

UltraGen: High-Resolution Video Generation with Hierarchical Attention

Teng Hu^{1*}, Jiangning Zhang^{2*}, Zihan Su¹, Ran Yi^{1†}

¹Shanghai Jiao Tong University

²Zhejiang University

{hu-teng, ranyi}@sjtu.edu.cn, 186368@zju.edu.cn, szhzephyr@gmail.com

Abstract

Recent advances in video generation have made it possible to produce visually compelling videos, with wide-ranging applications in content creation, entertainment, and virtual reality. However, most existing diffusion transformer based video generation models are limited to low-resolution outputs ($\leq 720P$) due to the quadratic computational complexity of the attention mechanism with respect to the output width and height. This computational bottleneck makes native high-resolution video generation (1080P/2K/4K) impractical for both training and inference. To address this challenge, we present **UltraGen**, a novel video generation framework that enables *i) efficient* and *ii) end-to-end native high-resolution* video synthesis. Specifically, UltraGen features a hierarchical dual-branch attention architecture based on global-local attention decomposition, which decouples full attention into a local attention branch for high-fidelity regional content and a global attention branch for overall semantic consistency. We further propose a spatially compressed global modeling strategy to efficiently learn global dependencies, and a hierarchical cross-window local attention mechanism to reduce computational costs while enhancing information flow across different local windows. Extensive experiments demonstrate that UltraGen can effectively scale pre-trained low-resolution video models to 1080P and even 4K resolution for the first time, outperforming existing state-of-the-art methods and super-resolution based two-stage pipelines in both qualitative and quantitative evaluations.

Code — <https://sjtuplayer.github.io/projects/UltraGen>

1 Introduction

The field of video generation (Huang et al. 2024b; Villegas et al. 2022; Xue et al. 2025; Hu et al. 2025a) has undergone rapid development in recent years, unlocking a diverse array of downstream applications, including video customization (Hu et al. 2025c,b; Liu et al. 2025), editing (Liang et al. 2025; Chen et al. 2025; Yi et al. 2024), and motion control (Hu 2024; Hu et al. 2024, 2025e). With the emergence of powerful diffusion-based generative models (Ho,

*These authors contributed equally.

†Corresponding author.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.



Figure 1: Typical video generation models exhibit significant 🙄 *quality degradation* and 🕒 *increased processing time* with higher resolutions, whereas our UltraGen delivers 😊 *superior video quality* at resolutions beyond 2K while achieving 🚀 *4.78× speedup* compared to the popular Wan-T2V-1.3B baseline (Wang et al. 2025a) (81 frames, $4\times H20$ GPUs). Enlarge for better visual effects.

Jain, and Abbeel 2020), the quality, coherence, and diversity of generated videos have significantly improved, narrowing the gap between synthetic and real-world content. Based on diffusion transformers (Peebles and Xie 2023), state-of-the-art models such as Wan (Wang et al. 2025a) and HunyuanVideo (Kong et al. 2024) have demonstrated impressive capabilities in synthesizing temporally consistent and semantically rich videos, making remarkable progress in high-quality video generation.

Despite these advancements, current video generation models still suffer from a critical limitation: restricted resolution. Since the advanced video generation models (Wang et al. 2025a; Kong et al. 2024) are based on diffusion transformers (Peebles and Xie 2023), they inherently suffer from the quadratic computational complexity of the full-attention mechanism with respect to the spatiotemporal size of the input, *i.e.*, $\mathcal{O}((T \cdot H \cdot W)^2)$, where T , H , and W denote the temporal length, height, and width of the video, respectively. For instance, doubling the width and height will result in a 16-fold increase in computational cost, making high-resolution video generation prohibitively expensive for both training and inference. To mitigate this, existing

approaches (He et al. 2022; Blattmann et al. 2023b) often resort to a two-stage pipeline that first generates low-resolution videos and subsequently applies video super-resolution models. However, this paradigm merely enhances visual clarity and fails to introduce enough visual details, leading to the synthesis of pseudo high-resolution content with limited authenticity and richness.

To address these challenges, we propose **UltraGen**, a hierarchical attention-based framework for native high-resolution video generation. *UltraGen* offers an efficient and scalable solution that transforms pre-trained low-resolution video diffusion models into end-to-end high-resolution generators with significantly reduced computational overhead. Concretely, we propose a dual-branch video generation architecture that decouples the full attention mechanism into *local* and *global attention* branches. The local attention branch focuses on generating fine-grained content within individual local spatial windows, while the global attention branch captures holistic video semantics and ensures coherence across different local windows. To efficiently model global dependencies without incurring prohibitive costs, we design a **spatially compressed global modeling module** that compresses spatial information via frame-wise convolutions before applying attention, so that the self-attention is conducted at a smaller spatial size, followed by 3D convolutions to restore spatial fidelity and enhance temporal continuity. Furthermore, to ensure effective information flow across different local windows, we propose a **hierarchical cross-window local attention mechanism**. By partitioning the local windows of adjacent layers differently and creating intersections between them, our model enables seamless interaction and consistency across spatial local windows, further improving the video generation quality.

We conduct extensive experiments by extending the Wan-1.3B model to support native 1080P and 4K video generation, which is the first to achieve native high-quality 4K synthesis in the field. Comparisons against state-of-the-art models, including Wan and Hunyuan Video, as well as two-stage pipelines (low-resolution generation + super resolution), demonstrate that *UltraGen* significantly outperforms existing methods both qualitatively and quantitatively, validating the effectiveness and scalability of our approach.

- We propose **UltraGen**, a novel high-resolution video generation framework based on global-local attention decomposition, which enables scalable extension of low-resolution pre-trained video diffusion models to support 1080P and 4K resolution in an end-to-end manner.
- We design a **Spatially Compressed Global Attention Mechanism** that significantly reduces computation cost of global context modeling. By compressing spatial information via frame-wise convolution, conducting self-attention at a smaller spatial size, and decoding through 3D convolution, our method efficiently captures holistic semantics while keeping temporal coherence.
- We introduce a **Hierarchical Cross-window Local Attention Mechanism** that facilitates efficient interaction among local regions. By allowing intersecting regions between attention windows of adjacent layers, it ensures

smooth content transitions and enhances local detail.

- **UltraGen** is the first model to achieve **native high-quality 4K video generation**. Extensive experiments demonstrate its superior ability in HD video generation.

2 Related Work

2.1 Video Generation Foundation Models

The advent of diffusion models (Ho, Jain, and Abbeel 2020) has greatly advanced video generation. Early methods (Guo et al. 2023; Blattmann et al. 2023a) typically extend text-to-image diffusion models (Rombach et al. 2022) by adding temporal modules to capture frame dynamics. While somewhat effective, these approaches often separate spatial and temporal modeling, limiting their ability to capture holistic spatiotemporal dependencies and resulting in less coherent videos. With DiT (Labs 2024), transformer-based architectures have become the leading paradigm in video generation (Yang et al. 2024; Zheng et al. 2024). These models treat videos as spatiotemporal volumes, flattening them into 1D token sequences across time, height, and width. Full self-attention is then used to jointly model spatial and temporal relationships, leading to notable improvements in temporal consistency and spatial detail. Recent work has further advanced video generation by leveraging large transformer backbones and massive video datasets. Notably, models like Wan (Wang et al. 2025a) and HunyuanVideo (Kong et al. 2024) show that scaling up model size and data significantly enhances video quality and diversity. These models achieve impressive text-to-video synthesis, producing videos with rich content and improved temporal consistency. However, *due to the quadratic complexity of self-attention, they remain limited to relatively low resolutions (e.g., 720P), and scaling to higher resolutions is still a major challenge.*

2.2 High-resolution Video Generation

To enable high-resolution generation, some existing methods such as Wan (Wang et al. 2025a) and HunyuanVideo (Kong et al. 2024) train their models on videos of various resolutions, allowing them to scale to arbitrary output sizes. However, when generating videos at resolutions beyond 2K, these approaches often produce blurry results, as illustrated in Fig. 1. In contrast to directly modeling high-resolution generation, other methods (He et al. 2022; Blattmann et al. 2023b; Singer et al. 2022; Ho et al. 2022; Wang et al. 2024), such as Align-Your-Latents (Blattmann et al. 2023b), adopt a two-stage process: they first generate low-resolution videos and then apply super-resolution (Du et al. 2024; Zhou et al. 2024; Zhang and Yao 2024) to up-scale the output. However, super-resolution primarily improves visual sharpness without introducing sufficient new details, resulting in pseudo high-resolution content that lacks authenticity and richness. Some recent works (Wang et al. 2025b; Dalal et al. 2025) have made progress in long video generation by leveraging linear attention mechanisms (Gu and Dao 2023) or test-time training (Zhang et al. 2025); however, they have paid limited attention to scaling up the spatial resolution of videos. To address these challenges, we investigate native high-definition (HD) video generation,

aiming to overcome the high computational costs while producing high-quality HD videos.

3 Preliminaries

Most state-of-the-art Diffusion Transformer (DiT) based video generation models (e.g., Wan (Wang et al. 2025a) and HunyuanVideo (Kong et al. 2024)) adopt a full-attention-based framework, which builds upon the Transformer architecture to model spatiotemporal dependencies in video sequences. Typically, a 3D variational autoencoder (3D-VAE) is first used to encode an input video into a latent representation of shape $D \times T \times H \times W$, where D denotes the hidden dimension, T , H , and W represent the temporal frames, height, and width, respectively. This downsampling strategy effectively reduces the sequence length and makes training tractable for medium-sized videos. Then, the video latents are reshaped into a 1D token sequence with sequence length $N = T \times H \times W$ via a patchify module.

Once the token sequence is obtained, video generation models apply full self-attention mechanisms across the entire sequence. For a sequence of N tokens, the self-attention module computes an $N \times N$ attention map, which scales quadratically with the sequence length. The computational complexity of self-attention is $\mathcal{O}(N^2 \cdot D)$, which becomes prohibitively expensive as the video resolution increases. For instance, doubling the height and width of the video leads to a four-fold increase in the number of tokens and a sixteen-fold increase in the size of the attention map. This quadratic scaling severely limits the feasibility of generating HD videos (e.g., 1080P and even 4K) using existing full-attention modules in terms of training and inference costs.

4 UltraGen: Born for HD Video Generation

4.1 Time-Aware Global-Local Attention

As discussed in Sec. 3, in DiT-based video generation, the computational complexity of full attention is $\mathcal{O}((TWH)^2 \times D)$, which grows quadratically with the spatial size ($W \times H$) of the generated video. To address this, we restrict attention to a fixed local region by introducing an attention window of size (W_0, H_0) . This ensures that, regardless of the overall spatial dimensions, attention is computed only within each (W_0, H_0) window. By applying this *local attention mechanism* to cover the entire frame, the total computational cost increases only linearly with the number of windows, rather than quadratically with frame size. Thus, the overall complexity is reduced to $\mathcal{O}((TW_0H_0)^2 \times D)$ up to a constant factor, effectively avoiding quadratic scaling. However, relying solely on local attention ignores dependencies across windows, potentially leading to isolated or inconsistent content. To address this, we introduce a *global attention mechanism* that connects all local windows, *enabling the model to capture long-range dependencies and maintain semantic consistency across the frame, thereby supporting high-resolution video generation with coherent semantics.*

Therefore, we propose a novel **global-local attention mechanism** that decomposes the original full attention module into two complementary components: *global attention* and *local attention*. Specifically, the local attention module

partitions the video sequence into multiple independent sub-regions and applies attention within each region separately, significantly reducing the overall computational cost. In parallel, the global attention module models the interactions across different local regions, injecting holistic spatiotemporal information into each local branch. This hierarchical design enables efficient and scalable attention modeling while preserving both local detail and global coherence.

Local Attention Mechanism. For a video latent representation $z_v \in \mathbb{R}^{B \times (T \cdot W \cdot H) \times D}$, we aim to reduce the computational burden of self-attention by introducing a **local attention mechanism** that approximates full self-attention with lower computational cost.

We partition the video latent z_v along the spatial dimensions (H and W) into m non-overlapping, equally sized local windows, each with dimensions $B \times (T \cdot W_0 \cdot H_0) \times D$. For each local window, self-attention is applied independently, and the results are aggregated along the spatial dimensions to produce the final local attention output with the original resolution:

$$\begin{aligned} \{z_l^{(i)}\}_{i=1}^m &= \text{Partition}(z_v), \\ z_l^{(i)} &= \text{Self-Attention}(z_l^{(i)}), \\ z_l &= \text{Aggregate}(\{z_l^{(i)}\}_{i=1}^m), \end{aligned} \quad (1)$$

where $\text{Partition}(\cdot)$ divides z_v into m local windows, $\text{Self-Attention}(\cdot)$ is applied within each window, and $\text{Aggregate}(\cdot)$ concatenates the outputs along the spatial dimensions to reconstruct the local attention result $z_l \in \mathbb{R}^{B \times T \times H \times W \times D}$ (detailed designs are in Sec. 4.3).

Global Attention Mechanism. Local attention reduces computational cost but limits focus to individual windows, potentially causing semantic inconsistencies. For example, a prompt describing "a dog" might lead to multiple independent versions across windows.

To address this, we introduce a **global attention module** to capture long-range dependencies and ensure semantic consistency. We compress the spatial information of the video latent z_v into a lower-resolution $z_g \in \mathbb{R}^{B \times (T \cdot H_g \cdot W_g) \times D}$ using a convolution module, apply global self-attention at this reduced size, and decompress the result to the original resolution:

$$\begin{aligned} z_g' &= E_g(z_v), \\ z_g'' &= \text{Self-Attention}(W_Q^g z_g', W_K^g z_g', W_V^g z_g'), \\ z_g &= D_g(z_g''). \end{aligned} \quad (2)$$

where E_g is the compression encoder, and D_g is the decompression function, ensuring z_g matches the original video latent size (detailed designs are in Sec. 4.2).

Time-aware Global-Local Composition. The local and global attention mechanisms yield two latent representations: the local latent z_l , capturing fine-grained details, and the global latent z_g , providing semantically coherent global context. To produce videos that are both globally consistent and locally detailed, we introduce a **global-local fusion module** that combines these representations using a learnable fusion factor α .

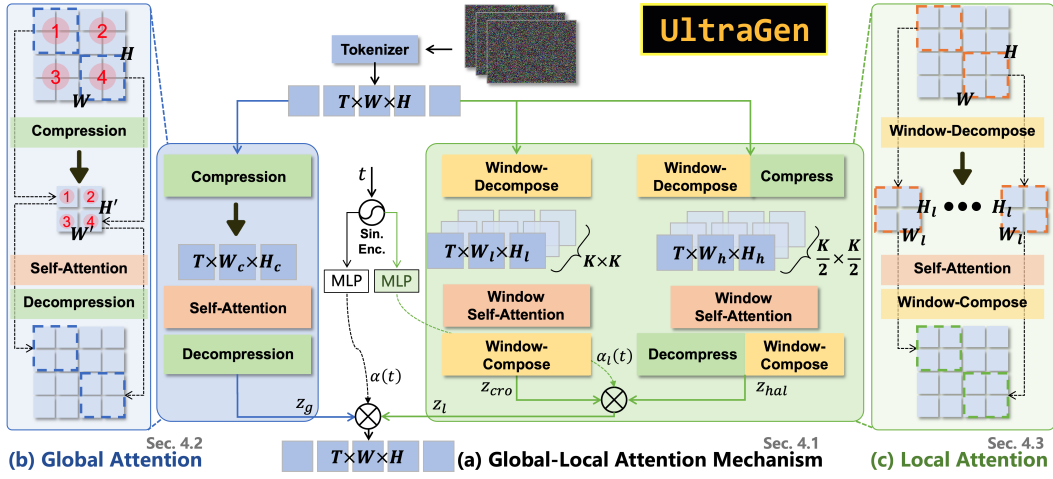


Figure 2: **Overview of our UltraGen** that decomposes the full-attention into a global attention branch (Sec. 4.2) for overall semantic consistency and a local attention branch (Sec. 4.3) for high-fidelity regional content, boosting high-efficiency and high-resolution video generation.

During the diffusion process, different denoising timesteps t focus on various video aspects: early timesteps emphasize global structure, while later ones refine details. Thus, the fusion factor α should dynamically adjust with the timestep, shifting focus from global to local information. To achieve this, we predict α based on timestep t . We embed t into a 256-dimensional time feature vector using Sinusoidal Encoding, then project it into a D -dimensional fusion factor via an MLP to fuse z_l and z_g :

$$\alpha(t) = MLP(\text{SinEncode}(t)), \quad \mathcal{R}^1 \rightarrow \mathcal{R}^D \quad (3)$$

$$z_{\text{fused}} = \alpha(t) \cdot z_g + (1 - \alpha(t)) \cdot z_l.$$

4.2 Spatially-Compressed Global Attention

In this section, we detailedly introduce our *spatially-compressed global attention* module, which is designed to capture global video context while maintaining computational efficiency. *The key idea is to compress the spatial dimensions of video latents before performing attention, so that the self-attention is conducted at a smaller spatial size, and then decompress them back to the original resolution using spatiotemporal convolution.* This reduces the attention cost without sacrificing global modeling capability.

Spatial Compression. A video can be considered as a sequence of consecutive images, and it is well-known that images can be spatially downsampled to lower resolutions while preserving global semantics at the cost of some local details. Leveraging this property, we propose to spatially compress the video latent by downsampling its width and height by a factor of k . This aligns the computational cost of global attention with that of our local attention module.

Specifically, given a video latent $z \in \mathbb{R}^{B \times T \times H \times W \times D}$, we apply a $k \times k$ 2D convolution with stride k along the spatial dimensions (H, W) to obtain a compressed latent $z_c \in \mathbb{R}^{B \times T \times H' \times W' \times D}$, where $H' = \frac{H}{k}$ and $W' = \frac{W}{k}$. To reduce the number of parameters and computational cost in the compression layer, we adopt a channel-wise (i.e., depth-

wise) convolution mechanism, where each hidden dimension is processed by a separate convolution kernel with a single input and output channel. Moreover, to ensure training stability at the early stage, we initialize the convolutional kernel weights to be $1/(k \times k)$, which initially behaves as average pooling.

Global Attention with Domain-aware LoRA. Once we obtain the compressed video latent $z_c \in \mathbb{R}^{B \times T \times H' \times W' \times D}$, we proceed to apply global self-attention over it. However, employing both local and global attention mechanisms requires maintaining two attention weights for each, which significantly increases computational overhead. To address this, we propose a **domain-aware LoRA** mechanism, which adapts the local attention parameters for global modeling. Specifically, for each projection weight $W \in \{W_Q, W_K, W_V\}$ and the FFN parameters W_{FFN} , we introduce a lightweight, trainable low-rank residual that specializes in global attention. The adapted weight is defined as:

$$W^{\text{global}} = W + \Delta W_{\text{LoRA}} = W + A_W B_W, \quad (4)$$

where $A_W \in \mathbb{R}^{d \times r}$ and $B_W \in \mathbb{R}^{r \times d}$ are low-rank matrices with rank $r \ll d$, and d is the input/output dimension. The same formulation is applied to W_{FFN} .

Spatiotemporal Decompression. After obtaining the globally modeled compressed latent $z_c^{\text{global}} \in \mathbb{R}^{B \times T \times H' \times W' \times D}$, we need to decompress it back to the original video resolution $T \times H \times W$.

Specifically, we first apply bilinear interpolation to up-sample the spatial resolution from $H' \times W'$ to $H \times W$. Then, to mitigate the over-smoothing effect caused by interpolation, we apply a convolutional refinement module. Since video frames exhibit not only spatial but also temporal continuity, spatial-only operations may lead to temporal discontinuities. Therefore, we utilize a 3D convolution to perform joint spatio-temporal processing to ensure temporally consistent decompression. The overall process is formulated as:

$$z_g = \text{Conv3D}(\text{BilinearUpsample}(z_c^{\text{global}})), \quad (5)$$

where z_g denotes the decompressed global latent, and Conv3D denotes a 3D convolution operation over the temporal and spatial dimensions. This enables effective restoration of spatial details while preserving temporal coherence.

4.3 Cross-window Hierarchical Local Attention

In order to avoid the quadratic increase in computational complexity as video resolution grows, we design *local attention mechanism* to partition the video latents into non-overlapping spatial windows and then conduct self-attention in local windows. However, this partition makes it difficult to model fine-grained relationships at the boundaries between adjacent local windows. To address this issue, we propose *Cross-window Hierarchical Local Attention*, which can effectively model local dependencies within each window and captures interactions between neighboring windows.

Local Attention. Concretely, we first reshape the video latent z_v into a new tensor of shape $B \times T \times H \times W \times D$. We then partition the spatial dimensions (H, W) into $K \times K$ non-overlapping local windows, resulting in a set of local video latent groups $\{v_{i,j}\}_{i=1,j=1}^{K,K}$, where each $v_{i,j} \in \mathbb{R}^{B \times T \times \frac{H}{K} \times \frac{W}{K} \times D}$ corresponds to a spatiotemporal sub-volume of the original video latent:

$$v_{i,j} = z_v[:, :, i \cdot \frac{H}{K} : (i+1) \cdot \frac{H}{K}, j \cdot \frac{W}{K} : (j+1) \cdot \frac{W}{K}, :]. \quad (6)$$

For each local video latent $v_{i,j}$, we apply self-attention within it to model the spatiotemporal dependencies:

$$v'_{i,j} = \text{Self-Attention}(W_Q^s v_{i,j}, W_K^s v_{i,j}, W_V^s v_{i,j}). \quad (7)$$

By applying self-attention only within each local window, the computational complexity is reduced from $\mathcal{O}((TWH)^2 \cdot D)$ to $\mathcal{O}(K^2 \cdot (\frac{TWH}{K^2})^2 \cdot D) = \mathcal{O}((TWH)^2 \cdot D / K^2)$. As the number of windows increases (i.e., window size decreases), the complexity decreases accordingly. In the extreme case, it reduces the complexity to $\mathcal{O}(TWH \cdot D)$ when each token forms an independent local group, enabling high-resolution video generation at significantly reduced cost.

After computing self-attention within each local window, we aggregate all locally updated features $\{v'_{i,j}\}$ and restore them to the original video latent resolution:

$$z_l = \text{Rearrange}(\{v'_{i,j}\}_{i,j=1}^K). \quad (8)$$

This rearranged z_l preserves the original spatial-temporal resolution of the video while significantly reducing the computation required during attention modeling, which ensures that local details are efficiently captured.

Cross-window Attention. Despite incorporating global information modeling, direct communication between local attention windows remains limited, especially at the boundaries, where discontinuities frequently occur. To address this, we propose a *Cross-window Local Attention* to enhance inter-window interaction across local attention windows.

Given that the model is composed of multiple layers of transformer blocks, we apply alternating local attention schemes at adjacent layers, where adjacent layers have different partition blocks and window boundaries. For an even-numbered layer i ($i \bmod 2 = 0$), the spatial domain of the

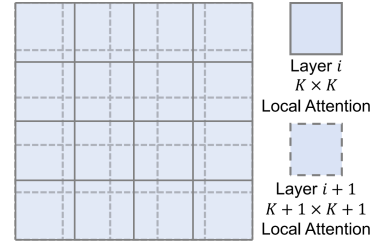


Figure 3: Cross-window Attention.

video latent is partitioned into non-overlapping $K \times K$ windows. For an odd-numbered layer i ($i \bmod 2 = 1$), we apply a shifted window strategy with $(K+1) \times (K+1)$ partitions that partially overlap with the even-layer windows. This cross-window local attention strategy enables hierarchical interaction across neighboring windows between adjacent transformer layers.

As a result, boundary information in the $K \times K$ windows at layer i is propagated through overlapping regions in the $(K+1) \times (K+1)$ windows at layer $i+1$, and vice versa. This enhances continuity across local attention boundaries and improves consistency in the generated outputs. Formally, the attention computation in layer i can be described as:

$$z_{cro}^{(i)} = \text{LocalAttn}_{(K+(i \bmod 2)) \times (K+(i \bmod 2))}(z^{(i)}). \quad (9)$$

Hierarchical Local Attention. While the proposed cross-window local attention enhances information exchange across adjacent local attention windows, it may still be hard to capture fast-moving small objects, which can simultaneously span multiple local windows across frames. In such cases, the limited overlapping in cross-window attention is insufficient, and global attention lacks the resolution to model fine-grained local details. To address this, we introduce a *Hierarchical Local Attention* (HLA) mechanism, which divides the full attention into $(K/2) \times (K/2)$ coarse windows (each twice the size as the local window), and performs local attention within each coarse window at an intermediate scale. This approach effectively compensates for the inability of global attention to capture fine-grained details, while also overcoming the limited receptive field inherent in conventional local attention mechanisms.

Specifically, we first compress the latent features within each local window using a strategy similar to our spatial-compressed global attention. The local latent z_c^{hla} within each coarse window of size $\frac{2H}{K} \times \frac{2W}{K}$ is downsampled via strided convolution. To effectively model the hierarchical attention, we apply a *domain-aware LoRA* adaptation to the pretrained attention weights (including W_Q , W_K , W_V , and FFN) to ensure they are appropriately adapted for hierarchical attention computation:

$$W_{hla} = W_{\text{local}} + \Delta W^{\text{HLA}}, \quad (10)$$

where ΔW^{HLA} is the domain-specific LoRA adaptation for hierarchical attention.

Similar to the cross-window local attention design, we employ an alternating shift mechanism between adjacent

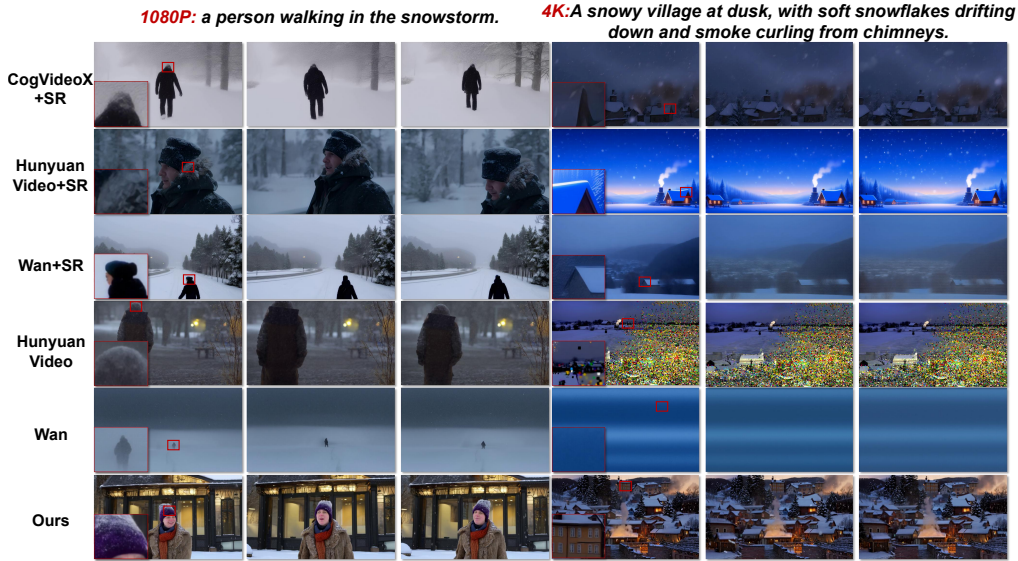


Figure 4: Comparison results of existing state-of-the-art video generation methods on 1080P video generation. The red boxes highlight zoomed-in regions, where our model produces the clearest high-resolution videos with the most fine-grained details.

transformer layers to ensure information flow across hierarchical attention windows. That is, for layer i , hierarchical attention is computed with non-overlapping $(K/2) \times (K/2)$ windows; for layer $i+1$, we partition the spatial domain into $(K/2+1) \times (K/2+1)$ non-overlapping windows, making the windows of adjacent layers intersect with each other and thus enabling boundary information propagation. The attention operation at each layer can be described as:

$$z_{hla}^{(i)} = \text{HierAttn}_{(\frac{K}{2}+(i \bmod 2)) \times (\frac{K}{2}+(i \bmod 2))}(z^{(i)}), \quad (11)$$

where $\text{HierAttn}_{k \times k}(\cdot)$ denotes attention over a $k \times k$ partitioned hierarchical window.

This hierarchical structure, combined with cross-layer shift design and domain-aware adaptation, enables efficient fine-grained motion modeling of fast-moving small objects and enhances the robustness of local attention modeling in dynamic video scenes. To fuse the results from both the *Cross-window Local Attention* z_{cro} and the *Hierarchical Local Attention* z_{hla} , we employ a time-aware alpha α_{local} to fuse the two results, which is the same as the *Time-aware Global-Local Composition*.

5 Experiments

5.1 Implementation Details

Baselines. We compare our model with state-of-the-art methods, including Wan (Wang et al. 2025a), Hunyuan-Video (Kong et al. 2024), and CogVideo-X (Yang et al. 2024). For each method, we generate two sets of videos: 1) one by directly generating videos at the target resolution, and 2) the other by first generating videos at the default resolution and then applying a super-resolution method (Zhang and Yao 2024) to upscale them to the target size. Note that CogVideoX cannot support HD video generation; therefore, we directly combine it with video super-resolution.

Evaluation Metrics. Conventional metrics such as FVD are inadequate for evaluating the quality of high-resolution video generation, as they rely on pretrained low-resolution video encoders that fail to capture high-resolution features. To address this limitation, we introduce three novel metrics specifically designed for high-resolution video evaluation: 1) **HD-FVD** measures the similarity between generated and real high-resolution videos, while 2) **HD-MSE** and 3) **HD-LPIPS** assess the fine-grained pixel-level and semantic-level details of the generated videos, respectively. Additional CLIP score (Radford et al. 2021) and temporal consistency (Huang et al. 2024a) are included for a more comprehensive evaluation. Further details and more Vbench metrics are provided in the appendix (Hu et al. 2025d).

5.2 Comparison Results

Qualitative Comparison. We compare our model with state-of-the-art methods on both 1080P and 4K video generation tasks. The comparison results are shown in Fig. 4. As can be seen, the Wan model is unable to directly generate 1080P videos, resulting in blurry outputs with little to no semantic content. HunyuanVideo is capable of generating 1080P videos, but often produces results with incorrect semantics that are inconsistent with the given prompt. Methods that combine super-resolution models can generate text-aligned videos; however, the outputs after super-resolution tend to be overly smooth and lack fine details. Among these, only HunyuanVideo+SR produces relatively good results, but the level of detail is still significantly lower than that of our model, as highlighted in the zoomed-in red boxes. Therefore, our model is able to generate high-resolution videos with fine-grained details while faithfully following the given prompt, demonstrating its superior performance in high-resolution video generation. Additional results generated by our model can be found in the Appendix.

Resolution	Method	SR	HD-FVD ↓	HD-MSE ↑	HD-LPIPS ↑	CLIP-L ↑	Temporal Consis ↑
1080P	CogVideoX	✓	394.82	97.21	0.3060	0.2834	0.9468
	HunyuanV	✓	238.75	126.68	0.3590	0.2883	0.9614
	Wan	✓	309.10	163.86	0.3499	0.2747	0.9750
	HunyuanV	✗	237.89	207.68	0.4911	0.2636	0.9752
	Wan	✗	821.54	42.93	0.4290	0.2528	0.9768
	Ours	✗	214.12	390.19	0.5455	0.2654*	0.9827
4K	CogVideoX	✓	574.10	68.94	0.2645	0.2436	0.9449
	HunyuanV	✓	453.41	276.76	0.4066	0.2576	0.9684
	Wan	✓	471.56	77.67	0.2782	0.2455	0.9697
	HunyuanV	✗	805.42	102.36	0.3858	0.2151	0.9679
	Wan	✗	1272.08	29.45	0.4270	0.2123	0.9705
	Ours	✗	424.61	386.01	0.6450	0.2444*	0.9710

Table 1: Quantitative comparisons. Our UltraGen demonstrates superior high-quality HD video generation capabilities. **Bold** indicates the best performance and * indicates the best performance among all the non-SR methods.

Resolution	HunyuanVideo	Wan	UltraGen (Ours)	Speedup (Ours)
1080P	43 min	35 min	13 min	×2.69
4K	11h 36min	8h 46min	1h 50min	×4.78

Table 2: Comparison of inference time. Our model archives a **4.78 × speedup** compared to the baseline Wan model.

Quantitative Comparison. We compare our method with state-of-the-art approaches in Tab. 1. For HD evaluation metrics, our model achieves the lowest HD-FVD scores on both 1080P and 4K video generation, indicating superior quality and diversity in the generated videos. Furthermore, we obtain the best HD-MSE and HD-LPIPS, demonstrating that our generated videos contain the most fine-grained details and validating the effectiveness of our HD video generation ability. Our model also achieves the best temporal consistency, which demonstrates the smoothness of the generated videos and the coherence across frames. In terms of prompt following, we observe that directly generating HD videos without super-resolution leads to a relatively lower CLIP score due to the difficulty in high-resolution video generation. Since our model is based on Wan 1.3B, its CLIP score cannot surpass that of Wan+SR. Nevertheless, we still achieve the best CLIP score among methods that natively generate high-resolution videos, highlighting the strong prompt-following capability of our model.

Time Comparison. Finally, we compare the inference time of our model with HunyuanVideo and Wan at different resolutions, as shown in Tab. 2. Our model achieves a $2.7\times$ speedup for 1080P generation and a **4.78× speedup** for 4K generation compared to the baseline Wan model, demonstrating the high efficiency of our approach for high-resolution video generation.

5.3 Ablation Studies

We conduct ablation studies on five variants: (1) without global attention, (2) without hierarchical attention, (3) without domain-aware LoRA, (4) without cross-window local attention, and (5) employing Swin-Attention (Liu et al. 2021) for local attention modeling. As shown in Fig. 5, the model without global attention tends to generate disjoint content, exemplified by the isolated 16 golden fishes in the rightmost

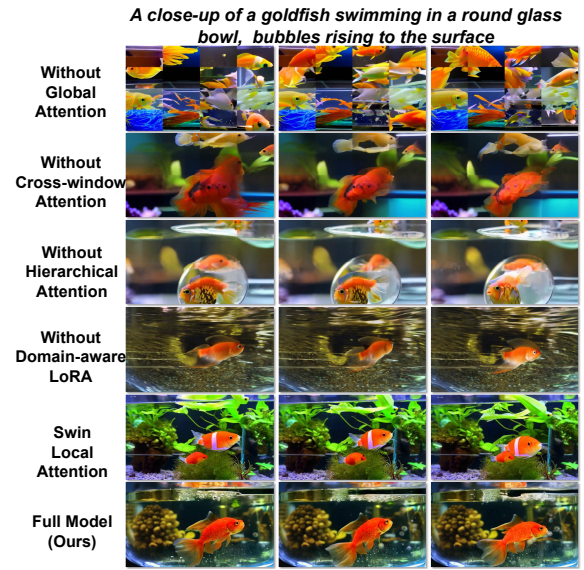


Figure 5: Ablation study on the proposed modules.

case. Models lacking either cross-window local attention or hierarchical attention can capture global relationships only coarsely and still exhibit inconsistencies at window boundaries. The model without domain-aware LoRA alleviates boundary inconsistency but suffers from reduced generation quality, producing somewhat blurry results. This is due to the limited capacity of a single set of attention weights to model three distinct attention mechanisms (global, local, and hierarchical). Moreover, when replacing hierarchical cross attention with Swin-Attention for local attention modeling, we observe that although adjacent windows can be connected smoothly, Swin-Attention struggles to effectively capture hierarchical features. As a result, the model often generates semantically inconsistent content across windows. For example, it may produce two goldfish in adjacent windows where only one should appear, indicating a lack of semantic coherence. In contrast, the full model generates high-quality videos, effectively resolves boundary inconsistencies, and captures global semantics well.

6 Conclusion

In this work, we propose UltraGen, a novel framework for efficient, end-to-end native high-resolution video generation. By leveraging a hierarchical dual-branch attention architecture, UltraGen effectively decouples local and global attention, enabling the synthesis of high-fidelity regional details while maintaining overall semantic consistency. Our spatially compressed global modeling and hierarchical cross-window local attention mechanisms further reduce computational complexity, making high-resolution video generation (up to 4K) feasible for both training and inference. Extensive experiments demonstrate that UltraGen not only scales pre-trained low-resolution models to 1080P and 4K resolutions, but also consistently outperforms existing state-of-the-art methods and super-resolution pipelines in both qualitative and quantitative evaluations.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 62302297, 72192821, 62272447, 62472285, 62472282), the Fundamental Research Funds for the Central Universities (YG2023QNB17, YG2024QNA44), National Key R&D Program of China (2024YFE0115500), Young Elite Scientists Sponsorship Program by CAST (2022QNRC001), Beijing Natural Science Foundation (L222117).

References

- Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.
- Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22563–22575.
- Chen, Y.; Zhang, J.; Hu, T.; Zeng, Y.; Xue, Z.; He, Q.; Wang, C.; Liu, Y.; Hu, X.; and Yan, S. 2025. IVEBench: Modern Benchmark Suite for Instruction-Guided Video Editing Assessment. *arXiv preprint arXiv:2510.11647*.
- Dalal, K.; Koceja, D.; Xu, J.; Zhao, Y.; Han, S.; Cheung, K. C.; Kautz, J.; Choi, Y.; Sun, Y.; and Wang, X. 2025. One-minute video generation with test-time training. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17702–17711.
- Du, Y.; Hu, T.; Yi, R.; and Ma, L. 2024. LD-BFR: Vector-Quantization-Based Face Restoration Model with Latent Diffusion Enhancement. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2852–2860.
- Gu, A.; and Dao, T. 2023. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; and Dai, B. 2023. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*.
- He, Y.; Yang, T.; Zhang, Y.; Shan, Y.; and Chen, Q. 2022. Latent video diffusion models for high-fidelity long video generation. *arXiv preprint arXiv:2211.13221*.
- Ho, J.; Chan, W.; Saharia, C.; Whang, J.; Gao, R.; Gritsenko, A. A.; Kingma, D. P.; Poole, B.; Norouzi, M.; Fleet, D. J.; and Salimans, T. 2022. Imagen Video: High Definition Video Generation with Diffusion Models. *ArXiv*, abs/2210.02303.
- Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.
- Hu, L. 2024. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8153–8163.
- Hu, T.; Yu, Z.; Zhang, G.; Su, Z.; Zhou, Z.; Zhang, Y.; Zhou, Y.; Lu, Q.; and Yi, R. 2025a. Harmony: Harmonizing Audio and Video Generation through Cross-Task Synergy. *arXiv preprint arXiv:2511.21579*.
- Hu, T.; Yu, Z.; Zhou, Z.; Liang, S.; Zhou, Y.; Lin, Q.; and Lu, Q. 2025b. HunyuanCustom: A Multimodal-Driven Architecture for Customized Video Generation. *arXiv:2505.04512*.
- Hu, T.; Yu, Z.; Zhou, Z.; Zhang, J.; Zhou, Y.; Lu, Q.; and Yi, R. 2025c. PolyVivid: Vivid Multi-Subject Video Generation with Cross-Modal Interaction and Enhancement. *arXiv preprint arXiv:2506.07848*.
- Hu, T.; Zhang, J.; Su, Z.; and Yi, R. 2025d. UltraGen: High-Resolution Video Generation with Hierarchical Attention. *arXiv preprint arXiv:2510.18775*.
- Hu, T.; Zhang, J.; Yi, R.; Huang, H.; Wang, Y.; and Ma, L. 2025e. High-efficient diffusion model fine-tuning with progressive sparse low-rank adaptation. In *13th International Conference on Learning Representations, ICLR 2025*, 92066–92078. International Conference on Learning Representations, ICLR.
- Hu, T.; Zhang, J.; Yi, R.; Wang, Y.; Huang, H.; Weng, J.; Wang, Y.; and Ma, L. 2024. Motionmaster: Training-free camera motion transfer for video generation. *arXiv preprint arXiv:2404.15789*.
- Huang, Z.; He, Y.; Yu, J.; Zhang, F.; Si, C.; Jiang, Y.; Zhang, Y.; Wu, T.; Jin, Q.; Chanpaisit, N.; et al. 2024a. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21807–21818.
- Huang, Z.; Tang, F.; Zhang, Y.; Cun, X.; Cao, J.; Li, J.; and Lee, T.-Y. 2024b. Make-your-anchor: A diffusion-based 2d avatar generation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6997–7006.
- Kong, W.; Tian, Q.; Zhang, Z.; Min, R.; Dai, Z.; Zhou, J.; Xiong, J.; Li, X.; Wu, B.; Zhang, J.; et al. 2024. Hunyuan-video: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*.
- Labs, B. F. 2024. FLUX. <https://github.com/black-forest-labs/flux>.
- Liang, S.; Yu, Z.; Zhou, Z.; Hu, T.; Wang, H.; Chen, Y.; Lin, Q.; Zhou, Y.; Li, X.; Lu, Q.; et al. 2025. OmniV2V: Versatile Video Generation and Editing via Dynamic Content Manipulation. *arXiv preprint arXiv:2506.01801*.
- Liu, L.; Ma, T.; Li, B.; Chen, Z.; Liu, J.; Li, G.; Zhou, S.; He, Q.; and Wu, X. 2025. Phantom: Subject-consistent video generation via cross-modal alignment. *arXiv preprint arXiv:2502.11079*.
- Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.
- Peebles, W.; and Xie, S. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4195–4205.

- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmlR.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Singer, U.; Polyak, A.; Hayes, T.; Yin, X.; An, J.; Zhang, S.; Hu, Q.; Yang, H.; Ashual, O.; Gafni, O.; et al. 2022. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*.
- Villegas, R.; Babaeizadeh, M.; Kindermans, P.-J.; Moraldo, H.; Zhang, H.; Saffar, M. T.; Castro, S.; Kunze, J.; and Erhan, D. 2022. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*.
- Wang, A.; Ai, B.; Wen, B.; Mao, C.; Xie, C.-W.; Chen, D.; Yu, F.; Zhao, H.; Yang, J.; Zeng, J.; et al. 2025a. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*.
- Wang, H.; Ma, C.-Y.; Liu, Y.-C.; Hou, J.; Xu, T.; Wang, J.; Juefei-Xu, F.; Luo, Y.; Zhang, P.; Hou, T.; Vajda, P.; Jha, N. K.; and Dai, X. 2025b. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2578–2588.
- Wang, Y.; Chen, X.; Ma, X.; Zhou, S.; Huang, Z.; Wang, Y.; Yang, C.; He, Y.; Yu, J.; Yang, P.; et al. 2024. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 1–20.
- Xue, Z.; Zhang, J.; Xie, X.; Cai, Y.; Liu, Y.; Li, X.; and Tao, D. 2025. AdaVideoRAG: Omni-Contextual Adaptive Retrieval-Augmented Efficient Long Video Understanding. In *NeurIPS*.
- Yang, Z.; Teng, J.; Zheng, W.; Ding, M.; Huang, S.; Xu, J.; Yang, Y.; Hong, W.; Zhang, X.; Feng, G.; et al. 2024. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*.
- Yi, R.; Hu, T.; Xia, M.; Tang, Y.; and Liu, Y.-J. 2024. Feditnet++: Few-shot editing of latent semantics in gan spaces with correlated attribute disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhang, T.; Bi, S.; Hong, Y.; Zhang, K.; Luan, F.; Yang, S.; Sunkavalli, K.; Freeman, W. T.; and Tan, H. 2025. Test-time training done right. *arXiv preprint arXiv:2505.23884*.
- Zhang, Y.; and Yao, A. 2024. Realviformer: Investigating attention for real-world video super-resolution. In *European Conference on Computer Vision*, 412–428. Springer.
- Zheng, Z.; Peng, X.; Yang, T.; Shen, C.; Li, S.; Liu, H.; Zhou, Y.; Li, T.; and You, Y. 2024. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*.
- Zhou, S.; Yang, P.; Wang, J.; Luo, Y.; and Loy, C. C. 2024. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2535–2545.