ApertureAttn: Native 4K Video Generation with Image-Only Supervision

Ruonan Yu[1], **Zigeng Chen**[1], **Zhenxiong Tan**[1], **Songhua Liu**[2]*, **Xinchao Wang**[1]*
[1]National University of Singapore
[2]School of Artificial Intelligence, Shanghai Jiao Tong University
ruonan@u.nus.edu, liusonghua@sjtu.edu.cn, xinchao@nus.edu.sg

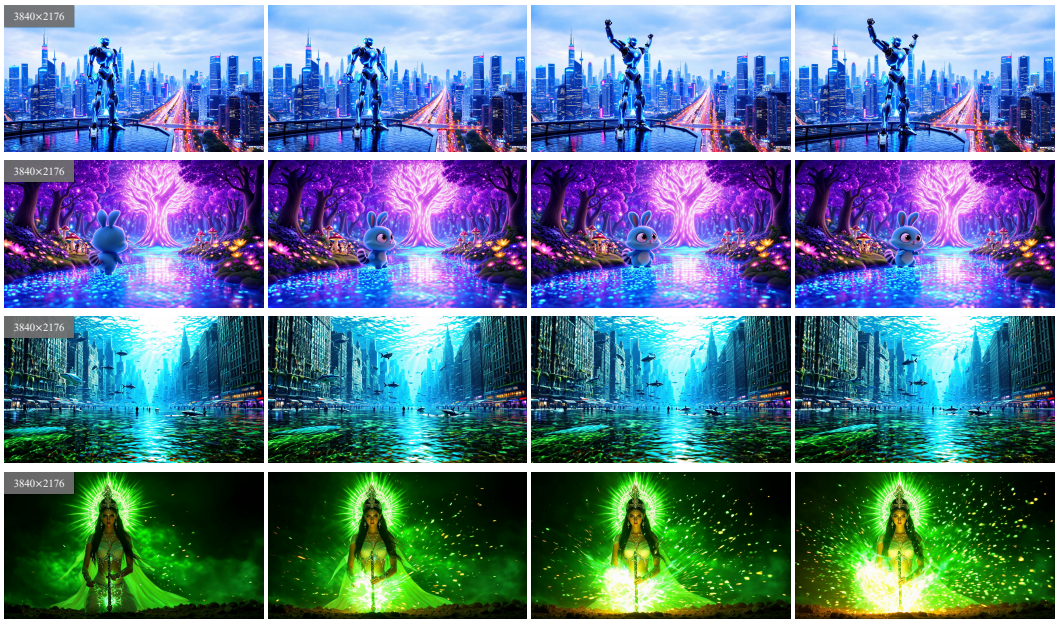

Figure 1: Visualization results of our proposed method. All results are sampled from high-resolution videos with resolution of 3840 × 2176 (4K) and 81 frames.

Abstract

Recent progress in video generation has been remarkable, yet state-of-the-art models remain largely confined to 720p, well short of the 4K resolution of modern displays. Scaling video generation to higher resolutions is constrained by the scarcity of high-resolution video data and the prohibitive cost of large-scale training. Beyond data and compute, higher resolutions place substantially greater demands on spatial modeling, particularly for fine-grained details, complex compositions, and local structures. Such high-resolution spatial supervision, however, is abundantly available in high-resolution images, which are easier to collect and train on at scale. This naturally raises a question: can image-only supervision unlock native 4K video generation? To answer this, we propose ApertureAttn, an inverted attention pyramid that concentrates adaptation on high-resolution fine-grained spatial modeling using only single-frame high-resolution images, without any video training data. We further introduce window-scaled extrapolation and

*Corresponding Authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

proximity-guided temporal attention to enhance spatiotemporal consistency. Extensive experiments show that our method enables native 4K video generation with strong visual fidelity and temporal coherence, while requiring only image supervision and minimal training cost. Despite this lightweight training setup, it remains competitive with state-of-the-art high-resolution video generation methods trained directly on high-resolution video data at substantially higher training cost.

1 Introduction

Recent diffusion video generation models [21, 1, 11, 25, 9, 5, 23, 3], such as Wan [35], HunyuanVideo [20], CogVideoX [41], and Open-Sora [52], have made rapid progress in producing high-quality videos with advanced temporal coherence, visual fidelity, and semantic alignment. However, most of these models are designed and trained for resolutions around 720p, falling short of the 4K resolution required by modern displays and professional production workflows.

Scaling video generation to 4K resolution introduces a set of compounding challenges. In practical, high-resolution video data remain limited in scale and availability, and training directly at such resolutions is prohibitively expensive, making end-to-end high-resolution training impractical for most settings. More fundamentally, even with sufficient data and computational resources, the architectural design of current video generation models poses inherent difficulties for high-resolution generation. At high resolutions, attention mechanisms operate over a much larger token space, which makes attention more diffuse and compromises both detail fidelity and global semantic coherence. Positional encodings, *e.g.*, RoPE [33], become less effective at high resolutions, as the enlarged spatial extent weakens their ability to distinguish fine-grained positional differences, leading to further detail loss and structural repetition.

To address these challenges, existing methods mainly fall into two categories: multi-stage generation and native high-resolution generation. Multi-stage methods [7, 37, 27, 19, 30, 8, 53, 26, 28] follow a coarse-to-fine paradigm, first generating videos at low resolution and then progressively increasing the resolution through iterative denoising and refinement. Although this design avoids the difficulty of training directly at high resolution, the cascaded pipeline inevitably introduces error accumulation and increases inference complexity. More importantly, high-resolution video generation is not simply a detail-enhanced extension of low-resolution generation, but follows a distinct visual distribution with richer textures, more complex spatial composition, and finer structural details. Consequently, methods that start from low-resolution generations and refine them progressively remain inherently bounded by the low-resolution regime. Native high-resolution generation [47, 14, 46, 43, 42, 4, 38] directly synthesizes videos at the target resolution. Existing native high-resolution generation methods, such as UltraWan [39], UltraGen [17] and T3-Video [45], achieve native 4K video generation with local attention designs. However, a common limitation of existing native high-resolution methods is their reliance on high-resolution video data, which remain extremely limited and expensive to obtain at scale. Training at such resolutions is also computationally demanding, for instance, UltraWan-4K reports 7.6K GPU hours for adaptation even under LoRA-based fine-tuning [16]. Such data and computation resources requirements make these approaches difficult to scale and broadly adopt.

Crucially, compared with 720p generation, 4K generation introduces additional challenges in spatial modeling, especially in capturing fine details, complex composition, and local structure. While high-resolution video data remains limited and costly to leverage for training, the spatial knowledge required for such synthesis is abundantly represented in high-resolution images, which are readily available at scale. This raises a natural question: *can high-resolution images alone provide sufficient supervision to unlock native 4K video generation?*

We show that native 4K video generation can be achieved from a pretrained 720p video model using only high-resolution image supervision, with substantially reduced training data and training computation costs. Specifically, we propose ApertureAttn, a counterintuitive inverted attention pyramid that progressively shifts from global semantic coordination to fine-grained spatial refinement across layer groups. Each local attention window is defined at a spatial scale within the pre-trained capacity, allowing the model to refine high-resolution content within a range where its learned spatiotemporal priors remain effective. During finetuning, we reduce attention computation to 2D spatial operations, concentrating updates on intra-frame spatial interactions and detailed high-resolution spatial modeling. We also introduce window-scaled extrapolation and proximity-guided

temporal attention to further enhance spatial fidelity and temporal coherence at ultra-high resolution by expanding the effective spatial receptive field and enabling distance-aware cross-frame modeling. By training exclusively on high-resolution images, we adapt a pretrained 720p video model to generate temporally coherent, high-fidelity, high-resolution videos without using any video data. Our contributions are summarized as follows:

- To the best of our knowledge, we are the first to achieve native 4K video generation trained solely on 2D images. We propose ApertureAttn, an inverted attention pyramid that progressively transitions from global semantic coordination to fine-grained spatial refinement, enabling effective high-resolution video generation under image-only supervision.
- We propose window-scaled extrapolation and proximity-guided temporal attention to further enhance spatiotemporal consistency by enlarging spatial receptive fields and modulating temporal attention based on frame distance.
- Extensive experimental results demonstrate that our method achieves native 4K video generation with strong visual fidelity and temporal coherence using only high-resolution image supervision and minimal training cost. Despite these modest requirements, it remains competitive with existing state-of-the-art native high-resolution generation methods that rely on high-resolution video data and substantially higher training costs.

2 Method

2.1 Preliminary

Modern video generation models, *e.g.*, Wan [35], adopt the Diffusion Transformer (DiT) architecture with 3D full attention and 3D Rotary Position Embedding (RoPE). Let $z \in \mathbb{R}^{f\times h\times w\times c}$ denote the video latent, where f, h, and w are the temporal and spatial dimensions in latent space, c is the feature dimensions of video embeddings. The latent representation forms a $N = f \times h \times w$ sequence of spatiotemporal tokens, which are processed by a stack of transformer blocks with 3D self-attention:

$$\text{Attn3D}(X) = \text{Attention}(X) \cdot W_O,\ \text{Attention}(X) = \text{Softmax}\left(\frac{XW_Q(XW_K)^T}{\sqrt{d_k}}\right) XW_V, \quad (1)$$

where $X \in \mathbb{R}^{N\times d}$ is the hidden states of the input video latent, d_k is the hidden dimension of transformer, W_Q, W_K, W_V, and W_O are learnable projection matrices. At higher resolutions, attention is distributed over many more spatial tokens, which often makes the softmax weights less concentrated and weakens detail fidelity.

To encode spatiotemporal positions, current video generation models employ 3D Rotary Position Embedding (RoPE). The head dimension is partitioned into three groups, each assigned to the temporal, height, and width axes respectively. For a token at position (i, j, k), the rotation angle along axis $u \in \{f, h, w\}$ at dimension index m is:

$$\theta_m^{(u)} = p_u \cdot \beta^{-2m/D_u}, \quad (2)$$

where $p_f = i$, $p_h = j$, and $p_w = k$. Here, D_u is the number of dimensions allocated to axis u, and β is the base frequency. When the generation resolution exceeds the training resolution, *e.g.*, at 4K, position indices along the spatial axes fall into unseen ranges, leading to extrapolation failure and degraded generation quality. In this paper, we address the above challenges without any high-resolution video data, enabling native 4K generation with only image supervision.

2.2 ApertureAttn

As discussed in Section 2.1, scaling video generation to ultra-high resolution causes the token sequence length to grow quadratically with spatial resolution. This makes full 3D self-attention over the entire sequence prohibitively expensive and results in increasingly diffuse attention distributions, degrading both fine-grained detail synthesis and global semantic coherence. A natural alternative is to replace full attention with local attention, which restricts the receptive field of each token to a smaller spatial region and thus allows attention to focus more effectively on local structures and details. However, purely local attention is inherently limited in global modeling and can easily lead to duplicated subjects, inconsistent scene structures, or weak global semantic coordination.

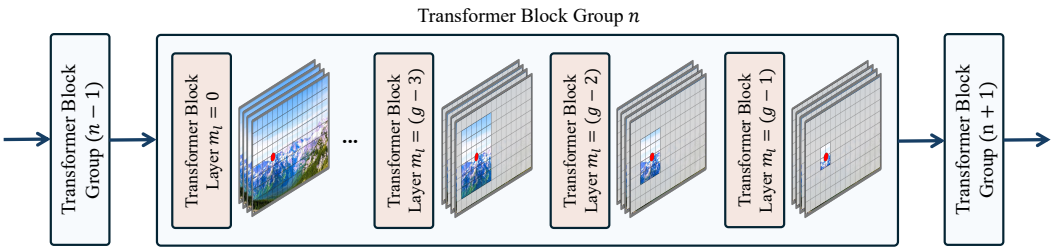

Figure 2: Overview of the ApertureAttn architecture. The transformer blocks are divided into groups of g layers. Each group begins with full attention to establish the global context, and then progressively narrows the receptive field to refine fine-grained details.

To overcome these limitations, we propose **ApertureAttn**, an inverted attention pyramid architecture, as shown in Fig. 2. Unlike the conventional local-to-global schedule, ApertureAttn introduces global context before progressively localizing attention. This mitigates premature semantic commitments within isolated local windows, which later global layers may struggle to reconcile and which can lead to duplicated subjects or incoherent scene structure. Concretely, we organize transformer blocks into groups of g layers, where each group starts with full attention to establish scene layout, subject count, and cross-frame relations, and then progressively restricts the receptive field to refine appearance, texture, and fine-grained structures. Formally, we divide the L transformer layers into consecutive groups of size g. For layer $l \in \{0, \dots, L-1\}$, its within-group index is defined as $m_l = l \bmod g$. The first layer in each group ($m_l = 0$) performs full attention to restore global coordination, while the remaining layers perform neighborhood attention with progressively shrinking spatial kernels. Concretely, for $m_l \in \{1, \dots, g-1\}$, the local attention kernel is parameterized by $\mathbf{R}^{(m_l)} = (R_t^{(m_l)}, R_h^{(m_l)}, R_w^{(m_l)})$, whose spatial components satisfy:

$$R_h^{(1)} > R_h^{(2)} > \cdots > R_h^{(g-1)}, \quad R_w^{(1)} > R_w^{(2)} > \cdots > R_w^{(g-1)}. \tag{3}$$

This inverted schedule is repeated across the network, enabling global context to be regularly reintroduced while deeper layers progressively specialize in local spatial refinement.

For local layers, we adopt neighborhood attention [13]. Unlike block-based partitioning schemes that confine attention to disjoint windows, neighborhood attention uses a token-centered receptive field that varies smoothly across spatial positions. This design avoids hard window boundaries and the associated artifacts that can disrupt spatial coherence, making it well suited for the local refinement stages of our inverted attention pyramid. Formally, for a token at spatiotemporal position (i, j, k), the attention neighborhood at layer l is defined as:

$$\mathcal{N}^{(l)}(i,j,k) = \left\{ (i',j',k') : |i'-i| \le R_t^{(m_l)},\ |j'-j| \le R_h^{(m_l)},\ |k'-k| \le R_w^{(m_l)} \right\}, \tag{4}$$

where $(i', j', k') \in [0, f-1] \times [0, h-1] \times [0, w-1]$. To preserve temporal coherence, we set $R_t^{(m_l)} = f$ for all local layers, making attention global along the temporal axis, while $R_h^{(m_l)}$ and $R_w^{(m_l)}$ follow the inverted-pyramid schedule in space.

$$\text{Attention}^{(l)}(i,j,k) = \text{Softmax}\left(\frac{X_{i,j,k}^{(l)} W_Q \left(X_{\mathcal{N}^{(l)}(i,j,k)}^{(l)} W_K \right)^\top}{\sqrt{d_k}} \right) X_{\mathcal{N}^{(l)}(i,j,k)}^{(l)} W_V, \tag{5}$$

$$\text{Attn3D}^{(l)}(i,j,k) = \text{Attention}^{(l)}(i,j,k) W_O. \tag{6}$$

The full attention output at layer l is obtained by applying the above operation to every token and stacking the outputs over all spatiotemporal positions:

$$\text{Attn3D}^{(l)}(X^{(l)}) = \left[\text{Attn3D}^{(l)}(i,j,k) \right]_{(i,j,k) \in [0,f-1] \times [0,h-1] \times [0,w-1]}. \tag{7}$$

Image-Only Finetuning. Based on ApertureAttn, we adapt the pretrained video generation model to higher resolutions using solely 2D images. The key is that ApertureAttn decomposes 4K generation into windowed computations whose token scale remains close to that encountered during video

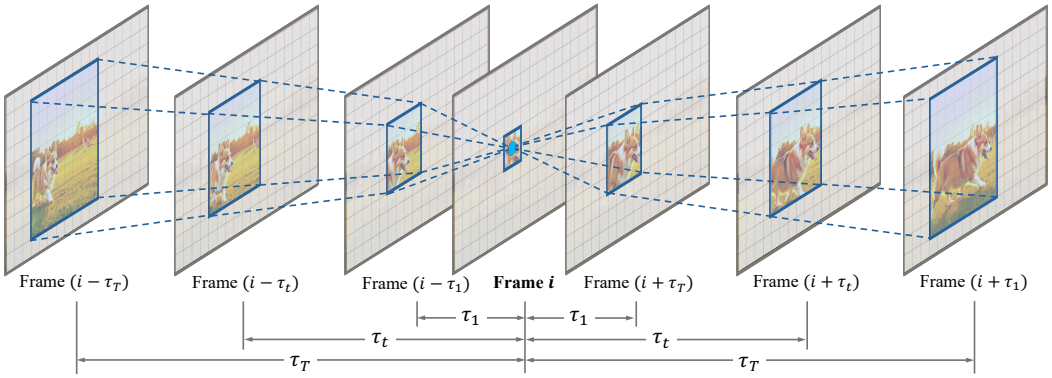

Figure 3: Illustration of proximity-guided temporal attention in a single transformer layer. Cross-frame attention uses larger windows for frames that are farther from the target frame, with the window size increased according to distance thresholds.

pretraining. This scale alignment allows the pretrained model to reuse its learned spatiotemporal priors, including motion patterns and cross-frame dynamics, without requiring additional high-resolution video supervision to relearn temporal modeling. During image-only adaptation, attention is explicitly computed in the 2D spatial domain, and the training signal is largely confined to intra-image spatial interactions and focuses on strengthening local spatial synthesis, *e.g.*, fine-grained structures, across windows of different sizes. To facilitate fine-detail learning, ApertureAttn intentionally adopts a relatively small-window schedule during adaptation. Overly large receptive fields at this stage may dilute fine-grained spatial supervision, amplify temporal noise, and interfere with the motion priors inherited from video pretraining. In contrast, small windows keep local computations within a stable local regime. The full-attention layer at the beginning of each group then coordinates information across local regions, helping preserve global layout and semantic consistency at 4K resolution.

2.3 Window-Scaled Extrapolation

Although small attention windows are beneficial for learning fine-grained details, directly applying the same small-window schedule to native 4K generation is insufficient. As the spatial canvas expands, each attention window captures only a limited portion of the frame, reducing long-range interactions across distant regions and potentially leading to fragmented layouts. To mitigate this issue, we introduce window-scaled extrapolation, which enlarges windows responsible for long-range scene coordination while keeping refinement windows close to their adapted receptive fields. This preserves fine-detail synthesis while restoring sufficient global context for semantic continuity at native 4K resolution. To better align inference with pretrained prior, we additionally apply NTK-aware RoPE scaling at the target resolution in spatial axes, *i.e.*, $u \in \{h, w\}$:

$$\theta_m^{(u)\prime} = p_u \cdot (\beta'_u)^{-2m/D_u},\ \beta'_u = \beta \cdot s^{D_u/(D_u-2)}, \tag{8}$$

where s denotes the ratio between the target resolution and the pretrained 720p resolution.

Importantly, window-scaled extrapolation requires no additional training. Built on neighborhood attention, ApertureAttn maintains a token-centered interaction pattern across window sizes. Enlarging the window only expands the set of neighboring tokens included in the softmax, while preserving the underlying spatial organization. In contrast, fixed block partitioning shifts window boundaries as the scale changes, thereby altering the relative position of each token within its attention region. Moreover, adapted parameters such as LoRA operate on token-wise feature transformations, whereas the attention window only determines which tokens interact. This decoupling enables the learned adaptation to generalize accordingly when the window size is increased at inference time.

2.4 Proximity-Guided Temporal Attention

While window scaling effectively resolves block-level fragmentation, we observe a subtler failure mode at 4K: visually coherent frames may still contain duplicated subjects at distant spatial locations,

Table 1: The evaluation results of baselines and our proposed method on VBench. Here, $^+$ means applying NTK RoPE scaling, proportional attention, and a larger timestep shift to the original Wan2.2.

Method/Setting	Aesthetic Quality	Background Consistency	Imaging Quality	Motion Smoothness	Overall Consistency	Subject Consistency
Original Wan2.2-TI2V-5B	42.02%	98.72%	39.79%	99.42%	17.07%	97.33%
Wan2.2-TI2V-5B$^+$	57.45%	98.94%	61.78%	98.81%	23.58%	96.80%
SDEdit	60.16%	98.53%	57.87%	98.17%	26.77%	96.26%
RealViformer	61.55%	98.48%	62.69%	97.80%	26.89%	96.02%
FlashVSR	62.08%	98.37%	65.60%	97.73%	26.70%	95.72%
Direct Tune	55.77%	98.70%	56.25%	99.09%	22.70%	98.09%
UltraWan	57.32%	98.73%	56.63%	99.23%	24.46%	98.09%
T3-Video	59.63%	98.39%	65.62%	98.54%	25.66%	96.60%
ApertureAttn (Ours)	**66.62%**	**99.24%**	**66.94%**	**98.84%**	**26.82%**	**97.72%**

such as two nearly identical pedestrians on opposite sides of a street. Unlike block artifacts, these frames remain spatially continuous and well-rendered, yet contain redundant instances that correspond to the same semantic entity. We attribute this to insufficient global semantic coordination along the temporal axis. The fact that frames remain spatially continuous indicates that intra-frame coverage is already sufficient. The duplication instead originates from the cross-frame attention pattern: when distant frames are connected only through localized cross-frame windows, each window pair forms an isolated coordination region. Within such isolated regions, distant spatial locations can independently commit to plausible subject placements that satisfy the prompt at a local level, leading to redundant instances of the same semantic subject across the canvas.

To address this issue, we propose proximity-guided temporal attention (PGTAttn), as shown in Fig. 3, motivated by the temporal proximity inherent in video. Adjacent frames usually share highly similar content and involve small spatial displacement, for which a tight local attention window is sufficient and helps reduce interference from irrelevant regions. As temporal distance increases, spatial displacement becomes larger, making broader windows necessary to capture meaningful correspondences. Based on this observation, we scale the cross-frame spatial attention window with temporal distance at inference time. Formally, for the token with spatiotemporal position (i, j, k), the attention mask can be defined as follows:

$$\mathcal{N}^{(l)}(i,j,k) = \left\{ (i',j',k') : |i'-i| \le R_t^{(m_l)},\ |j'-j| \le \tilde{R}_h^{(m_l,\delta)},\ |k'-k| \le \tilde{R}_w^{(m_l,\delta)} \right\}, \tag{9}$$

where $\delta = |i' - i|$ is denoted as the temporal distance. Here, when m_l is not 0:

$$\tilde{R}^{(m_l,\delta)} = R^{\pi(m_l,\delta)},\ \pi(m_l,\delta) = \max(m_l - \phi(\delta), 1),\ \phi(\delta) = \sum_{t=1}^{T} \mathbb{I}[\delta > \tau_t], \tag{10}$$

$\mathbf{T} = \{\tau_1, \tau_2, \ldots, \tau_T\}$ is the monotonically increasing sequence of frame distance thresholds. To further improve generation quality, we adopt proportional attention to sharpen the attention distribution and counteract detail dilution. We also increase the timestep shift in the flow-matching scheduler to allocate more denoising steps to the high-noise regime.

3 Experiments

3.1 Settings and Implementation Details

In this paper, we build on the open-source Wan2.2-TI2V-5B [35] model to demonstrate the effectiveness of our method. We curate 3K synthetic 2K-resolution images generated by FLUX1.1 [Pro] Ultra for finetuning. Our method requires only 30 epochs of LoRA fine-tuning on 4 H100 GPUs with a batch size of 4, taking less than 8 hours in total. During finetuning, we set $g = 5$ and use window-size schedule of $[\text{Full}, 32, 24, 20, 16]$ for ApertureAttn. At 4K inference, we adopt window-scaled extrapolation and expand window-size schedule to $[\text{Full}, 64, 40, 24, 16]$. The cross-frame window size schedule of Proximity-Guided Temporal Attention (PGTAttn) starts from the same-frame window size of the current layer, and progressively expands to larger values in ApertureAttn window-size schedule ($[64, 40, 24, 16]$) as the frame distance crosses the thresholds $\mathbf{T} = \{\tau_1 = 2, \tau_2 = 5, \tau_3 = 10\}$.



Figure 4: Qualitative comparison with baseline methods. Our method produces more coherent native 4K videos with higher-fidelity details and fewer artifacts.

To demonstrate the effectiveness of our method, we conduct comparisons against existing state-of-the-art native 4K generation methods, UltraWan [39], UltraGen [17] and T3-Video [45]. We also compare our method with representative multi-stage approaches, including SDEdit [24], state-of-the-art video super-resolution methods, FlashVSR [54] and RealViFormer [49]. For evaluation, we adopt VBench [18] as the benchmark and generate one video for each prompt with seed 0. All generated videos contain 81 frames with a resolution of 3840×2176 (4K). For more implementation and evaluation details, please refer to the Sec. A in supplementary.

3.2 Baseline Comparison

We compare our method with both multi-stage pipelines, including SDEdit, FlashVSR, and RealViFormer, and native high-resolution video generation methods, including UltraWan and T3-Video. We also report two controlled baselines: Direct-Tune, which fine-tunes Wan2.2-TI2V-5B directly on our image data, and Wan2.2-TI2V-5B$^+$, which applies inference-time modifications, including NTK-aware RoPE scaling, proportional attention, and timestep-shift adjustment. Quantitative and qualitative results are shown in Table 1 and Fig. 4. Our method achieves strong performance in both visual fidelity and temporal coherence. It attains high aesthetic and imaging quality, surpassing super-resolution-based baselines designed for visual enhancement. Meanwhile, its favorable consistency

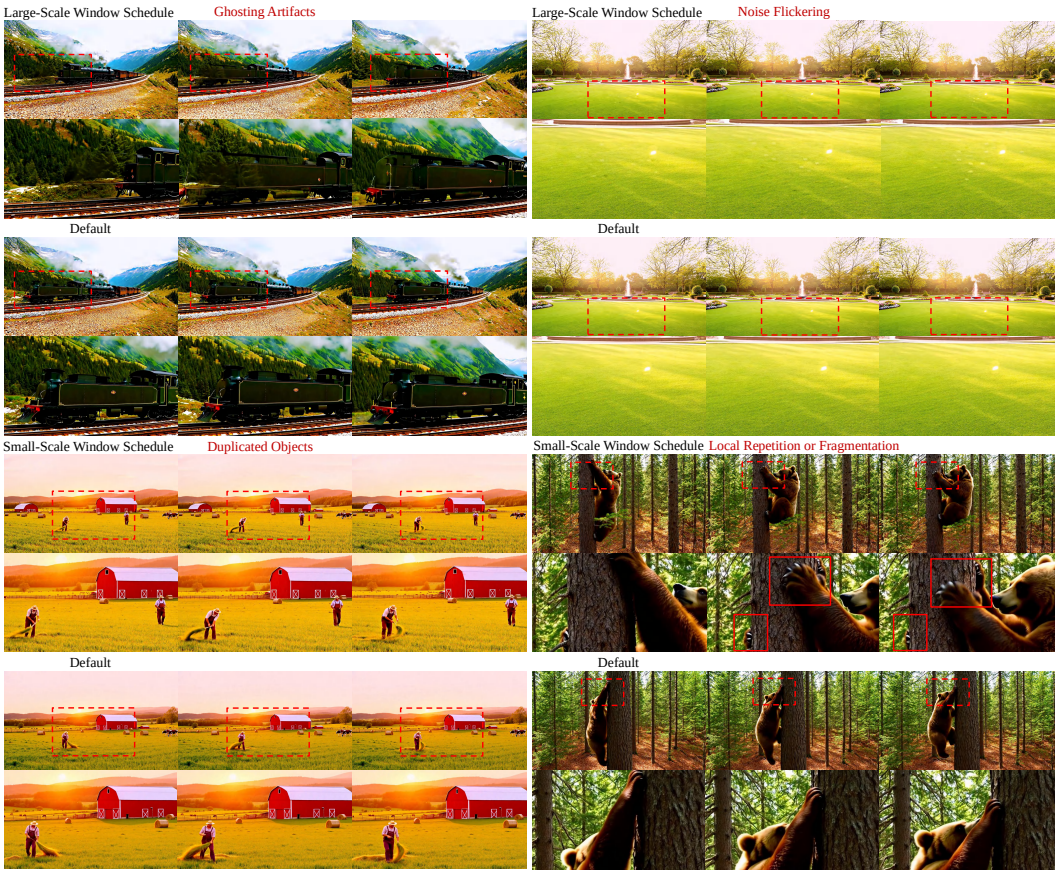

Figure 5: Ablation study on window-size schedule for ApertureAttn.

and motion smoothness demonstrate that image-only supervision can support detailed native 4K video generation without compromising temporal consistency.

3.3 Ablation Study

Impact on Window-Size Schedule for ApertureAttn. We conduct ablation study on the ApertureAttn window-size schedule. We use $[\text{Full}, 48, 32, 24, 16]$ as the small-scale schedule and $[\text{Full}, 64, 48, 40, 32]$ as the large-scale schedule. The default setting, consistent with our main experiments, is $[\text{Full}, 64, 40, 24, 16]$. All variants use the same proximity-guided temporal attention configuration as the default model. The results are shown in Fig. 5. As shown in the visualizations, small windows provide insufficient intra-frame global context and cross-frame temporal correspondence, leading to duplicated objects, fragmented body parts, and artifacts in window-crossing motion. Increasing the window size alleviates these issues by providing broader spatiotemporal context. However, excessively large windows make attention less focused on temporally relevant cues and introduce interference from distant frames, often resulting in ghosting or stretching artifacts. Moreover, the increased modeling burden may exceed the model effective generation capacity, leading to degraded fine details and stronger temporal flickering.

Impact on Cross-Frame Window Configuration for PGTAttn. We ablate the cross-frame window configuration of PGTAttn along two dimensions, initial cross-frame window size and the temporal expansion schedule. For initial cross-frame window, we set $R^{(m_l)}$-anchored starts from the same-frame window size, and $R^{(m_l)}$-skipped starts from the next larger window in the schedule. Temporal expansion schedule is defined by the threshold set $\mathbf{T}$. We compare a short-tier configuration, $\tau_1 = 2, \tau_2 = 5, \tau_3 = 10$, which reaches larger windows more quickly, with a long-tier configuration, $\tau_1 = 5, \tau_2 = 10, \tau_3 = 15$, which expands more gradually. The evaluation results are shown in Table 2. The ApertureAttn window schedule is default as main experiments. From the results, PGTAttn



Figure 6: Ablation study on each component of our method. Components are progressively added from left to right, illustrating the contribution of each design choice to native 4K generation quality.

Table 2: The results of different cross-frame window configurations for PGTAttn on VBench.

Method/Setting	Aesthetic Quality	Background Consistency	Imaging Quality	Motion Smoothness	Overall Consistency	Subject Consistency
$R^{(m_l)}$-Anchored+Short-Tier	66.62%	99.24%	66.94%	98.84%	26.82%	97.72%
$R^{(m_l)}$-Anchored+Long-Tier	66.02%	99.08%	67.08%	98.84%	26.76%	97.68%
$R^{(m_l)}$-Skipped+Short-Tier	66.25%	99.07%	67.17%	98.83%	26.75%	97.74%
$R^{(m_l)}$-Skipped+Long-Tier	66.75%	99.06%	67.12%	98.82%	26.67%	97.76%

performs consistently across different cross-frame window allocations. Using more small windows slightly improves aesthetic quality through finer local modeling, but may mildly reduce consistency due to weaker long-range interaction.

Impact on Each Component. We further conduct component-wise ablation study of our method, with the results shown in Fig. 6. Directly applying Wan2.2 to native 4K video generation leads to severe degradation. The sharp increase in token count pushes the spatial position indices beyond the range seen during training, causing RoPE extrapolation failure and inaccurate positional modeling. Meanwhile, attention operates over a much larger token set produces overly diffuse distributions and weakens fine-detail fidelity. We first apply several training-free adaptations, including NTK-aware spatial RoPE scaling, proportional attention, and a larger timestep shift in the flow-matching scheduler to allocate more denoising steps to the high-noise regime. As shown in the second group of example, these modifications partially recover local details, but the results remain noisy and often suffer from duplicated subjects, fragmented structures, and poor global coherence. We then introduce ApertureAttn with image-only supervision, brings a clear improvement in detail synthesis, as illustrated by the third group of example. However, the relatively small window schedule used at this stage still limits long-range context, resulting in slight block-level fragmentation. Next, we introduce window-scaled extrapolation, which restores sufficient long-range spatial context at 4K and largely resolves these issues while keeping fine-grained details. However, a subtler failure mode remains: frames may appear globally coherent while still containing semantically duplicated subjects at distant spatial locations. Finally, we introduce proximity-guided temporal attention, which further improves cross-frame consistency and suppresses such semantic duplication, leading to more coherent motion and faithful 4K generation.

4 Related Works

High-resolution visual generation has been studied in both image and video synthesis. For image generation, existing methods can be broadly categorized into multi-stage generation and native high-resolution generation. Multi-stage methods first synthesize low-resolution images or intermediate results, and then progressively enhance [7, 37, 27, 19, 30, 8, 10, 53, 26, 28, 51] and integrate [40, 22] them into ultra-high-resolution outputs. In contrast, native high-resolution generation directly

produces high-resolution images without an explicit low-to-high-resolution cascade. This direction includes inference-time extrapolation methods [47, 31, 14, 12, 34], which extend pretrained diffusion models beyond their original training resolution, and training-based methods [46, 43, 42, 4, 38, 50, 29, 6], which train or adapt models directly at the target resolution.

Compared with high-resolution image generation, high-resolution video generation faces additional constraints from temporal consistency and the rapidly growing spatiotemporal token space. To mitigate these constraints, some high-resolution video generation methods [36, 32, 44, 15, 2, 48] adopt cascaded pipelines, where low-resolution videos are progressively refined or upsampled to higher resolutions. Recently, several methods have explored native ultra-high-resolution video generation. UltraWan [39] constructs a large-scale UHD-4K video dataset and uses it to finetune UltraWan-4K for native high-resolution video synthesis. UltraGen [17] adapts pretrained low-resolution video models with hierarchical attention for native 4K generation, and T3-Video [45] retrofits pretrained full-attention transformers for efficient 4K generation. Existing native high-resolution methods are generally adapted with high-resolution video data. Our method instead learns high-resolution spatial priors from images only, enabling native 4K video generation without any video training data.

5 Conclusion

In this paper, we address the key challenges of native 4K video generation. We propose ApertureAttn, which enables high-resolution spatial adaptation using only image supervision. By decomposing high-resolution generation into scale-aligned windowed computations, ApertureAttn allows pretrained video models to learn high-fidelity spatial structures and fine-grained details while preserving temporal priors. Further, we introduce window-scaled extrapolation and proximity-guided temporal attention to improve spatiotemporal consistency over the enlarged 4K canvas. Extensive experiments demonstrate that our method achieves high-quality native 4K video generation with strong visual fidelity, temporal coherence, and substantially reduced training cost.

Acknowledgments

This project is supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2026, WBS: A-8004345-00-00), the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office (Award: CRPO-GC1-NTU-002), as well as the National Natural Science Foundation of China (Grant No. 62606320).

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[2] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[3] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.

[4] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[5] Junsong Chen, Yuyang Zhao, Jincheng Yu, Ruihang Chu, Junyu Chen, Shuai Yang, Xianbang Wang, Yicheng Pan, Daquan Zhou, Huan Ling, et al. Sana-video: Efficient video generation with block linear diffusion transformer. *arXiv preprint arXiv:2509.24695*, 2025.

[6] Junyu Chen, Wenkun He, Yuchao Gu, Yuyang Zhao, Jincheng Yu, Junsong Chen, Dongyun Zou, Yujun Lin, Zhekai Zhang, Muyang Li, et al. Dc-videogen: Efficient video generation with deep compression video autoencoder. *arXiv preprint arXiv:2509.25182*, 2025.

[7] Ruoyi Du, Dongliang Chang, Timothy Hospedales, Yi-Zhe Song, and Zhanyu Ma. Demofusion: Democratising high-resolution image generation with no $$$. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6159–6168, 2024.

[8] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. *arXiv preprint arXiv:2410.07536*, 2024.

[9] Yu Gao, Haoyuan Guo, Tuyen Hoang, Weilin Huang, Lu Jiang, Fangyuan Kong, Huixia Li, Jiashi Li, Liang Li, Xiaojie Li, et al. Seedance 1.0: Exploring the boundaries of video generation models. *arXiv preprint arXiv:2506.09113*, 2025.

[10] Lanqing Guo, Yingqing He, Haoxin Chen, Menghan Xia, Xiaodong Cun, Yufei Wang, Siyu Huang, Yong Zhang, Xintao Wang, Qifeng Chen, et al. Make a cheap scaling: A self-cascade diffusion model for higher-resolution adaptation. In *European conference on computer vision*, pages 39–55. Springer, 2024.

[11] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

[12] Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. Elasticdiffusion: Training-free arbitrary size image generation through global-local content separation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6603–6612, 2024.

[13] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6185–6194, 2023.

[14] Yingqing He, Shaoshu Yang, Haoxin Chen, Xiaodong Cun, Menghan Xia, Yong Zhang, Xintao Wang, Ran He, Qifeng Chen, and Ying Shan. Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2023.

[15] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

[17] Teng Hu, Jiangning Zhang, Zihan Su, and Ran Yi. Ultragen: High-resolution video generation with hierarchical attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4923–4931, 2026.

[18] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.

[19] Younghyun Kim, Geunmin Hwang, Junyu Zhang, and Eunbyung Park. Diffusehigh: Training-free progressive high-resolution image synthesis through structure guidance. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 4338–4346, 2025.

[20] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[21] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[22] Zhihang Lin, Mingbao Lin, Meng Zhao, and Rongrong Ji. Accdiffusion: An accurate method for higher-resolution image generation. In *European Conference on Computer Vision*, pages 38–53. Springer, 2024.

[23] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.

[24] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[25] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[26] Haonan Qiu, Ning Yu, Ziqi Huang, Paul Debevec, and Ziwei Liu. Cinescale: Free lunch in high-resolution cinematic visual generation. *arXiv preprint arXiv:2508.15774*, 2025.

[27] Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16893–16903, 2025.

[28] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra high-resolution image synthesis to new peaks. *Advances in Neural Information Processing Systems*, 37:111131–111171, 2024.

[29] Jingjing Ren, Wenbo Li, Zhongdao Wang, Haoze Sun, Bangzhen Liu, Haoyu Chen, Jiaqi Xu, Aoxue Li, Shifeng Zhang, Bin Shao, et al. Turbo2k: Towards ultra-efficient and high-quality 2k video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18155–18165, 2025.

[30] Shuwei Shi, Wenbo Li, Yuechen Zhang, Jingwen He, Biao Gong, and Yinqiang Zheng. Resmaster: Mastering high-resolution image generation via structural and fine-grained guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6887–6895, 2025.

[31] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. Freeu: Free lunch in diffusion u-net. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4733–4743, 2024.

[32] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[33] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

[34] Tobias Vontobel, Seyedmorteza Sadat, Farnood Salehi, and Romann Weber. Hiwave: Training-free high-resolution image generation via wavelet-based diffusion sampling. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pages 1–11, 2025.

[35] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[36] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 133(5):3059–3078, 2025.

[37] Haoning Wu, Shaocheng Shen, Qiang Hu, Xiaoyun Zhang, Ya Zhang, and Yanfeng Wang. Megafusion: Extend diffusion models towards higher-resolution image generation without further tuning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3944–3953. IEEE, 2025.

[38] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, et al. Sana: Efficient high-resolution image synthesis with linear diffusion transformers. *arXiv preprint arXiv:2410.10629*, 2024.

[39] Zhucun Xue, Jiangning Zhang, Teng Hu, Haoyang He, Yinan Chen, Yuxuan Cai, Yabiao Wang, Chengjie Wang, Yong Liu, Xiangtai Li, et al. Ultravideo: High-quality uhd video dataset with comprehensive captions. *arXiv preprint arXiv:2506.13691*, 2025.

[40] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and Bin Cui. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. *arXiv preprint arXiv:2401.11708*, 2024.

[41] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

[42] Tian Ye, Song Fei, and Lei Zhu. Ultraflux: Data-model co-design for high-quality native 4k text-to-image generation across diverse aspect ratios. *arXiv preprint arXiv:2511.18050*, 2025.

[43] Ruonan Yu, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Ultra-resolution adaptation with ease. *arXiv preprint arXiv:2503.16322*, 2025.

[44] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, 133(4):1879–1893, 2025.

[45] Jiangning Zhang, Junwei Zhu, Teng Hu, Yabiao Wang, Donghao Luo, Weijian Cao, Zhenye Gan, Xiaobin Hu, Zhucun Xue, and Chengjie Wang. Transform trained transformer: Accelerating naive 4k video generation over 10×. *arXiv preprint arXiv:2512.13492*, 2025.

[46] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23464–23473, 2025.

[47] Shen Zhang, Zhaowei Chen, Zhenyu Zhao, Yuhao Chen, Yao Tang, and Jiajun Liang. Hidiffusion: Unlocking higher-resolution creativity and efficiency in pretrained diffusion models. In *European Conference on Computer Vision*, pages 145–161. Springer, 2024.

[48] Shilong Zhang, Wenbo Li, Shoufa Chen, Chongjian Ge, Peize Sun, Yifu Zhang, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Ping Luo. Flashvideo: Flowing fidelity to detail for efficient high-resolution video generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12735–12743, 2026.

[49] Yuehan Zhang and Angela Yao. Realviformer: Investigating attention for real-world video super-resolution. In *European conference on computer vision*, pages 412–428. Springer, 2024.

[50] Zhengqiang Zhang, Ruihuang Li, and Lei Zhang. Frecas: Efficient higher-resolution image generation via frequency-aware cascaded sampling. *arXiv preprint arXiv:2410.18410*, 2024.

[51] Chen Zhao, Jiawei Chen, Hongyu Li, Zhuoliang Kang, Shilin Lu, Xiaoming Wei, Kai Zhang, Jian Yang, and Ying Tai. Luve: Latent-cascaded ultra-high-resolution video generation with dual frequency experts. *arXiv preprint arXiv:2602.11564*, 2026.

[52] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[53] Yifan Zhu, Guanting Chen, Bing Wei, and Haoran Luo. Hiflow: Hierarchical feedback-driven optimization for constrained long-form text generation. *arXiv preprint arXiv:2603.04996*, 2026.

[54] Junhao Zhuang, Shi Guo, Xin Cai, Xiaohui Li, Yihao Liu, Chun Yuan, and Tianfan Xue. Flashvsr: Towards real-time diffusion-based streaming video super-resolution. *arXiv preprint arXiv:2510.12747*, 2025.

A More Implementation and Evaluation Details

A.1 More Architecture Details

In this paper, we adopt the open-source Wan2.2-TI2V-5B [35] model to demonstrate the effectiveness of our method. Wan2.2-TI2V-5B is pretrained for 720p video generation, our proposed method enables native 4K synthesis without using any video data. For finetuning, we set $g = 5$ and use the ApertureAttn schedule $[\text{Full}, 32, 24, 20, 16]$. For 4K inference, we apply window-scaled extrapolation, expanding the schedule to $[\text{Full}, 64, 40, 24, 16]$. For Proximity-Guided Temporal Attention (PGTAttn), the cross-frame window size schedule starts from the same-frame window size of the current layer, and progressively grows to larger values in ApertureAttn window-size schedule ($[64, 40, 24, 16]$) when the frame distance exceeds $\mathbf{T} = \{\tau_1 = 2, \tau_2 = 5, \tau_3 = 10\}$. For instance, with $R_h^{(m_l)} = R_w^{(m_l)} = 16$, the cross-frame window grows as $16 \to 24 \to 40 \to 64$ across the four frame-distance tiers. The intra- and cross-frame window schedules are shown in Table 3. We also adopt proportional attention during inference, and set 4.0 for NTK-aware RoPE scaling, 10.0 for time shifting during inference.

Table 3: The intra- and cross-frame window size schedule within transformer block group during both finetuning and inference.

Layer	Intra-Frame Window Size	Cross-Frame Window Size			
		$\delta \in [0, \tau_1]$	$\delta \in (\tau_1, \tau_2]$	$\delta \in (\tau_2, \tau_3]$	$\delta \in (\tau_3, f]$
Finetuning					
$m_l = 0$	Full	Full	Full	Full	Full
$m_l = 1$	$R_h^{(m_l)} = R_w^{(m_l)} = 32$	$R_h^{(m_l)} = R_w^{(m_l)} = 32$	$R_h^{(m_l)} = R_w^{(m_l)} = 32$	$R_h^{(m_l)} = R_w^{(m_l)} = 32$	$R_h^{(m_l)} = R_w^{(m_l)} = 32$
$m_l = 2$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$
$m_l = 3$	$R_h^{(m_l)} = R_w^{(m_l)} = 20$	$R_h^{(m_l)} = R_w^{(m_l)} = 20$	$R_h^{(m_l)} = R_w^{(m_l)} = 20$	$R_h^{(m_l)} = R_w^{(m_l)} = 20$	$R_h^{(m_l)} = R_w^{(m_l)} = 20$
$m_l = 4$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$
Inference					
$m_l = 0$	Full	Full	Full	Full	Full
$m_l = 1$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$
$m_l = 2$	$R_h^{(m_l)} = R_w^{(m_l)} = 40$	$R_h^{(m_l)} = R_w^{(m_l)} = 40$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$
$m_l = 3$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 40$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$
$m_l = 4$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$	$R_h^{(m_l)} = R_w^{(m_l)} = 16$	$R_h^{(m_l)} = R_w^{(m_l)} = 24$	$R_h^{(m_l)} = R_w^{(m_l)} = 40$	$R_h^{(m_l)} = R_w^{(m_l)} = 64$

A.2 More Finetuning Details

For training data, we curate 3K synthetic 2K-resolution images generated by FLUX1.1 [Pro] Ultra for finetuning. We set $g = 5$ and use the ApertureAttn schedule $[\text{Full}, 32, 24, 20, 16]$ during finetuning. Specifically, we use LoRA with rank 32 and apply it to the attention projections and feed-forward layers, including q, k, v, o, ffn.0, and ffn.2. During training, we disable NTK RoPE scaling, proportional attention, and time-shift adjustment. The learning rate is set to 1×10^{-4}. Our method requires only 30 epochs of LoRA fine-tuning on 4 H100 GPUs with a batch size of 4, taking less than 8 hours in total.

A.3 More Evaluation Details

To demonstrate the effectiveness of our method, we conduct comparisons against existing state-of-the-art native 4K generation methods, UltraWan [39], UltraGen [17] and T3-Video [45]. We also compare our method with representative multi-stage approaches, including SDEdit [24], state-of-the-art video super-resolution methods, FlashVSR [54] and RealViFormer [49]. We also report two controlled baselines: Direct-Tune and Wan2.2-TI2V-5B$^+$. Diret-Tune fine-tunes Wan2.2-TI2V-5B directly on our image data. Wan2.2-TI2V-5B$^+$ applies inference-time modifications, including NTK-aware RoPE scaling (set 4.0 as ours), proportional attention, and timestep-shift adjustment (set 10.0 as ours). Specifically, for multi-stage pipelines, we use Wan2.2-TI2V-5B to generate low-resolution 720p videos (1280×704). For SDEdit, the generated videos are directly resized to 4K (3840×2176), and generation is performed with denoising strength of 0.7. For super-resolution methods, to ensure an exact $4\times$ upscaling factor, we first resize the 720p videos to 960×544 and then apply $4\times$ upscaling. Here, we use FlashVSR-v1.1-Full as the FlashVSR super resolution model. For evaluation, we adopt VBench [18] as the benchmark and generate one video for each prompt with seed 0. All

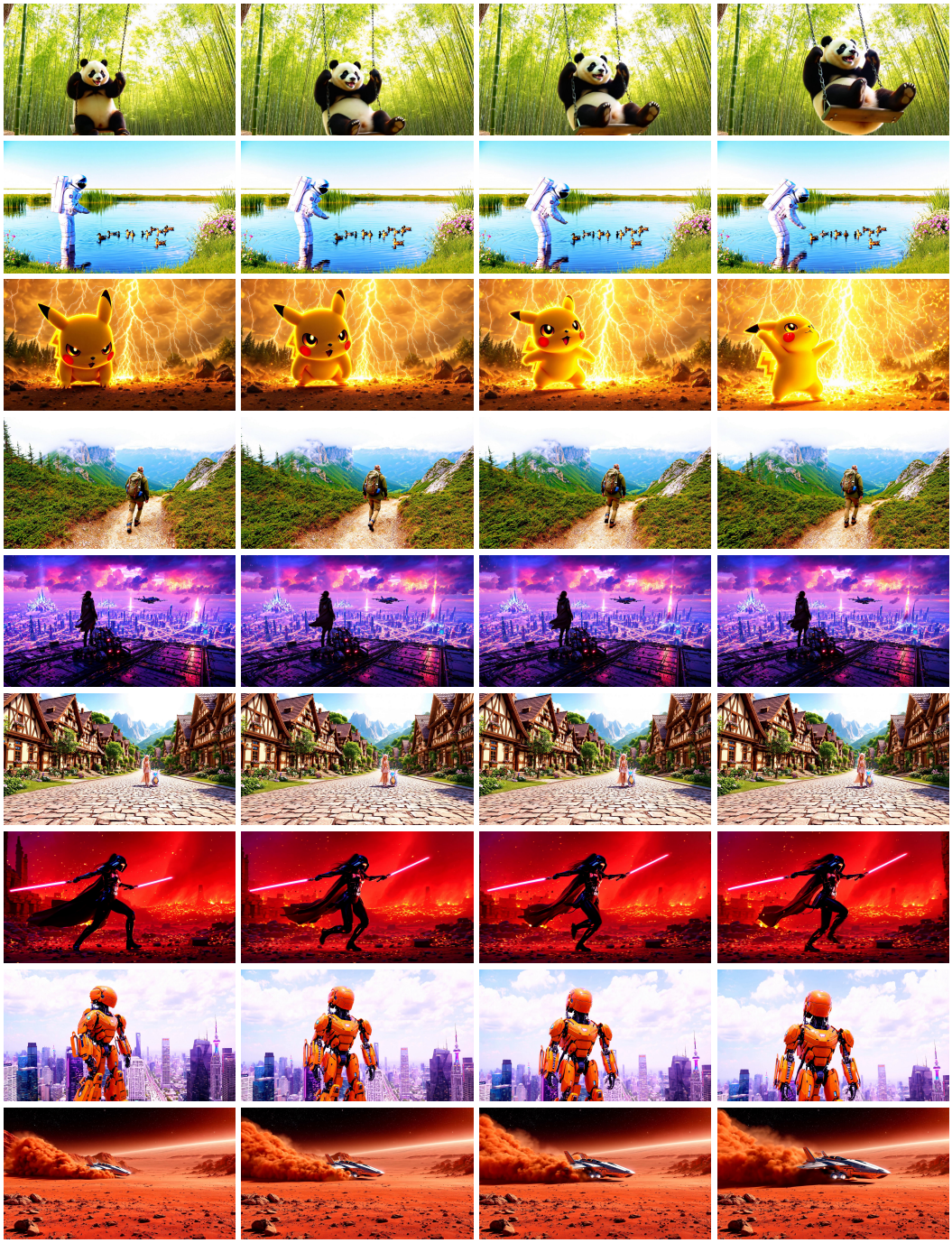

Figure 7: More visualization results (3840×2176) of our proposed method.

generated videos contain 81 frames with a resolution of 3840×2176 (4K). For VBench evaluation, we evaluate six dimensions: aesthetic quality, background consistency, imaging quality, motion smoothness, overall consistency, and subject consistency. These metrics provide a comprehensive evaluation of video generation quality, covering both frame-level visual fidelity and cross-frame temporal consistency.

B Broader Impacts

Our work enables native 4K video generation under modest training computational costs and data requirements, with several positive implications for the broader community. First, it lowers the training barrier for high-resolution video synthesis, allowing researchers and creators without access to massive 4K video corpora or large-scale training infrastructure to produce content at standards matching modern displays. Second, the ability to generate content at cinematic resolutions opens new avenues for creative applications, including film pre-visualization, storyboard prototyping, and concept development for media production, where high spatial fidelity is essential for conveying visual intent. Third, by demonstrating that high-resolution images can substitute for high-resolution videos during training, our approach reduces reliance on the costly and time-consuming process of curating large-scale 4K video datasets, making high-resolution generation research more accessible. Finally, image-only training substantially reduces the storage and training resource demands compared to video-based pipelines, since high-resolution images require far less disk space, I/O bandwidth, and memory costs.

C Limitations

Due to limited computational resources, our experiments are mainly conducted on the 5B-parameter Wan2.2-TI2V-5B model, using a training set of 3K high-quality, 2K-resolution images generated by the FLUX1.1 [Pro] Ultra model. While the experimental results demonstrate the effectiveness of the proposed method under this setting, its generalizability to larger models, such as Wan2.2-T2V-A14B, larger-scale datasets, and more diverse data distributions, such as synthetic data generated by other high-quality image generation models, remains to be further investigated.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully discloses all the information needed to reproduce the main experimental results of the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The paper provides open access to the code.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies all the training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports the information about the statistical significance of the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper provides sufficient information on the computer resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the societal impacts in Appendix.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets used in the paper have been properly credited, and the license is explicitly mentioned in the original credited asset, and the terms of use have been properly respected.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: New assets introduced in the paper are well documented and the documentation provided alongside the assets

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.