dual-Path with Cross-Attention override*, the best overall score is achieved. Fig. 7

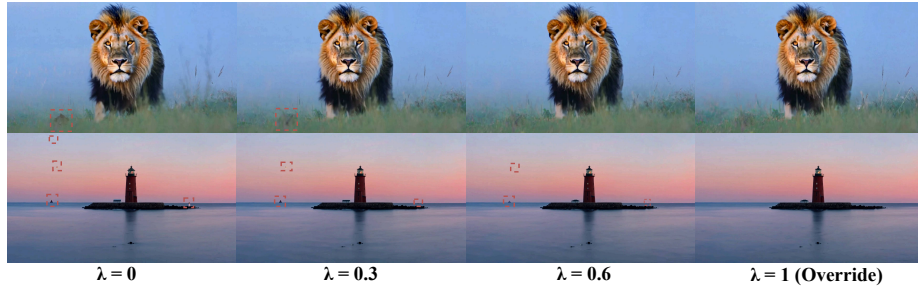


Fig. 7: Qualitative comparison of different λ under different cases. Repeated patterns are highlighted with red boxes.

further demonstrates that increasing λ helps mitigate repetition, although no single value is optimal across all cases. For instance, at $\lambda = 0.3$, both examples still exhibit repetition, whereas at $\lambda = 0.6$, the first example shows no repeated lions, but the second still displays duplicated beacons. Interestingly, the quality of fine-grained details remains largely unaffected by changes in λ . Given these results, we choose $\lambda = 1$, as it provides the best robustness across all scenarios and aligns with our *Dual-Path with Cross-Attention override* strategy.

Models	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Overall Consistency	Overall Score
$\lambda=0.0$	94.3%	<u>96.9%</u>	98.2%	45.9%	40.9%	16.8%	65.5%
$\lambda=0.3$	<u>91.5%</u>	95.8%	96.2%	58.9%	62.5%	23.5%	71.4%
$\lambda=0.6$	91.1%	96.8%	96.3%	<u>57.6%</u>	63.0%	<u>24.2%</u>	<u>71.7%</u>
$\lambda=1.0$	94.7%	97.2%	<u>97.7</u> %	57.9%	<u>63.4%</u>	<u>25.1%</u>	72.7%

Table 4: Quantitative comparison of different λ .

Exploration of Reusing Mechanisms. In this section, we study the effect of the proposed *Full-Branch Feature Reuse strategy*. As shown in Fig. 9, we first observe that the unconditional branch does not benefit from the dual-path strategy. Removing it reduces computation and results in a $1.2\times$ speedup over the dual-path baseline, even surpassing direct inference in terms of effi-

ciency. Furthermore, when using the cache, at $P = 2, 5$, and 8 , visual content is generated that maintains both fine-grained details and global structural consistency, achieving lossless acceleration compared to the *w/o cache* implementation. As shown in Fig. 8, we then examine the recompute interval P of cross-attention from the full branch. Increasing P reduces the refresh frequency and increases feature reuse, thereby improving inference efficiency. At higher resolutions, the computational cost of the full branch grows, making this caching and reuse strategy more critical for efficiency. At 4K resolution, $P = 8$ achieves up to a $2.8\times$ speedup. To balance visual quality and efficiency at any resolution, we adopt $P = 2$ by default as the most robust version. To clarify the specific acceleration logic, we provide the following time calculation formulas as shown in Eq. 3:

$$\begin{aligned}
 T_{w/o \text{ cache}} &= 2tT_w + 2tT_f, \\
 T_{w/o \text{ CFG-Full}} &= 2tT_w + tT_f, \\
 T_{w/ \text{ cache}}(P) &= 2tT_w + \left(\left\lfloor \frac{t}{P} \right\rfloor + 1 \right) T_f.
 \end{aligned} \tag{3}$$

where - $T_{w/o \text{ cache}}$ denotes the total time for inference without caching. - $T_{w/o \text{ CFG-Full}}$ denotes the total time for inference without caching and without CFG-Full. - $T_{w/ \text{ cache}}(P)$ denotes the total time for inference with caching, where P represents the recompute interval. - t denotes the number of iterations or steps. - T_w denotes the time spent on the weighted operation (e.g., feature computation). - T_f denotes the time spent on feature reuse or caching. - P denotes the recompute interval for cross-attention in the full branch, which determines how often the cached features are refreshed.

The effectiveness of this reuse and caching strategy is further confirmed by the quantitative results in Tabs. 1(bottom) and 3.

VRAM Usage Analysis for Different Processing Pipelines. We analyze the VRAM consumption of our processing pipeline across different resolutions. For the $\{1080P, 2K, 3K, 4K\}$ resolutions, the VRAM usage for our method in the *w/o cache* version is $\{10.9 \text{ GB}, 18.0 \text{ GB}, 30.3 \text{ GB}, 37.3 \text{ GB}\}$, respectively. While the Full-Branch Feature Cache and Reuse increase VRAM overhead, offloading modes and features to the CPU allows us to effectively control VRAM

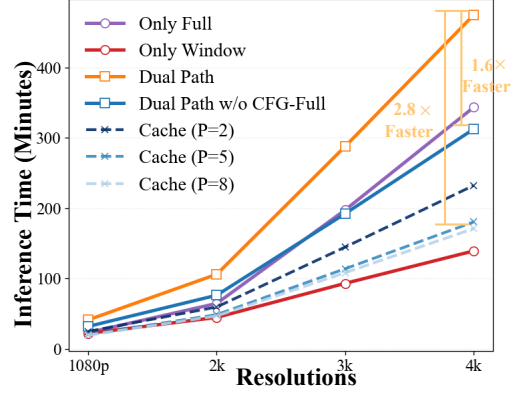


Fig. 8: Impact of different processing pipelines on inference time.

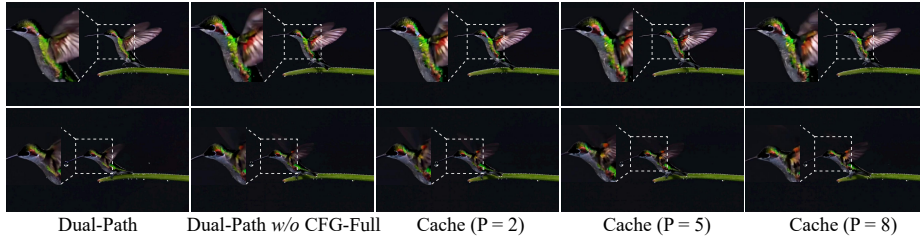


Fig. 9: Five processing pipelines are presented, where the inference speed increases from left to right. However, visual details are preserved without loss, ensuring a consistent global layout throughout.

usage, keeping it at {10.9 GB, 18.0 GB, 30.3 GB, 37.3 GB} for the {1080P, 2K, 3K, 4K} resolutions.

5 Conclusion

In this paper, we introduce FreeSwim, a novel training-free paradigm aimed at ultra-high-resolution video generation by leveraging pre-trained text-to-video diffusion transformers trained only at low resolution scales. By introducing an inward sliding-window attention scheme and a window-attention and full-attention dual-path approach with cross-attention override, FreeSwim effectively addresses the underlying challenges in high-resolution generation, such as repetitive patterns and quality degradation, without the need for any data collection or model training. Moreover, our method supports accelerating inference by caching and reusing cross-attention from the full-attention branch, resulting in negligible performance degradation. Experimental results demonstrate the superiority and efficiency of FreeSwim in ultra-high-resolution video generation, surpassing existing methods in video quality.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
2. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., Rombach, R.: Stable video diffusion: Scaling latent video diffusion models to large datasets (2023), <https://arxiv.org/abs/2311.15127>
3. Bu, J., Ling, P., Zhou, Y., Zhang, P., Wu, T., Dong, X., Zang, Y., Cao, Y., Lin, D., Wang, J.: Hiflow: Training-free high-resolution image generation with flow-aligned guidance (2025)
4. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., Weng, C., Shan, Y.: Videocrafter1: Open diffusion models for high-quality video generation (2023), <https://arxiv.org/abs/2310.19512>
5. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation (2024)
6. Chen, Z., Fang, G., Ma, X., Yu, R., Wang, X.: dparallel: Learnable parallel decoding for dllms. In: International Conference on Learning Representations (2026)
7. Chow, W., Li, L., Kong, L., Li, Z., Xu, Q., Song, H., Ye, T., Wang, X., Bai, J., Xu, S., Li, X., Pan, J., Liu, S., Zhou, R., Yang, T., Liu, S.: Editmgt: Unleashing potentials of masked generative transformers in image editing (2026)
8. Dong, J., Feng, B., Guessous, D., Liang, Y., He, H.: Flex attention: A programming model for generating optimized attention kernels (2024), <https://arxiv.org/abs/2412.05496>
9. Du, R., Chang, D., Hospedales, T., Song, Y.Z., Ma, Z.: Demofusion: Democratising high-resolution image generation with no \$\$\$ (2023)
10. Du, R., Liu, D., Zhuo, L., Qi, Q., Li, H., Ma, Z., Gao, P.: I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow (2024)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., Rombach, R.: Scaling rectified flow transformers for high-resolution image synthesis (2024), <https://arxiv.org/abs/2403.03206>
12. Gao, P., Zhuo, L., Liu, D., Du, R., Luo, X., Qiu, L., Zhang, Y., Lin, C., Huang, R., Geng, S., Zhang, R., Xi, J., Shao, W., Jiang, Z., Yang, T., Ye, W., Tong, H., He, J., Qiao, Y., Li, H.: Lumina-t2x: Transforming text into any modality, resolution, and duration via flow-based large diffusion transformers (2024), <https://arxiv.org/abs/2405.05945>
13. 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., Bibi, O.: Ltx-video: Realtime video latent diffusion (2024), <https://arxiv.org/abs/2501.00103>
14. Hassani, A., Walton, S., Li, J., Li, S., Shi, H.: Neighborhood attention transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6185–6194 (2023)
15. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models (2023)

16. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation (2023)
17. He, Z., Wu, Y., Wang, X., Liu, S.: Bend the basics: Degradation-aware deformable tokenization for all-in-one image restoration. In: Forty-third International Conference on Machine Learning (2026), <https://openreview.net/forum?id=ZF0A8DZqhl>
18. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* **33**, 6840–6851 (2020)
19. Ho, J., Salimans, T.: Classifier-free diffusion guidance (2022)
20. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022)
21. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers (2022)
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models (2021), <https://arxiv.org/abs/2106.09685>
23. Hu, T., Zhang, J., Su, Z., Yi, R.: Ultragen: High-resolution video generation with hierarchical attention (2025), <https://arxiv.org/abs/2510.18775>
24. Hu, Y., Gong, H., Yang, C., An, Z., Xu, Y., Liu, S.: Multianimate: Pose-guided image animation made extensible (2026)
25. Huang, Z., He, Y., Yu, J., Zhang, F., Si, C., Jiang, Y., Zhang, Y., Wu, T., Jin, Q., Chanpaisit, N., Wang, Y., Chen, X., Wang, L., Lin, D., Qiao, Y., Liu, Z.: Vbench: Comprehensive benchmark suite for video generative models (2023), <https://arxiv.org/abs/2311.17982>
26. Issachar, N., Yariv, G., Benaim, S., Adi, Y., Lischinski, D., Fattal, R.: Dype: Dynamic position extrapolation for ultra high resolution diffusion (2026), <https://arxiv.org/abs/2510.20766>
27. Jin, Y., Sun, Z., Li, N., Xu, K., Xu, K., Jiang, H., Zhuang, N., Huang, Q., Song, Y., Mu, Y., Lin, Z.: Pyramidal flow matching for efficient video generative modeling (2025)
28. Jin, Z., Shen, X., Li, B., Xue, X.: Training-free diffusion model adaptation for variable-sized text-to-image synthesis. *Advances in Neural Information Processing Systems* **36**, 70847–70860 (2023)
29. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., Wu, K., Lin, Q., Yuan, J., Long, Y., Wang, A., Wang, A., Li, C., Huang, D., Yang, F., Tan, H., Wang, H., Song, J., Bai, J., Wu, J., Xue, J., Wang, J., Wang, K., Liu, M., Li, P., Li, S., Wang, W., Yu, W., Deng, X., Li, Y., Chen, Y., Cui, Y., Peng, Y., Yu, Z., He, Z., Xu, Z., Zhou, Z., Xu, Z., Tao, Y., Lu, Q., Liu, S., Zhou, D., Wang, H., Yang, Y., Wang, D., Liu, Y., Jiang, J., Zhong, C.: Hunyuanvideo: A systematic framework for large video generative models (2025)
30. Labs, B.F., Batifol, S., Blattmann, A., Boesel, F., Consul, S., Diagne, C., Dockhorn, T., English, J., English, Z., Esser, P., Kulal, S., Lacey, K., Levi, Y., Li, C., Lorenz, D., Müller, J., Podell, D., Rombach, R., Saini, H., Sauer, A., Smith, L.: Flux.1 kontext: Flow matching for in-context image generation and editing in latent space (2025), <https://arxiv.org/abs/2506.15742>
31. Liu, F., Zhang, S., Wang, X., Wei, Y., Qiu, H., Zhao, Y., Zhang, Y., Ye, Q., Wan, F.: Timestep embedding tells: It’s time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108* (2024)
32. Liu, S., Yu, R., Wang, X.: Universalbooth: Model-agnostic personalized text-to-image generation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 18314–18324 (October 2025)

33. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
34. Ma, X., Fang, G., Bi Mi, M., Wang, X.: Learning-to-cache: Accelerating diffusion transformer via layer caching. *Advances in Neural Information Processing Systems* **37**, 133282–133304 (2024)
35. Ma, X., Fang, G., Wang, X.: Deepcache: Accelerating diffusion models for free (2023)
36. Ma, Y., He, Y., Cun, X., Wang, X., Chen, S., Li, X., Chen, Q.: Follow your pose: Pose-guided text-to-video generation using pose-free videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 4117–4125 (2024)
37. Ma, Y., Liu, Y., Zhu, Q., Yang, A., Feng, K., Zhang, X., Li, Z., Han, S., Qi, C., Chen, Q.: Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207* (2025)
38. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073* (2021)
39. Millon, E.: Krea realtime 14b: Real-time video generation (2025), <https://github.com/krea-ai/realtime-video>
40. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
41. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023)
42. Peng, B., Quesnelle, J.: Ntk-aware scaled rope allows llama models to have extended (8k+) context size without any fine-tuning and minimal perplexity degradation (2023)
43. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis (2023), <https://arxiv.org/abs/2307.01952>
44. Qiu, H., Yu, N., Huang, Z., Debevec, P., Liu, Z.: Cinescale: Free lunch in high-resolution cinematic visual generation (2025)
45. Qiu, H., Zhang, S., Wei, Y., Chu, R., Yuan, H., Wang, X., Zhang, Y., Liu, Z.: Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16893–16903 (2025)
46. Ren, J., Li, W., Wang, Z., Sun, H., Liu, B., Chen, H., Xu, J., Li, A., Zhang, S., Shao, B., et al.: Turbo2k: Towards ultra-efficient and high-quality 2k video synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 18155–18165 (2025)
47. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
48. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
49. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International conference on machine learning. pp. 2256–2265. pmlr (2015)

50. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
51. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., Zeng, J., Wang, J., Zhang, J., Zhou, J., Wang, J., Chen, J., Zhu, K., Zhao, K., Yan, K., Huang, L., Feng, M., Zhang, N., Li, P., Wu, P., Chu, R., Feng, R., Zhang, S., Sun, S., Fang, T., Wang, T., Gui, T., Weng, T., Shen, T., Lin, W., Wang, W., Wang, W., Zhou, W., Wang, W., Shen, W., Yu, W., Shi, X., Huang, X., Xu, X., Kou, Y., Lv, Y., Li, Y., Liu, Y., Wang, Y., Zhang, Y., Huang, Y., Li, Y., Wu, Y., Liu, Y., Pan, Y., Zheng, Y., Hong, Y., Shi, Y., Feng, Y., Jiang, Z., Han, Z., Wu, Z.F., Liu, Z.: Wan: Open and advanced large-scale video generative models (2025)
52. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-esrgan: Training real-world blind super-resolution with pure synthetic data (2021), <https://arxiv.org/abs/2107.10833>
53. Wang, Z., Fang, G., Ma, X., Yang, X., Wang, X.: Sparsed: Sparse attention for diffusion language models. In: *International Conference on Learning Representations* (2026)
54. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., Yin, D., Zhang, Y., Wang, W., Cheng, Y., Xu, B., Gu, X., Dong, Y., Tang, J.: Cogvideox: Text-to-video diffusion models with an expert transformer (2025)
55. Ye, F., Zhao, Z., Mu, Y., Shen, J., Li, R., Wang, K., Sun, D., Agarwal, S., Lee, M., Cao, T., Akella, A., Krishnamurthy, A., Ng, T.S.E., Tu, Z., Wang, Y.: Supergen: An efficient ultra-high-resolution video generation system with sketching and tiling (2025)
56. Zhang, S., Chen, Z., Zhao, Z., Chen, Y., Tang, Y., Liang, J.: Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models (2024)
57. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution (2023), <https://arxiv.org/abs/2312.06640>
58. Zou, C., Liu, X., Liu, T., Huang, S., Zhang, L.: Accelerating diffusion transformers with token-wise feature caching (2025)