ViBe: Ultra-High-Resolution Video Synthesis Born from Pure Images

Yunfeng Wu*[1,2], Hongying Cheng*[1,3], Zihao He[1], and Songhua Liu†[1]

[1] School of Artificial Intelligence, Shanghai Jiao Tong University
[2] Xi'an Jiaotong-Liverpool University
[3] Jilin University

Fig. 1: Ultra-resolution results generated by our method built upon Wan2.2 [46]. 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. Transformer-based video diffusion models rely on 3D attention over spatial and temporal tokens, which incurs quadratic time and memory complexity and makes end-to-end training for ultra-high-resolution videos prohibitively expensive. To overcome this bottleneck, we propose a pure image adaptation framework that upgrades a video Diffusion Transformer pre-trained at its native scale to synthesize higher-resolution videos. Unfortunately, naively fine-tuning with high-resolution images alone often introduces noticeable noise due to the image–video modality gap. To address this, we decouple the learning objective to separately handle modality alignment and spatial extrapolation. At the core of our approach is *Relay LoRA*, a two-stage adaptation strategy. In the

* Equal contribution.
† Corresponding author (`liusonghua@sjtu.edu.cn`).

first stage, the video diffusion model is adapted to the image domain using low-resolution images to bridge the modality gap. In the second stage, the model is further adapted with high-resolution images to acquire spatial extrapolation capability. During inference, only the high-resolution adaptation is retained to preserve the video generation modality while enabling high-resolution video synthesis. To enhance fine-grained detail synthesis, we further propose a *High Frequency Awareness Training Objective*, which explicitly encourages the model to recover high-frequency components from degraded latent representations via a dedicated reconstruction loss. Extensive experiments demonstrate that our method produces ultra-high-resolution videos with rich visual details without requiring any video training data, even outperforming previous state-of-the-art models trained on high-resolution videos by 0.8 on the VBench benchmark. Code is available here.

1 Introduction

Empowered by the attention mechanism's ability to capture complex token-wise dependencies [45], Diffusion Transformers (DiTs) [37] have shown remarkable performance in generating high-quality images [5–7, 11, 12, 16, 29, 48] and videos [6, 7, 13, 19, 22, 25, 32, 50]. 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.

We report how VRAM usage and iteration time scale with the number of frames under different resolutions to provide an intuitive comparison between training with high-resolution images and videos. As the resolution increases, VRAM consumption rises sharply, as shown in Fig. 2(a); as the number of frames grows, the iteration time increases approximately linearly, as shown in Fig. 2(b). 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 video generators pre-trained only at lower resolution, such as 480P, by training solely with image data?*

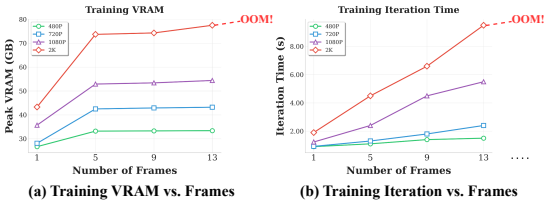
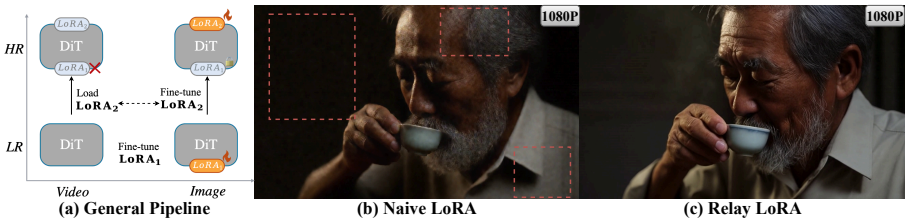


Fig. 2: The two plots show how training VRAM consumption and iteration time scale with the number of frames under different resolutions.



(a) General Pipeline (b) Naive LoRA (c) Relay LoRA

Fig. 3: The plots on the left shows the general pipeline of our Relay LoRA. The two panels on the right provide a qualitative analysis on Wan2.2 [46]: naive high-resolution image fine-tuning (Naive LoRA) introduces noticeable artifacts, whereas ours methods (Relay LoRA) effectively removes them.

However, this is a highly non-trivial problem–direct training on images while conducting video inference can yield noticeable noise, as shown in Fig. 3(b) and (c), which we attribute to the gap between the image and video modalities. This observation highlights the need for a principled approach that can bridge the modality gap while enabling spatial extrapolation from image data.

In this paper, we reveal that the core to alleviate this drawback lies decoupling the learning objective to retain only the benefits of spatial extrapolation while mitigating the adverse effects caused by the modality gap between images and videos. We thus propose a two-stage adaptation strategy termed *Relay LoRA*, which explicitly separate the two adaptation processes: modality alignment and spatial extrapolation, as shown in Fig. 3(a). Specifically, in Stage 1, we fine-tune a base video DiT on *low-resolution images* with LoRA [20] and obtain LoRA_1, which familiarizes the video model with the image modality. Building upon LoRA_1, Stage 2 fine-tunes the model on *high-resolution images* to obtain another LoRA, termed LoRA_2, which equips the model with spatial extrapolation capability. Since inference is conducted for *high-resolution videos* rather than images, only LoRA_2—the spatial extrapolation adapter—is attached to the base DiT, while LoRA_1—the video-to-image adapter—is discarded.

Moreover, to enhance fine-grained detail rendering, we introduce a *High Frequency Awareness Training Objective (HFATO)*. HFATO explicitly encourages the model to recover high-frequency components from degraded latent representations, thereby improving its ability to reconstruct subtle textures and sharp structural patterns. To better exploit the degraded latents during training, we couple the latent degradation module with a dedicated reconstruction loss. This additional supervision regularizes the optimization toward high-frequency preservation. Empirically, we observe that this design further improves fine-grained detail fidelity and structural clarity, particularly in challenging high-resolution scenarios.

We conduct extensive experiments on modern DiT-based video generation models, *e.g.*, Wan2.2 [46]. As shown in Fig. 1, our method synthesizes 4K-resolution videos with strong semantic coherence and visually compelling fine-

grained details. Quantitatively, it achieves state-of-the-art results on VBench [23], even surpassing prior training-based methods that are trained on real high-resolution video data in both generation quality and consistency. Our contributions are summarized as follows:

- Conceptually, we delve into the field of ultra-high-resolution video generation and, to the best of our knowledge, present the first method tailored to DiT architectures such as Wan2.2 [46] that achieves ultra-high-resolution video generation using pure image data for training.
- Technically, we propose a Relay LoRA strategy that mitigates the noise induced by image-only training, augmented with a high frequency awareness training objective to strengthen fine-detail processing.
- Experimentally, extensive experiments and user evaluations confirm that our method achieves superior performance in ultra-high-resolution video generation, outperforming training-based counterparts.

2 Related Work

2.1 Diffusion Models for Video Generation

Diffusion models (DMs) [17, 42] have been widely studied for video generation and have attracted substantial attention. In contrast to conventional UNet-based video diffusion models [2, 4, 15, 18], which rely on convolutional backbones, Diffusion Transformer (DiT) [37] uses a Transformer architecture as the primary denoising backbone. This architectural shift allows DiT to model long-range dependencies more effectively and to represent more complex spatiotemporal relationships in video sequences. Early studies [24, 33, 50] demonstrated the effectiveness of spatiotemporal Transformers that employ self-attention with a global receptive field. More recently, LTX-Video [13] improved the coupling between the Video-VAE and the denoising Transformer, and Wan2.1 [46] and Hunyuan [25], both trained on large-scale video datasets, achieved strong performance in realistic video generation. Despite the substantial gains in generation quality, the native resolution of these models remains limited for high-quality applications.

2.2 High-Resolution Visual Generation

In the image domain, training-free high-resolution visual generation has been widely studied. Most existing approaches [38] are built on U-Net [41] and often suffer from repetitive patterns in high-resolution synthesis due to limited local receptive fields. Several works [1, 9, 14] enlarge receptive fields through dilated convolutions and fuse local and global patches to mitigate repetition. With the emergence of recent foundational diffusion models [26], DiT has become the dominant architecture, benefiting from the capacity of attention mechanisms to model complex token-wise dependencies. Training-free methods [3, 10] built on DiT can preserve global layout consistency while synthesizing fine-grained visual details.

In the video domain, high-resolution synthesis with DiT-based models introduces additional challenges, including substantial computational overhead, blur, and structural distortion. Some works [51] focus on the system infrastructure required for efficient high-resolution video generation, but the generation quality remains constrained. Most current video generation methods [21, 39, 40] rely on fine-tuning with high-resolution data, which imposes prohibitive computational demands. In this work, we propose a fully training-based approach that uses pure image data tailored for modern DiT architectures, while ensuring global layout consistency and producing fine-grained details in high-resolution synthesis.

3 Methods

Generating 4K videos is challenging because it requires both fine-grained detail synthesis and coherent high-level semantic structure under a tight compute budget. Following prior work [34], we adopt a coarse-to-fine pipeline: we first generate a video at the model's native resolution from the text prompt to establish global layout, object identities, and motion semantics. We then produce a high resolution version based on the initial output, so that the second stage focuses on recovering high-frequency details while preserving the coarse structure.

We implement this pipeline within a flow-matching framework that specifies both the training objective and the sampling pipelines (Sec. 3.1). On top of this backbone, we introduce three key components. (i) **Relay LoRA** (Sec. 3.2) decomposes image-based fine-tuning into two stages to mitigate modality-induced artifacts. (ii) **Global Coarse Local Fine Attention (GCLFA)** (Sec. 3.3) balances global semantic consistency and local detail modeling, enabling effective long-range interactions at high resolution. (iii) **High Frequency Awareness Training Objective (HFATO)** (Sec. 3.4) enhances high-frequency detail synthesis by explicitly training the model to reconstruct clean latents from degraded latents. Together, these designs improve 4K detail fidelity while maintaining coherent semantics and fine-grained details. An overview of the method is shown in Fig. 4.

3.1 Flow Matching in Video Generation

Flow matching [28] provides a convenient parameterization for diffusion-based video generation by learning a time-dependent velocity field whose integral curve transports a simple base distribution to the data distribution. Given a clean sample $x_0 \sim p_{\text{data}}$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, flow matching constructs a continuous interpolation

$$x_t \coloneqq \alpha_t x_0 + \sigma_t \epsilon, \quad t \in [0, 1], \tag{1}$$

which smoothly bridges the data at $t = 0$ and a noise distribution at $t = 1$. The model v_θ is trained to approximate the ground-truth velocity along this trajectory by minimizing the weighted mean squared error

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0,\epsilon,t}\Big[w(t)\,\|v_\theta(x_t, t) - v_t\|^2\Big], \tag{2}$$

where $t \sim \mathcal{U}(0,1)$ and $w(t)$ is a time-dependent reweighting that balances learning across different noise levels. Under the linear noise schedule [30] defined by $\alpha_t = 1 - t$, $\sigma_t = t$, and $T = 1$, we have

$$\frac{\mathrm{d}x_t}{\mathrm{d}t} = \dot{\alpha}_t x_0 + \dot{\sigma}_t \epsilon = (-1)\, x_0 + (1)\, \epsilon, \tag{3}$$

thus the target velocity satisfies

$$v_t \coloneqq \frac{\mathrm{d}x_t}{\mathrm{d}t} = \epsilon - x_0.$$

After training, sampling is performed by integrating the probability flow ordinary differential equation (PF-ODE) [43],

$$\mathrm{d}x_t = v_\theta(x_t, t)\, \mathrm{d}t, \quad x_T \sim \mathcal{N}(0, I), \quad t : T \to 0, \tag{4}$$

which deterministically maps a noise sample to a data sample. In practice, (4) is solved with a numerical ODE solver (e.g., Euler or Heun) using a small number of discretization steps.

3.2 Relay LoRA

Nevertheless, in video generation, a model must capture complex 3D interactions across both spatial and temporal dimensions, whereas image generation involves only spatial structure. As shown in Fig. 3(b) and (c), naively fine-tuning the model on high-resolution images and directly using the resulting LoRA [20] for video inference introduces noticeable artifacts due to the modality gap between images and videos, which can severely degrade visual quality.

To address this problem, motivated by this observation, we speculate that introducing an auxiliary adaptation mechanism to mitigate the modality gap before fine-tuning for new capabilities, i.e., high-resolution generation, can effectively address this issue. To this end, the naive high-resolution image fine-tuning procedure is decomposed into two stages, denoted as Relay LoRA. In the first stage, a LoRA [20] module is trained on low-resolution images used for video pretraining, adapting DiTs to single low-resolution frame generation, denoted as LoRA_1. The learned LoRA_1 is merged with the base model to produce a new set of weights,

$$W^{(1)} = W^{(0)} + \Delta W_{\mathrm{LoRA}_1}, \qquad \Delta W_{\mathrm{LoRA}_1} = \frac{\alpha}{r} B_1 A_1, \tag{5}$$

where $W^{(0)}$ denotes the frozen base-model weights and (A_1, B_1) are the low-rank factors learned in the first stage, r is the LoRA rank, and α is a scaling hyperparameter, which may still introduce image-induced artifacts. In the second stage, high-resolution images are used to train a second LoRA [20] module, denoted as LoRA_2, on top of the merged weights $W^{(1)}$, adapting DiTs to single high-resolution frame generation. This two-stage design effectively mitigates modality-induced artifacts. During inference, we load only LoRA_2 into the base

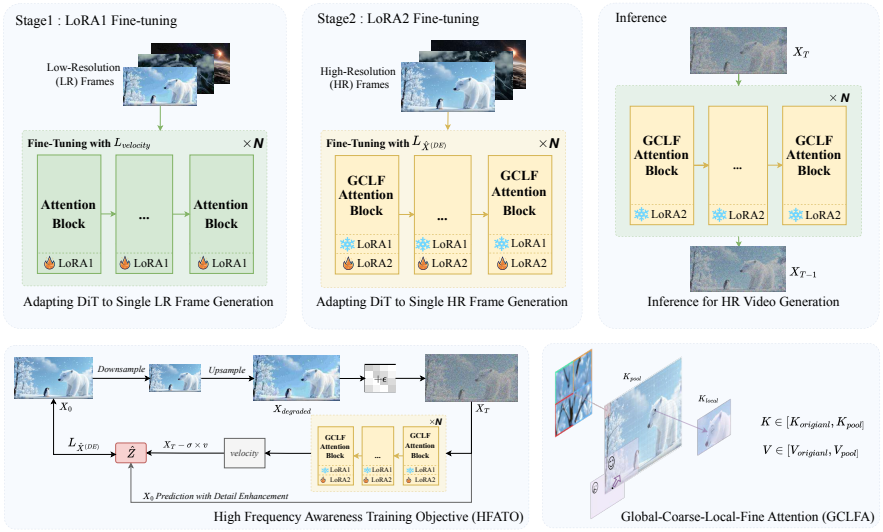


Fig. 4: Upper Left: Stage-1 fine-tuning trains LoRA_1 to adapt the DiT backbone to single low-resolution frame generation. **Upper Middle**: Stage-2 first merges LoRA_1 into the base model and freezes the merged weights, then trains a Relay LoRA_2 to adapt the DiT backbone to single high-resolution frame generation. **Upper Right**: During inference, only LoRA_2 is loaded on the base model for video generation. **Bottom Left**: Global Coarse Local Fine Attention combines an inward sliding-window local attention with pooled coarse attention. **Bottom Right**: The training objective first degrades the latents via a downsample–upsample operation, then add noise on the degraded sample. The model applies the predicted flow and computes the loss against the clean latents.

model, which is decoupled and learns only the ability to handle spatial high-resolution frame generation. Concretely, for each LoRA-injected weight matrix, we apply

$$W_{\text{infer}} = W^{(0)} + \Delta W_{\text{LoRA}_2}, \qquad \Delta W_{\text{LoRA}_2} = \frac{\alpha}{r} B_2 A_2, \tag{6}$$

where $W^{(0)}$ denotes the base-model weights and (A_2, B_2) are the low-rank factors learned in the second stage, r is the LoRA rank, and α is a scaling hyperparameter. We notice that [31] adopts a similar high-level spirit in the NLP field.

3.3 Global Coarse Local Fine Attention

Based on our observation, we can design any customized attention variants in the second stage for visual enhancement. Inspired by FreeSwim [49], which demonstrates that local attention significantly improves fine-grained details, we introduce a *Global Coarse Local Fine Attention (GCLFA)* in Stage 2, as illustrated in Fig. 4 (Bottom Left). The local branch enhances high-frequency details, while the global branch preserves semantic coherence.

Local Fine Branch. In the fine local branch, naive sliding-window attention truncates the attention window for boundary tokens due to limited spatial context, resulting in an inconsistent number of interactable key/value tokens across queries. To ensure a uniform receptive field for all tokens, including those near boundaries, we propose an inward sliding-window attention mechanism. Specifically, when a query's attention window extends beyond the video boundary, the window is shifted inward according to the token's spatial location (e.g., using a 480P window for a 480P model). This design preserves the receptive-field size used during training and eliminates boundary truncation, which can be formulated as:

$$\begin{aligned} \Delta_w^{(q)} &= \max\Big(\frac{w}{2} - x_q, \frac{w}{2} + x_q - W + 1, 0\Big), \\ \Delta_h^{(q)} &= \max\Big(\frac{h}{2} - y_q, \frac{h}{2} + y_q - H + 1, 0\Big), \\ M_{qk} &= \begin{cases} 1, & \text{if } |x_q - x_k| \le \frac{w}{2} + \Delta_w^{(q)} \\ & \quad \text{and } |y_q - y_k| \le \frac{h}{2} + \Delta_h^{(q)}, \\ 0, & \text{otherwise,} \end{cases} \end{aligned} \tag{7}$$

Global Coarse Branch. For the global branch, we first apply RoPE [44] to Q and K to inject positional information. The key and value are then pooled to produce downsampled coarse tokens:

$$\bar{K} = \mathrm{Pool}(K), \quad \bar{V} = \mathrm{Pool}(V). \tag{8}$$

These coarse tokens provide a compact global representation aligned with the global branch. We concatenate them with the original key–value sequence:

$$\tilde{K} = \mathrm{Concat}(K, \bar{K}), \quad \tilde{V} = \mathrm{Concat}(V, \bar{V}), \tag{9}$$

so that each query attends to both fine local tokens and coarse global tokens:

$$O = \mathrm{Softmax}\left(\frac{Q\tilde{K}^\top \cdot M}{\sqrt{d}}\right)\tilde{V}. \tag{10}$$

Here, x_* and y_* denote spatial token coordinates; W and H are the target spatial dimensions, while w and h are the native window sizes; Q, K, V, and M denote the query, key, value, and attention mask matrices; d is the feature dimension; and O is the attention output. We replace the original 3D full attention with this scheme in all self-attention layers, which serve as the primary token interaction modules in modern DiT-based video generators such as Wan2.2 [46].

3.4 High Frequency Awareness Training Objective

To enhance high-frequency detail synthesis, we propose a *High Frequency Awarenss Training Objective (HFATO)* that explicitly enforces clean-latent reconstruction from degraded noisy inputs. Unlike standard flow-matching training [26], which

perturbs the clean latent with noise and learns to predict the flow at timestep t, we first apply a downsample–upsample degradation operator to corrupt high-frequency components before injecting Gaussian noise.

Under this formulation, the model predicts the flow from latents that are first degraded and then corrupted by noise, rather than from clean latents directly perturbed by noise. This shift in training dynamics encourages the model to operate under high-frequency-deficient conditions and explicitly learn detail restoration. To promote high-frequency generation, we introduce a dedicated reconstruction objective that supervises the recovery of the original clean latent. Specifically, the model estimates the flow on degraded and noisy latents and reconstructs the clean latent accordingly. We minimize a reconstruction loss to enforce consistency between the restored latent and the ground-truth clean latent. Empirically, we find that directly using the clean latent as the supervision target yields better high-frequency recovery. The overall training procedure can be summarized as:

$$\begin{aligned} &\tilde{x}_0 = \mathrm{DU}(x_0), \qquad x_t = \tilde{x}_0 + \sigma_t \epsilon,\ \epsilon \sim \mathcal{N}(0, I), \\ &\hat{x}_0 = x_t - \sigma_t v_\theta(x_t, t), \qquad \mathcal{L} = \left\| \hat{x}_0 - x_0 \right\|^2 . \end{aligned} \tag{11}$$

Here, x_0 denotes the clean latent, $\mathrm{DU}(\cdot)$ is the downsample–upsample degradation operator, $\tilde{x}_0$ is the degraded latent, ϵ is standard Gaussian noise, σ_t is the noise level at timestep t, $v_\theta(\cdot, \cdot)$ is the flow predictor, $\hat{x}_0$ is the reconstructed clean latent, and $\mathcal{L}$ is the training loss.

4 Experiments

4.1 Settings and Implementation Details

We conduct experiments on Wan2.2 [46], a state-of-the-art DiT-based video generation model. Wan2.2 includes a 5B variant trained only at 1280×704 resolution with a VAE downsampling factor of $16 \times 16 \times 4$, and a 14B variant trained on a mixed-resolution dataset of 832×480 and 1280×720 with a VAE downsampling factor of $8 \times 8 \times 4$. Additional results on other DiT models are provided in the Appendix. We implement the proposed GCLFA using FlexAttention [8] in PyTorch [36], leveraging efficient low-level optimizations for sparse attention. We fine-tune the parameters in the attention layers on 2.3K samples at 2752×1536 resolution, generated by FLUX 1.1 Pro Ultra [26], for 3K iterations in Stage 1 with the standard flow matching loss, and for another 3K iterations in Stage 2 using the loss defined in Eq. 11. Other hyper-parameters, including LoRA rank, schedulers and optimizers, follow the default settings of DiffSynth-Studio [35]. Two stage Training is conducted on a single A100 GPU and finishes in about one day. Unless otherwise specified, all inference is performed on a single A100 GPU.

The method is evaluated with VBench [23], which assesses both visual quality and semantic coherence. For 1920×1088 (1080P) video and 3820×2176 (4K) generation, 100 prompts are randomly selected from the standard prompt suite

of VBench [23], and the official evaluation protocol is followed. Each method generates five videos per prompt using five different random seeds. The official VBench metrics are then applied to ensure a fair comparison across all methods.

4.2 Main Comparisons

We evaluate our method on Wan2.2 [46] across both 1080P and 4K resolutions. We consider the following baselines: (1) **Low-Level Super Resolution Methods**, including Real-ESRGAN [47] and Upscale-A-Video [54], whose base videos are derived from Wan2.2 [46]; (2) **Training-Free High-Resolution Generation Methods**, including I-Max [10], HiFlow [3]; and (3) **Training-Based High-Resolution Generation Methods**, including CineScale [39] and T3-Video [52], both of which adapt models using 1080P and 4K real video data.

Quantitative comparison. As shown in Tabs. 1, our method achieves competitive performance across both 1080P and 4K 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.

Qualitative comparison. To visually demonstrate the superiority of our approach, as illustrated in Fig. 5, we compare our method with the baselines [10, 39, 46, 47, 52, 54] designed for high-resolution visual generation on Wan2.2-5B [46]. 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. Although Real-ESRGAN [47] and Upscale-A-Video [54], two video super-resolution methods, enhances video clarity and quality to some extent, it struggles to synthesize finer details. While I-Max [10] and HiFlow [3] are capable of preserving the overall video structure in a training-free paradigm, they often synthesize low-fidelity details, leading to overly limited detail creation. In comparison, the proposed methods consistently yields aesthetically pleasing and semantically coherent outcomes. Furthermore, our methods delivers high-resolution images of exceptional quality, outperforming leading training-based models such as CineScale [39] and T3-Video [52], underscoring its outstanding generative capabilities. In particular, in the fifth 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.

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 Wan2.2-5B [46]: the

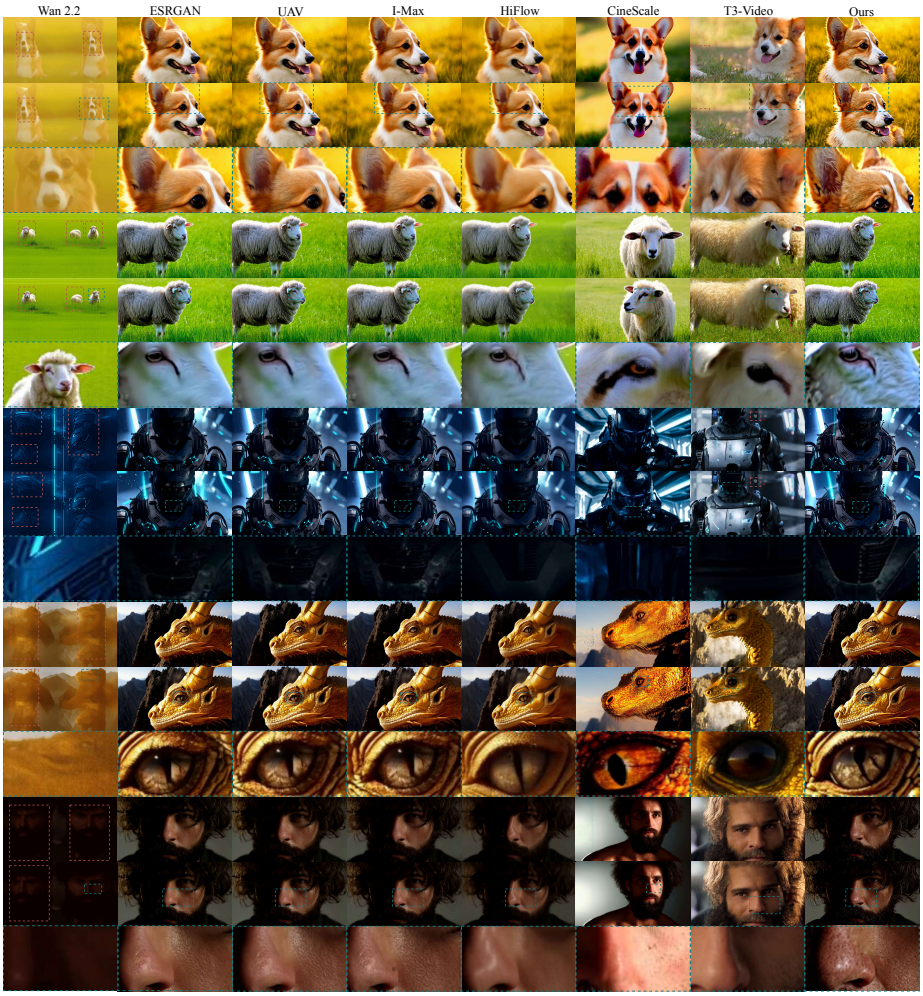


Fig. 5: Qualitative comparison. ViBe yields high-resolution videos characterized by high-fidelity details and coherent structure. The red boxes highlight regions with incorrect semantics or layout. The blue boxes provide zoomed-in views.

training-based high resolution approach CineScale [39] and T3-Video [52] and our proposed method. All methods are evaluated using the same prompts from VBench [23]. 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

Resolution (W × H)	Model	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Overall Consistency	Overall Score
1920 × 1088	Wan2.2 [46]	94.9%	95.4%	98.5%	59.1%	64.0%	23.3%	72.5%
	ESRGAN [47]	95.3%	93.8%	96.9%	60.4%	63.6%	25.7%	72.6%
	UAV [54]	95.2%	93.2%	96.8%	58.7%	62.5%	25.2%	71.9%
	I-Max [10]	**95.7%**	95.3%	97.7%	60.2%	62.0%	24.0%	72.5%
	HiFlow [3]	95.5%	95.6%	97.4%	60.1%	59.6%	25.4%	72.3%
	CineScale [39]	93.9%	96.8%	97.4%	59.1%	60.7%	23.5%	71.9%
	T3-Video [52]	94.2%	91.8%	98.6%	37.8%	35.8%	20.8%	63.2%
	Ours	95.5%	**97.5%**	**98.9%**	**61.4%**	**65.7%**	**26.7%**	**74.3%**
3840 × 2176	Wan2.2 [46]	94.7%	94.8%	97.1%	59.1%	33.9%	13.6%	65.5%
	ESRGAN [47]	95.1%	**97.1%**	97.3%	59.8%	58.3%	24.3%	72.0%
	UAV [54]	94.9%	96.4%	97.3%	61.1%	65.7%	25.8%	73.5%
	I-Max [10]	94.7%	95.7%	98.9%	57.0%	61.8%	26.9%	72.5%
	HiFlow [3]	95.0%	96.7%	99.1%	56.6%	54.4%	27.0%	71.5%
	CineScale [39]	94.8%	97.0%	98.2%	60.1%	**66.3%**	25.1%	73.6%
	T3-Video [52]	92.8%	95.4%	98.1%	60.8%	64.8%	24.7%	72.8%
	Ours	**95.2%**	**97.1%**	**99.2%**	**61.4%**	66.1%	**27.1%**	**74.4%**

Table 1: Video comparison with both training-based and training-free models at 1080P and 4K resolutions. The highest value is **bold**, and the second-highest is underlined. Overall Score is computed as the mean of the six metrics in each row.

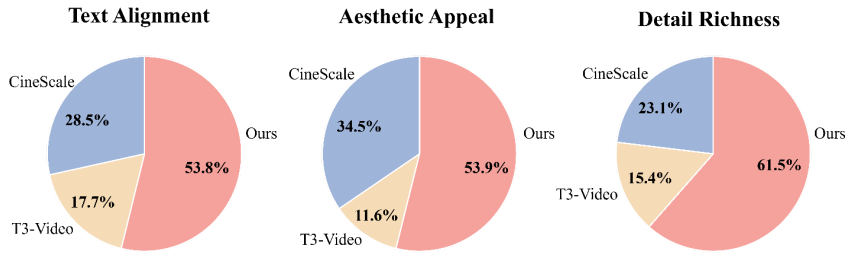


Fig. 6: Results of user study for high-resolution video generation. Participants were expected to select the best method based on Text Alignment, Aesthetic Quality, and Video Quality.

provided prompts. In total, 35 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

We conduct controlled ablations to quantify the contribution of each component in our framework, including (i) Relay LoRA, (ii) Global Coarse-Local Fine Attention (GCLFA), and (iii) High Frequency Awareness Training Objective

Fig. 7: Qualitative ablations. We perform controlled comparisons of our method with alternative variants under controlled settings. Best viewed zoomed in. The red boxes highlight regions with more noticeable noise, and the blue boxes indicate the corresponding zoomed-in views. Here, *w/o* Relay LoRA denotes the variant that equips the base model with only LoRA_1, without the relay design.

(HFATO). Quantitative results are reported in Tab. 2, and qualitative comparisons are shown in Fig. 7. we draw the following conclusions regarding the contribution of each component:

Relay LoRA. Relay LoRA is crucial for suppressing the noise induced by naïve high-resolution image fine-tuning. Compared to direct high-resolution LoRA adaptation, Relay LoRA yields markedly cleaner and more temporally stable videos (Fig. 7(a)→(b)). Without Relay LoRA, the model either fails to maintain coherent content at the target resolution or suffers from severe instability, establishing Relay LoRA as the foundation for reliable ultra-high-resolution generation.

Global Coarse Local Fine Attention (GCLFA). GCLFA further improves fine-grained details while preserving global structure (Fig. 7(b)→(c)). Removing GCLFA consistently degrades visual richness and semantic integrity,

Methods	Subject Consistency	Background Consistency	Motion Smoothness	Aesthetic Quality	Imaging Quality	Overall Consistency	Overall Score
w/o Relay LoRA	94.6%	**97.9%**	98.3%	46.7%	37.4%	21.1%	66.0%
w/o GCLFA	94.9%	95.4%	99.0%	52.2%	49.0%	24.3%	69.1%
w/o HFATO	95.0%	94.2%	99.1%	53.7%	59.0%	24.0%	70.8%
w/o X_0 Loss	**96.3%**	95.3%	<u>99.2%</u>	57.3%	59.2%	<u>26.7%</u>	72.3%
Ours (DR = 1/2)	<u>95.2%</u>	97.1%	<u>99.2%</u>	**61.4%**	**66.1%**	**27.1%**	**74.4%**
Ours (DR = 1/4)	95.0%	<u>97.2%</u>	**99.3%**	<u>61.3%</u>	<u>65.6%</u>	26.5%	<u>74.2%</u>

Table 2: Quantitative ablations. Ours Methods achieves the optimal overall video quality. The highest value is **bold**, and the second-highest is <u>underlined</u>.

indicating that coarse global context combined with local fine attention is essential for high-resolution detail synthesis.

High Frequency Awareness Training Objective (HFATO). HFATO consists of two decoupled factors: latent degradation and x_0-reconstruction supervision. When we apply only the downsample–upsample degradation without the x_0 loss, the model gains limited detail enhancement (Fig. 7(c)→(d)). Adding the x_0 reconstruction loss substantially boosts high-frequency details and improves robustness (Fig. 7(d)→(e)), suggesting that explicitly supervising the recovered clean latent is key to stable detail amplification.

Effect of the Downsample Ratio (DR). In GCLFA, we investigate two downsample ratios: 1/2 and 1/4. As shown in Tab. 2, both ratios are competitive in terms of aesthetic and imaging quality. Additionally, both ratios generate coherent content (Fig. 7(e)→(f)), with the 1/4 downsample ratio further enhancing fine details, such as wrinkles on the elderly's skin, especially when zoomed in. Consequently, we recommend a downsample ratio of 1/2 for more robust results, and 1/4 for results with greater detail.

4.5 Empirical Study

To further validate the generalization of our framework, we conduct an empirical study across different scenarios. As illustrated in Fig. 8(a), experiments show that our LoRA_2 can be directly integrated into few steps distilled models [27] to significantly accelerate inference speed while maintaining high-fidelity visual details with negligible degradation. Furthermore, as illustrated in Fig. 8(b), experiments show that our LoRA_2 can be directly integrated into Image-to-Video (I2V) generation tasks to upgrading existing I2V models to synthesize ultra-high-resolution content.

5 Conclusion

In this paper, we introduce ViBe, a novel image-based training paradigm designed for ultra-high-resolution video generation. By introducing *Relay LoRA*, ViBe effectively addresses challenges in naively fine-tuning pre-trained text-to-video diffusion transformers, which are originally trained only on low-resolution

(a) 4 Steps Generation (b) Image-to-Video Generation

Fig. 8: Left: Illustration 4 steps generation results integrated with our framework. **Right**: Illustration Image-to-Video generation results integrated with our framework.

scales. Our method mitigates issues such as residual noise when using high-resolution images. Additionally, ViBe incorporates *Global Coarse Local Fine Attention*, which enhances visual details while preserving semantic consistency. Furthermore, we propose the *High Frequency Awareness Training Objective* to improve the model's capability to handle high-frequency details. Experimental results demonstrate that ViBe outperforms existing methods in ultra-high-resolution video generation, achieving superior 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. 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)
15. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation (2023)
16. 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=ZFOA8DZqhl
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
18. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models (2022)
19. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: Cogvideo: Large-scale pretraining for text-to-video generation via transformers (2022)
20. 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
21. Hu, T., Zhang, J., Su, Z., Yi, R.: Ultragen: High-resolution video generation with hierarchical attention (2025), https://arxiv.org/abs/2510.18775
22. Hu, Y., Gong, H., Yang, C., An, Z., Xu, Y., Liu, S.: Multianimate: Pose-guided image animation made extensible (2026)
23. 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
24. 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)
25. 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)
26. 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
27. Li, Q., Xing, Z., Wang, R., Zhang, H., Dai, Q., Wu, Z.: Magicmotion: Controllable video generation with dense-to-sparse trajectory guidance. arXiv preprint arXiv:2503.16421 (2025)
28. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling (2023), https://arxiv.org/abs/2210.02747
29. 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)
30. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow (2022), https://arxiv.org/abs/2209.03003
31. Liu, Y., Chang, S., Jaakkola, T., Zhang, Y.: Fictitious synthetic data can improve llm factuality via prerequisite learning (2024), https://arxiv.org/abs/2410.19290
32. 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)

33. 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)
34. 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)
35. ModelScope: Diffsynth-studio (2026), https://github.com/modelscope/DiffSynth-Studio, accessed: 2026-03-02
36. 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)
37. Peebles, W., Xie, S.: Scalable diffusion models with transformers (2023)
38. 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
39. Qiu, H., Yu, N., Huang, Z., Debevec, P., Liu, Z.: Cinescale: Free lunch in high-resolution cinematic visual generation (2025)
40. 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)
41. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation (2015)
42. 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)
43. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations (2021), https://arxiv.org/abs/2011.13456
44. Su, J., Lu, Y., Pan, S., Murtadha, A., Wen, B., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding (2023), https://arxiv.org/abs/2104.09864
45. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
46. 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)
47. 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
48. Wang, Z., Fang, G., Ma, X., Yang, X., Wang, X.: Sparsed: Sparse attention for diffusion language models. In: International Conference on Learning Representations (2026)

49. Wu, Y., Song, J., Tan, Z., He, Z., Liu, S.: Freeswim: Revisiting sliding-window attention mechanisms for training-free ultra-high-resolution video generation (2025), https://arxiv.org/abs/2511.14712
50. 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)
51. 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)
52. Zhang, J., Zhu, J., Hu, T., Wang, Y., Luo, D., Cao, W., Gan, Z., Hu, X., Xue, Z., Wang, C.: Transform trained transformer: Accelerating naive 4k video generation over 10× (2025), https://arxiv.org/abs/2512.13492
53. Zhang, S., Chen, Z., Zhao, Z., Chen, Y., Tang, Y., Liang, J.: Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models (2024)
54. 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