AtlasVid: Efficient Ultra-High-Resolution Long Video Generation via Decoupled Global-Local Modeling

Ziyang Mai[*] Dartmouth College

Yuyao Zhang[*] Dartmouth College

Yu-Wing Tai Dartmouth College

Project page: `https://ziyang1106.github.io/AtlasVid/`

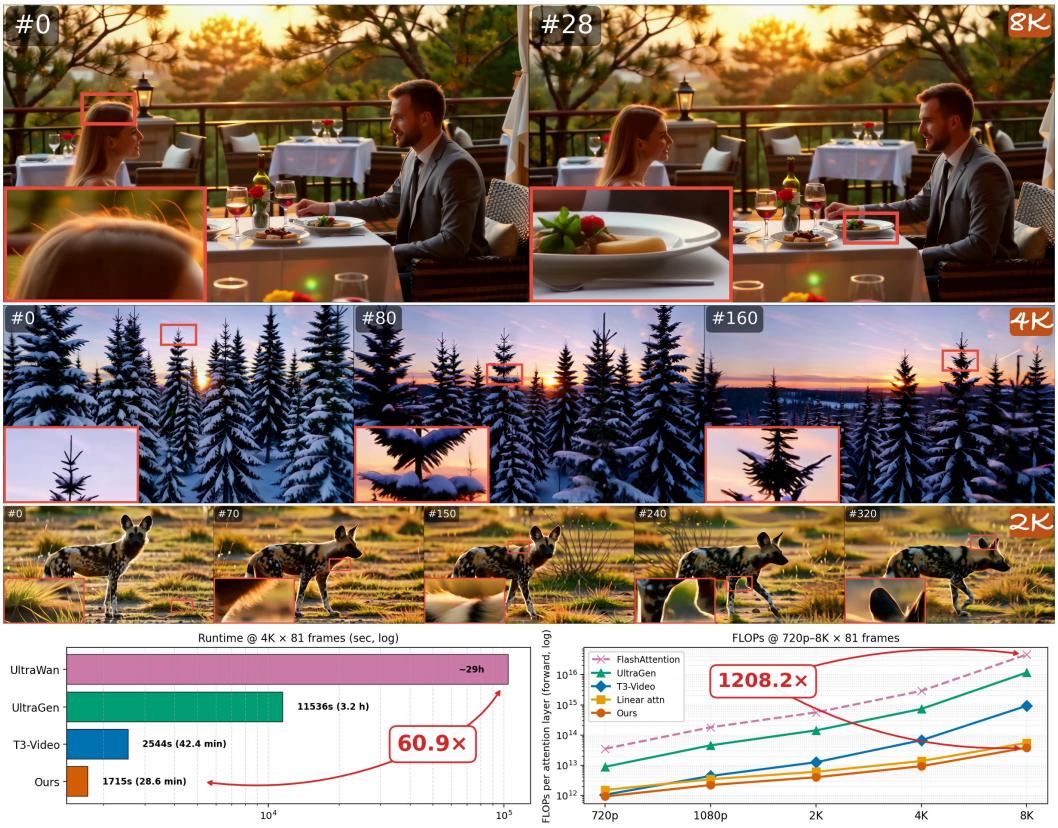

Figure 1: **AtlasVid** enables the generation of ultra-high-resolution and long-duration videos in different settings, including 8K 29 frames, 4K 161 frames and 2K 321 frames. Frame index indicated in the top-left corner and the output resolution in the top-right corner. **Bottom:** AtlasVid runs 60.9× faster than UltraWan at 4K × 81 frames (left) and reduces per-layer attention FLOPs by up to 1208.2× over FlashAttention from 720p to 8K (right).

Abstract

Recent diffusion-based video generators have achieved remarkable visual fidelity and prompt controllability, yet scaling them to ultra-high-resolution (UHR) long videos remains prohibitively expensive. The difficulty is especially pronounced for

[*]Co-first authors, equal contribution.

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

long single-shot generation where a continuous scene must preserve global temporal coherence, and fine-grained spatial details without relying on clip transitions or autoregressive shot stitching. In this work, we revisit this challenge from the perspective of decoupled modeling. We argue that existing video diffusion models already encode strong local visual priors, while the main bottleneck lies in efficiently extending global spatiotemporal modeling as resolution and duration increase. Based on this insight, we propose **AtlasVid**, a decoupled global-local framework for efficient UHR long video generation. **AtlasVid** first generates a low-resolution and low-FPS global semantic proxy via temporally scaled RoPE, thereby extending the temporal horizon without increasing the training token count. Guided by this proxy, a high-resolution detail branch performs joint denoising with hierarchical locality-preserving attention. Reordered spatiotemporal windows preserve geometric locality and asymmetric global-local attention injects aligned semantic guidance and preserves the model's pretrained ability. This design enables resolution-agnostic training: the model is trained only at 720P with lightweight LoRA adaptation, yet generalizes directly to 4K and beyond for longer (>10s) video synthesis. Experiments show that **AtlasVid** substantially improves the efficiency of ultra-high-resolution long video generation, achieving high-quality UHR long video generation with $60.9\times$ speed up and significantly less training cost and even better performance than native 4K video generators.

1 Introduction

The field of video generation has advanced rapidly, driven by Diffusion Models [Ho et al., 2020] and the Diffusion Transformer (DiT) architecture [Peebles and Xie, 2023]. Recent text-to-video (T2V) systems [Arkhipkin et al., 2025, Chen et al., 2025a, Li et al., 2026, Kong et al., 2024, Team, 2025, Lin et al., 2024, HaCohen et al., 2024, 2026, Ma et al., 2025a, Wan et al., 2025] have significantly improved realism and visual fidelity, enabling compelling video synthesis from text or image prompts. As these models mature, the focus is shifting from prompt following and visual quality toward high-resolution, long-duration generation, motivated by applications in filmmaking, television, and professional content creation, where fine-grained spatial detail and long-range temporal coherence are both critical. To meet this demand, industrial systems [OpenAI, 2025, Google DeepMind, 2025, Team Seedance et al., 2026, Kling AI, 2026] rely on massive data and computation to support generation up to 2K resolution, typically in a clip-by-clip autoregressive manner. However, each clip is usually limited to less than 5 seconds.

Meanwhile, academic works have explored long-video streaming generation [Yin et al., 2024b, Huang et al., 2025, Chen et al., 2026, Teng et al., 2025] and planning-based methods for extended video creation [Huang et al., 2024a, Zheng et al., 2024, Zhao et al., 2024, Guo et al., 2025], yet genuinely long single-shot generation remains out of reach.

In real-world scenarios, however, high-resolution long single-shot videos are often required, such as cinematic long takes, music and dance performances, and documentary-style footage.

A major challenge is the inherent complexity of the DiT architecture, whose full-attention mechanism scales quadratically with the spatiotemporal input size, i.e., $O((HWT)^2)$, where H, W, and T denote the height, width, and temporal length of the video. Consequently, doubling all three dimensions increases the computational cost by $64\times$.

Moreover, directly training models at ultra-high resolution can introduce additional and unpredictable difficulties such as severe memory pressure, training instability at extremely long token sequences, and the scarcity of native 4K long-video training data. Existing methods [He et al., 2024, Xie et al., 2025, Zhuang et al., 2025] alleviate this issue with a "generation-then-super-resolution" pipeline. However, this "pseudo high-resolution" paradigm mainly improves sharpness and often fails to recover sufficiently rich high-frequency details. Another line of work [Chen et al., 2025b, Zhang et al., 2025c, Wang et al., 2025] improves efficiency through sparse or linearized attention mechanisms, although faster, they do not fundamentally resolve the substantial data bottleneck of high-resolution long-video generation.

These limitations prompt a rethinking of what fundamentally restricts high-resolution long-video generation. We hypothesize that pretrained video generation models already possess much of the

Table 1: Training resource comparison with state-of-the-art high-resolution video generation methods.

Method	GPUs	Training Stage	Max Frames	Max Reso.	Duration
UltraWan-4K [Xue et al., 2025]	7.6K H20-hours	LoRA on Wan-1.3B	29f	4K	$\sim$1.2s
UltraGen [Hu et al., 2026]	32$\times$H20	4K direct fine-tuning	29f	4K	$\sim$1.2s
T3-Video [Zhang et al., 2025a]	64	720P $\rightarrow$ 4K (two-stage)	81f	4K	$\sim$3.4s
AtlasVid (Ours)	**2$\times$RTX PRO 6000**	**720P (+ optional 4K)**	**321f**	**$>$4K**	**$\sim$13.4s**

visual knowledge required to synthesize plausible fine-grained details at higher resolutions, since pretraining exposes them to similar objects, structures, and scenes across different visual scales. From this perspective, the main bottleneck is not an inherent lack of high-resolution generation capacity, but the difficulty of modeling long-range dependencies efficiently as the spatiotemporal resolution grows.

Motivated by this view, we propose a hierarchical locality-preserving attention mechanism for efficient high-resolution long-video generation. Our design decouples the modeling of local neighborhood structure from global semantics, allowing the model to scale more effectively across resolutions. Furthermore, we introduce a temporal-scale RoPE strategy for long-video modeling, which enlarges the temporal modeling range without increasing the token count. By combining hierarchical locality-preserving attention with temporal-scale positional modeling, our framework enables efficient ultra-high-resolution video generation and reduces the data and resource demands of scaling.

- Through the decoupled modeling of global semantic and local details, we significantly reduce the computing complexity of ultra-high-resolution generation leading to a speedup of $60.9\times$ compare to Wan2.1-1.3B baseline.
- Enabled by our resolution-agnostic training paradigm, **AtlasVid** is the first method to jointly scale up both spatial resolution and temporal duration (i.e 4K, 321frames) without the data bottleneck to our knowledge.
- The full pipeline can be trained via LoRA fine-tuning at 720P resolution on as few as **2 NVIDIA RTX PRO 6000 Blackwell GPUs** and inference on 1 GPU within 29 minutes, with the learned capability transferring directly to 4K inference without any additional high-resolution training stage. In comparison to other method which requires massive training on more than 32 GPUs clusters, **AtlasVid** substantially lowers the resource barrier ultra-high-resolution long video generation.

2 Related Work

2.1 High Resolution Video Generation

Recent diffusion transformer based video generation models [Wan et al., 2025, Team, 2025, Hong et al., 2022] have demonstrated impressive synthesis quality, yet they are still largely trained at resolutions up to 720P. Scaling these models to ultra-high resolution (UHD), such as 4K, remains challenging due to the quadratic complexity of full attention with respect to token count and the scarcity of high-quality 4K video data. Training-free methods [He et al., 2023, Zhang et al., 2024c, Qiu et al., 2025, Wu et al., 2025] adapt pretrained models to higher resolutions by modifying attention patterns, receptive fields, or positional encodings at inference time. While computationally convenient, they rely entirely on low-resolution priors and often suffer from semantic repetition, over-smoothed textures, and content inconsistency at 4K. Video super-resolution methods [He et al., 2024, Xie et al., 2025, Zhuang et al., 2025] instead cascade low-resolution generation with a dedicated spatial upscaler, but such pipelines are mainly restricted to low-level texture enhancement and cannot reliably correct semantic errors or synthesize genuinely new high-frequency content. Native high-resolution fine-tuning approaches [Xue et al., 2025, Hu et al., 2026, Zhang et al., 2025a, Zhao et al., 2026] directly adapt foundation models on curated high-resolution datasets. However, these methods primarily scale spatial resolution and remain limited to short clips. In contrast, our proposed **AtlasVid** is resolution-agnostic: it scales effectively to higher resolutions without relying heavily on native 4K training data.

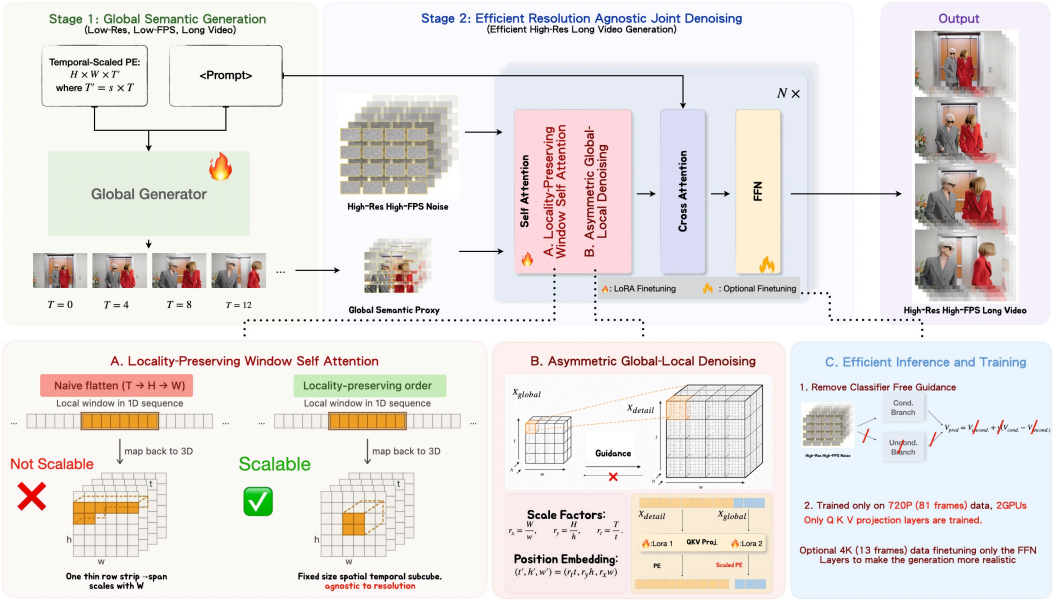

Figure 2: Pipeline of **AtlasVid**. It first employs a semantic generator to produce a low-resolution, low-frame-rate video that serves as a global semantic proxy. Conditioned on this reference, the second stage performs spatiotemporal detail generation through an efficient hierarchical locality-preserving attention mechanism, enabling ultra-high-resolution long-video synthesis(UHRL video) with substantially improved computational efficiency. Section A, B, and C demonstrate the detailed designs for scalable UHRL video generation.

Discussion with similar works. ***i)*** **Methodology.** UltraWan, UltraGen, and T3-Video all rely on large-scale 4K data for training. UltraGen adopts carefully designed hierarchical attention modules to reduce computation, while T3-Video uses transformed window attention for more efficient training and inference. In contrast, our method decouples global-local modeling and introduces locality-preserving attention, enabling efficient 4K inference while requiring training only at 720P with optional 4K finetuning for more realistic results. A detailed comparison with T3-Video is provided in Appendix F. ***ii)*** **Training resources.** As shown in Table 1, UltraGen uses 32 H20 GPUs, and T3-Video requires 64 GPUs for training, whereas our method requires only 2 GPUs. ***iii)*** **Results.** Our method is the first to enable long single-shot 4K video generation.

2.2 Efficient Video Generation

Efficient video generation has attracted increasing attention due to the heavy computational cost of diffusion-based video models, which becomes more prohibitive at ultra high spatial temporal resolutions. Existing methods mainly improve efficiency through several complementary directions. First, hardware-aware attention kernels such as FlashAttention [Dao, 2023, Shah et al., 2024] and SageAttention [Zhang et al., 2024b,a, 2025b] accelerate standard attention computation and reduce memory overhead without changing the model architecture. Second, caching-based methods exploit redundancy across denoising timesteps by reusing intermediate features or attention outputs, thereby reducing repeated computation during sampling [Liu et al., 2024, Ma et al., 2025b, Fan et al., 2025]. Third, many recent approaches improve efficiency at the architectural level by introducing complexity-reduced spatiotemporal modeling, such as training-free sparsification, trainable sparse attention, and linear-complexity designs [Zhang et al., 2025c, Wang et al., 2025, Zhang et al., 2025d]. In addition, distillation-based methods reduce the number of sampling steps for faster deployment [Luo et al., 2023, Yin et al., 2024a]. Different from these works that primarily target general video generation efficiency, our work focuses on native high-resolution video generation and develops sparse-attention-based optimization tailored to low reference guidance setting.

3 Method

To generate high-resolution long single-shot videos without incurring the prohibitive quadratic cost of standard self-attention or the substantial training burden of direct high-resolution long-video modeling, **AtlasVid** reformulates spatiotemporal video generation as an efficient hierarchical two-stage pipeline. As illustrated in Figure 2, the first stage employs a semantic generator to produce a low-resolution, low-frame-rate video that serves as a global semantic proxy. Conditioned on this reference, the second stage performs spatiotemporal detail generation through an efficient hierarchical locality-preserving attention mechanism, enabling high-resolution long-video synthesis with substantially improved computational efficiency. In the remainder of this section, Section 3.1 introduces the preliminaries of DiT-based video diffusion models, Section 3.2 presents our decoupled formulation of global semantic and local detail modeling for scalable and efficient generation, and Section 3.3 describes how the global semantic proxy guides high-resolution long-video denoising.

3.1 Preliminaries of DiT based Video Diffusion Models

Most recent DiT-based video generation models adopt a full-attention Transformer architecture to model spatiotemporal dependencies in latent video representations. Given an input video, a 3D-VAE first encodes it into a latent tensor of shape $D \times T \times H \times W$, where D is the channel dimension and T, H, W denote the temporal and spatial dimensions. The latent tensor is then patchified and flattened into a 1D token sequence of length $N = THW$. Full self-attention is applied over the entire sequence, producing an $N \times N$ attention map with complexity $O(N^2 D)$. As a result, the cost grows quadratically with the spatiotemporal token count, which quickly becomes prohibitive for high-resolution and long video generation.

3.2 Decoupled Modeling of Global Semantics and Local Details

Generating high-resolution long videos requires simultaneously modeling long-range spatiotemporal dependencies over extremely long token sequences while preserving fine-grained local fidelity. We observe that recent state-of-the-art video generation models already demonstrate strong visual synthesis ability across diverse objects, scales, and motion patterns, suggesting that pretrained models have largely internalized the local visual priors required for detail generation. The main challenge, instead, lies in efficiently maintaining coherent global semantics as the spatiotemporal extent grows. Based on this observation, we decouple global semantic modeling from local detail generation and address them separately in our framework.

Global Semantic Generation via Temporally Scaled Positions. The pretrained base model already possesses the capacity to generate low-resolution semantic proxies, but extending such semantic modeling to longer temporal horizons still requires adaptation. A straightforward solution is to finetune the base model for longer video generation directly. However, this quickly becomes prohibitively expensive due to the quadratic complexity of self-attention with respect to temporal token length.

Importantly, global semantic modeling does not require the full frame rate of the target video. Instead, a low-frame-rate video is sufficient to capture the coarse scene evolution and long-range temporal semantics. Based on this observation, we construct the first stage as a low-frame-rate semantic generator. Specifically, we enlarge the temporal indices used in RoPE by a factor of r_t, such that adjacent generated frames are interpreted as being separated by larger temporal intervals. This allows the model to represent a longer temporal horizon using the same number of frames. For example, when $r_t = 4$, a 4fps 20-second semantic proxy has the same temporal token length as an original 5-second video, reducing the self-attention cost by $16\times$ compared with directly modeling the same duration at the original frame rate. We implement this adaptation by finetuning the base model with LoRA, enabling efficient learning of long-horizon semantic generation while preserving the pretrained visual prior.

Locality-Preserving Efficient Attention for Detail Generation. Let the target video contain N latent tokens. Applying full self-attention over all tokens leads to a quadratic complexity of $O(N^2)$, which is prohibitive for high-resolution long-video generation. Existing efficient attention schemes typically reduce this cost by restricting attention to local windows, yielding a complexity of $O(Nb)$ with window size b. However, when windows are formed directly on the flattened token sequence,

sequence locality does not necessarily coincide with geometric locality in the original spatiotemporal video volume. As a result, tokens grouped within the same attention window may not belong to the same coherent local region in space and time.

We therefore introduce a locality-preserving efficient attention mechanism for detail generation. Figure 2 (A) demonstrates the differences between ours and naive local-attention. Specifically, given a latent video volume of size $T \times H \times W$, we partition it into spatiotemporal cubes of size $t \times h \times w$, chosen such that each cube remains within the local generation regime that the pretrained base model can effectively handle. We then reorder the flattened token sequence so that tokens from the same cube become contiguous. Window attention is applied on the reordered sequence with window size $b = thw$, ensuring that each sparse attention block corresponds to an actual local spatiotemporal neighborhood.

This construction preserves the computational efficiency of local attention while making the sparse attention pattern compatible with the geometric structure of the video. Consequently, it better preserves the pretrained model's ability to synthesize fine local details. In practice, we additionally allow attention across adjacent cubes to reduce blocking artifacts near cube boundaries.

3.3 Efficient Resolution Agnostic Joint Denoising

Given the global semantic proxy X_{global} with length n and the target noisy video tokens X_{detail} with length N, our next goal is to construct an efficient joint denoising architecture that allows X_{global} to guide the generation of X_{detail} while remaining scalable across resolutions.

Efficient Joint Denoising with Hierarchical Attention. To inject global semantic guidance into high-resolution detail generation, we concatenate the two token streams as $X = [X_{\text{detail}}; X_{\text{global}}]$, and perform joint self-attention followed by text cross-attention. The key design principle is an asymmetric hierarchical attention pattern in which the coarse global proxy provides semantic guidance, while the high-resolution detail branch performs denoising under both local detail interactions and global semantic conditioning.

In self-attention, X_{global} attends only to itself, while X_{detail} attends both to its own tokens through the locality-preserving attention described in Section 3.2 and to the aligned global semantic tokens. The self-attention is formulated as

$$\text{SelfAttn}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top + M}{\sqrt{d}}\right) V, M = \begin{bmatrix} M_{dd} & M_{dg} \\ -\infty & M_{gg} \end{bmatrix} \in \mathbb{R}^{(N+n)\times(N+n)}$$

Here, M_{dd} denotes the locality-preserving mask for detail-to-detail attention, M_{dg} denotes the mask for detail-to-global attention, and M_{gg} denotes unrestricted self-attention among global tokens. The $-\infty$ block prevents global tokens from attending to detail tokens, thereby enforcing one-way semantic guidance from the global proxy to the noisy detail branch. For text cross-attention, we only allow X_{detail} to attend to text tokens. This asymmetric design ensures that the global semantic proxy remains clean and is not corrupted by noisy detail features during denoising.

To adapt the pretrained DiT backbone to this new joint denoising pattern, we apply LoRA finetuning to all the DiT layers associated with global semantic conditioning and to only the query, key, and value projection layers in the detail branch. In contrast, we keep the feed-forward layers in the detail branch frozen in this joint denoising pattern learning process, so that the pretrained model's original local generative capability is preserved as much as possible while learning the detail-to-global mapping. As an optional refinement stage, we additionally finetune previously frozen FFN layers using a small and short (13-frame) videos to further improve local realism.

Scaled Spatial-Temporal RoPE for Global-Local Matching. To make the global semantic proxy provide accurate guidance for high-resolution detail generation, we align its positional encoding with the coordinate system of the target video. Demonstrated in Figure 2 (B), for a global proxy token located at spatial-temporal index (w', h', t'), we scale its RoPE coordinates by

$$r_x = \frac{W}{w}, \qquad r_y = \frac{H}{h}, \qquad r_t = \frac{T}{t},$$

so that its position is mapped to the corresponding location in the high-resolution long-video latent grid. In this way, each token in X_{global} serves as an anchor for the corresponding spatiotemporal cube in X_{detail}, enabling consistent global-to-local semantic guidance during denoising.



Figure 3: Qualitative comparison of long ultra-high-resolution video generation. Top: The first two rows compare **AtlasVid** with SkyReels-V2 at 720P and T3-Video at 4K. All three methods generate 161-frame results. T3-Video exhibits structural artifacts, color shifts, and quality degradation when extended to 161 frames, whereas **AtlasVid** maintains coherent structures and stable visual quality. Bottom: Rows 3–5 show the 81-frame comparison with UltraGen at 1080P and T3-Video at 4K. **AtlasVid** preserves 4K-level details and is the only method that scales reliably beyond 81 frames.

This positional alignment naturally complements the asymmetric hierarchical attention described above: the global proxy provides semantically aligned coarse guidance, while the detail branch focuses on synthesizing local high-frequency content within each spatiotemporal neighborhood. Since our framework explicitly decouples global semantic modeling from local detail generation, training only needs to learn the new locality-preserving attention pattern and the hierarchical global-local correspondence, rather than directly modeling the full target spatiotemporal scale end-to-end.

As a result, **AtlasVid** does not require long ultra-high-resolution videos (e.g., 4K, 321-frame videos) for training as illustrated in Figure 2(c) , making the overall training pipeline resolution-agnostic and scalable. Moreover, due to the existence of the global proxy as condition, we remove the Classifier-Free Guidance which significantly speeds up the inference.

4 Experiments

4.1 Implementation Details

Training details. We implement **AtlasVid** on top of Wan2.1-1.3B for both stage 1 and stage 2. Models are trained on UltraVideo using AdamW optimizer (batch size 1) on 2 NVIDIA RTX PRO 6000 Blackwell GPUs with gradient accumulation 4, and Stage 2 is trained for 15K iterations. Learning rates is 1e-4 with LoRA rank 16. Models are trained on 720p, 81 frames, and is successfully scaled beyond 4K, 321frames. Stage 1 fine-tunes the base model with temporal-scale RoPE (r_t=4) for long-horizon low frame-rate semantic generation. We set the spatial-temporal block size in stage 2 to be $(256, 256, 32)$ which is equivalent to $(8, 8, 4)$ in the latent space and adapts a smaller block-size

for the border cases when the spatial-temporal resolution cannot be divided by the block-size. We adapt the mask rule for flex attention to support dynamic length and resolution.

Baselines. We compare against: Wan2.1-T2V-1.3B [Wan et al., 2025], UltraWan [Xue et al., 2025], UltraGen [Hu et al., 2026], and T3-Video-T2V-1.3B [Zhang et al., 2025a]. Note that neither Wan2.1-T2V-1.3B, nor UltraGen have released official 4K-resolution checkpoints, we therefore evaluate them at their highest support resolution, 720p and 1080p separately.

4.2 Qualitative Results

Demonstration of UHR long videos. Figure 1 presents examples of ultra-high-resolution long videos generated by our method, reaching up to 8K resolution and 321 frames in temporal length. These results demonstrate the scalability of our framework in both spatial resolution and temporal duration, while maintaining coherent scene structure and fine-grained visual details.

Comparison on 4K long videos Figure 3 provides qualitative comparisons for long 4K video generation. The upper section compares long video generation ability on 161 frames with the long-video generator SkyReels-V2 under 720p and the native 4K generator T3-Video under 4K. Our method is the only one that produces plausible long 4K results. When extrapolated to longer temporal lengths, T3-Video exhibits noticeable quality degradation and prompt misalignment, such as color shift and incorrectly changing the prompted man into a woman. SkyReels-V2 can generate long video with consistency, but their result are limited to 720p. The lower section further compares our method with native UHR video generation methods, including UltraGen and T3-Video. UltraGen produces noticeably inconsistent results when using its released 1080P checkpoint, such as the feather on the head in green box. Though our method and T3-Video preserve detailed structures at 4K resolution, T3-Video still exhibits structural artifacts, such as distorted bird shape in green and yellow box, with extra legs in blue box, due to the lack of explicit global semantic control. In contrast, our decoupled design provides stable semantic guidance and more coherent high-resolution long-video synthesis.

4.3 Quantitative Results

Table 2 reports quantitative results using both VBench and high-resolution evaluation metrics, with detailed metric definitions provided in the appendix. We randomly sample 100 prompts from VBench as our test set. Wan2.1 and UltraGen are evaluated at 720P and 1080P respectively, as their 4K checkpoints are not publicly available, while UltraWan, T3-Video, and **AtlasVid** are evaluated at 4K resolution. **AtlasVid** achieves the best performance on most VBench metrics among existing 4K generation methods, and remains competitive with pretrained models evaluated at 720P resolution. These results demonstrate the strong capability of **AtlasVid** in synthesizing high-quality ultra-high-resolution videos while preserving both visual fidelity and temporal consistency. Since HD-FVD is computed on patch-level features whose distribution depends on the native resolution (Appendix C), HD-FVD values across different resolutions should be interpreted with caution. We therefore draw our main conclusions from the same-resolution 4K comparison among UltraWan, T3-Video, and **AtlasVid**, and include the 720P and 1080P rows only as contextual references.

Long-duration temporal coherence. On 321-frame ($\sim$20s) 2K videos, the CLIP image similarity between frame 0 and the frames at 25%, 50%, 75%, and 100% of the video is 0.950, 0.934, 0.932, and 0.922, decaying smoothly without abrupt jumps, while the CLIP-to-prompt score of 4 consecutive segments remains stable (0.308–0.315), indicating minimal long-range drift and no quality degradation over time.

Efficiency Comparisons. We also compare our model's efficiency against the other baselines. The bottom part of Figure 1 demonstrates the comparison among our attention and other attention mechanisms across 720P to 8K resolutions (81 frames), where our attention even beats linear attention in terms of floating point operations (FLOPs) and surpasses dense attention by 1208.2 times. The right part demonstrates the overall inference time comparison on 4K 81 frames, given the efficient designs we achieve 1.48× faster than T3-Video and 60.9× faster than baseline.

Table 2: Quantitative comparison on 4K long video generation. Our method achieves the best performance among 4K generation methods and remains competitive with 720P base models.

Method	4K HD Metrics			VBench Motion					VBench Semantics		
	HD-FVD↓	HD-LPIPS↑	CLIP↑	SC↑	BC↑	TF↑	MS↑	DD↑	AQ↑	IQ↑	Clr.↑
Wan2.1	512.32	0.6424	0.3076	96.34	97.29	99.49	97.44	**85.56**	62.43	66.51	89.58
UltraGen	432.71	0.6358	0.2848	94.19	96.33	98.90	98.72	71.34	58.13	65.88	97.13
UltraWan	372.39	0.6211	0.2903	95.71	**97.94**	98.86	99.06	62.22	59.52	67.39	**99.10**
T3-Video	311.29	0.6471	0.2972	97.02	96.23	95.12	99.17	78.92	62.96	68.17	97.42
Ours	**284.71**	**0.6695**	**0.3019**	**97.41**	96.54	**99.57**	**99.31**	82.21	**63.31**	**68.52**	98.31

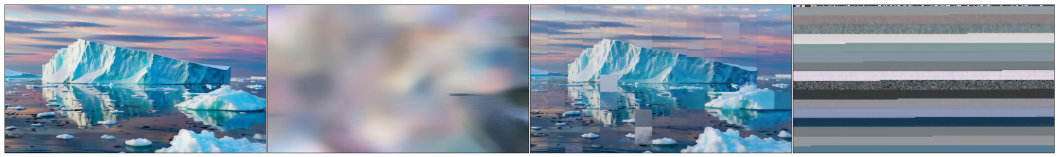

Figure 4: Ablation on the importance of our attention design. The first two columns demonstrate the effect of global guidance. The third and fourth columns evaluate our locality-preserving attention, where attention to neighboring spatiotemporal blocks is removed to better visualize the learned block structure. All models are trained at 720P and evaluated at 4K.

5 Ablation Study

Ablation on different attention patterns. We conduct ablation studies on different attention patterns to validate our design choices, with all variants trained only on 720P data. The left part shows that removing the global proxy guidance leads to inconsistent generation results, highlighting the importance of global semantic conditioning. The right part demonstrates the role of locality-preserving attention in extrapolating beyond the training resolution. For clearer visualization, we remove attention to neighboring blocks in this ablation so the "blocks" will be shown in the figure. With our locality-preserving attention, the model successfully scales to 4K generation, whereas naive block attention fails to generalize. These results confirm the effectiveness of our global-local design and locality-preserving attention mechanism.

Ablation on 4K data fine-tuning. Figure 5 compares videos generated by models trained with and without additional 4K real data. Even without 4K fine-tuning, our model already produces clear 4K textures, demonstrating the resolution extrapolation ability of our framework. Additional 4K fine-tuning further improves naturalness and local realism, and is therefore used only as an optional refinement stage. Importantly, this stage is substantially lighter than the native 4K training required by UltraGen and T3-Video: instead of using long 4K 81-frame videos, we only require short 4K 13-frame clips to adapt the feed-forward layers toward more natural high-resolution pixel synthesis.

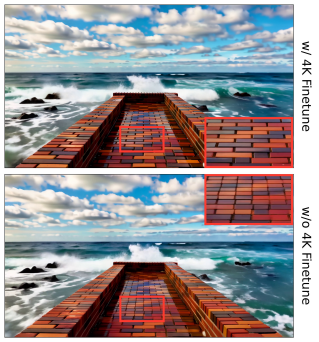

Figure 5: Ablation on 4K fine-tuning. Top: with 4K fine-tuning; bottom: without.

Ablation on the temporal scaling factor. Table 3 varies the temporal scaling factor r_t of the Stage-1 proxy, where proxy frames are obtained by stride sub-sampling. Reducing r_t from 4 to 2 brings no clear quality gain, while doubling the number of proxy frames needed to cover the same duration. Increasing r_t to 8 marks the breakdown point: HD-FVD rises to 308.29 and motion smoothness (MS) drops to 97.81, as the detail branch struggles to bridge the larger temporal gaps between proxy frames. This is consistent with the division of labor in our design: the proxy only provides sparse global anchors, while the detail branch synthesizes the intermediate motion and fine details between them. We therefore use r_t=4 by default.

Table 3: Ablation on the temporal scaling factor r_t of the Stage-1 proxy.

Metric	r_t=2	r_t=4 (**ours**)	r_t=8
HD-FVD↓	288.38	**284.71**	308.29
MS↑	99.24	**99.31**	97.81
CLIP↑	**0.3042**	0.3019	0.3025

Table 4: Ablation on the removal of classifier-free guidance (CFG).

Variant	**Inference**	**HD-FVD↓**	**HD-LPIPS↑**	**CLIP↑**
w/ 4K fine-tuning	w/ CFG	**273.64**	0.6612	0.3012
w/ 4K fine-tuning	w/o CFG (**ours**)	284.71	**0.6695**	**0.3019**
w/o 4K fine-tuning	w/ CFG	408.53	0.6398	0.3017
w/o 4K fine-tuning	w/o CFG	409.15	0.6401	0.3011

Ablation on removing classifier-free guidance (CFG). Since the global proxy already provides a deterministic semantic target for high-resolution denoising, we remove classifier-free guidance to improve inference efficiency. We evaluate this design both with and without optional 4K fine-tuning. As shown in Table 4, disabling CFG has no significant impact on the quantitative metrics in either setting. Meanwhile, it nearly halves the DiT inference cost by eliminating the unconditional generation branch. We therefore disable CFG in our final generation pipeline. Table 4 also quantifies the benefit of the optional 4K fine-tuning (Figure 5): without CFG, it reduces HD-FVD from 409.15 to 284.71.

Cross-cube spatial consistency. Although detail tokens do not attend to all other detail tokens, spatial consistency across cubes is maintained by two mechanisms: (i) each cube also attends to its adjacent cubes (Section 3.2), so information propagates across cube boundaries; and (ii) all detail tokens attend to the shared, coordinate-aligned global proxy (Section 3.3), which serves as a global consistency backbone for large objects, lighting, and scene layout. Figure 4 supports both roles: removing the global proxy yields inconsistent content across regions, while removing attention to neighboring cubes exposes the underlying block structure.

Limitations. **AtlasVid** relies on the Stage-1 global proxy. Fine details missing from the low-resolution proxy can be synthesized by the detail branch, but high-level semantic errors, such as missing objects or incorrect layouts, may propagate to the final output (Appendix D). Since the proxy is inexpensive to generate, it can serve as a preview before the costly 4K stage. Future work includes training with semantically corrupted proxies and combining our framework with orthogonal acceleration techniques such as distillation.

6 Conclusion

We presented **AtlasVid**, an efficient framework for ultra-high-resolution long video generation. By decoupling global semantic modeling from local detail synthesis, **AtlasVid** avoids directly applying full attention to prohibitively large spatiotemporal token sequences. A low-resolution, low-frame-rate semantic proxy captures long-range scene evolution, while a high-resolution detail branch performs coordinate-aligned joint denoising with hierarchical locality-preserving attention. This design enables resolution-agnostic training: with lightweight LoRA adaptation mainly at 720P, **AtlasVid** generalizes to 4K and beyond while substantially reducing computation. Experiments demonstrate that **AtlasVid** produces coherent, detailed ultra-high-resolution long videos with significantly improved efficiency, offering a scalable path toward accessible ultra-high-resolution video generation.

Acknowledgments and Disclosure of Funding

This work is supported by Dartmouth College faculty startup fund.

References

Vladimir Arkhipkin, Vladimir Korviakov, Nikolai Gerasimenko, Denis Parkhomenko, Viacheslav Vasilev, Alexey Letunovskiy, Nikolai Vaulin, Maria Kovaleva, Ivan Kirillov, Lev Novitskiy, et al. Kandinsky 5.0: A family of foundation models for image and video generation. *arXiv preprint arXiv:2511.14993*, 2025.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

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*, 2025a.

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*, 2025b.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. *arXiv preprint arXiv:2602.06028*, 2026.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

Zhentao Fan, Zongzuo Wang, and Weiwei Zhang. Taocache: Structure-maintained video generation acceleration. *arXiv preprint arXiv:2508.08978*, 2025.

Google DeepMind. Veo 3.1, 2025. URL `https://deepmind.google/models/veo/`.

Yuwei Guo, Ceyuan Yang, Ziyan Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17281–17291, 2025.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Jingwen He, Tianfan Xue, Dongyang Liu, Xinqi Lin, Peng Gao, Dahua Lin, Yu Qiao, Wanli Ouyang, and Ziwei Liu. Venhancer: Generative space-time enhancement for video generation. *arXiv preprint arXiv:2407.07667*, 2024.

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.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

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.

Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers. *arXiv preprint arXiv:2410.23775*, 2024a.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

Kling AI. Kling VIDEO 3.0, 2026. URL `https://kling.ai/quickstart/klingai-video-3-model-user-guide`.

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.

Debang Li, Zhengcong Fei, Tuanhui Li, Yikun Dou, Zheng Chen, Jiangping Yang, Mingyuan Fan, Jingtao Xu, Jiahua Wang, Baoxuan Gu, et al. Skyreels-v3 technique report. *arXiv preprint arXiv:2601.17323*, 2026.

Zhen Li, Zuo-Liang Zhu, Ling-Hao Han, Qibin Hou, Chun-Le Guo, and Ming-Ming Cheng. Amt: All-pairs multi-field transforms for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9801–9810, 2023.

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.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. *arXiv preprint arXiv:2411.19108*, 2024.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

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*, 2025a.

Zehong Ma, Longhui Wei, Feng Wang, Shiliang Zhang, and Qi Tian. Magcache: Fast video generation with magnitude-aware cache. *arXiv preprint arXiv:2506.09045*, 2025b.

OpenAI. Sora 2, 2025. URL `https://developers.openai.com/api/docs/models/sora-2`.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

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.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *Advances in Neural Information Processing Systems*, 37:68658–68685, 2024.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report, 2025. URL https://arxiv.org/abs/2511.18870.

Team Seedance et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026. URL https://arxiv.org/abs/2604.14148.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

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.

Hongjie Wang, Chih-Yao Ma, Yen-Cheng Liu, Ji Hou, Tao Xu, Jialiang Wang, Felix Juefei-Xu, Yaqiao Luo, Peizhao Zhang, Tingbo Hou, et al. Lingen: Towards high-resolution minute-length text-to-video generation with linear computational complexity. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2578–2588, 2025.

Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2024.

Yunfeng Wu, Jiayi Song, Zhenxiong Tan, Zihao He, and Songhua Liu. Freeswim: Revisiting sliding-window attention mechanisms for training-free ultra-high-resolution video generation. *arXiv preprint arXiv:2511.14712*, 2025.

Rui Xie, Yinhong Liu, Penghao Zhou, Chen Zhao, Jun Zhou, Kai Zhang, Zhenyu Zhang, Jian Yang, Zhenheng Yang, and Ying Tai. Star: Spatial-temporal augmentation with text-to-video models for real-world video super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17108–17118, 2025.

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.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6613–6623, 2024a.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast causal video generators. *arXiv e-prints*, pages arXiv–2412, 2024b.

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 10x. *arXiv preprint arXiv:2512.13492*, 2025a.

Jintao Zhang, Haofeng Huang, Pengle Zhang, Jia Wei, Jun Zhu, and Jianfei Chen. Sageattention2: Efficient attention with thorough outlier smoothing and per-thread int4 quantization. *arXiv preprint arXiv:2411.10958*, 2024a.

Jintao Zhang, Jia Wei, Haofeng Huang, Pengle Zhang, Jun Zhu, and Jianfei Chen. Sageattention: Accurate 8-bit attention for plug-and-play inference acceleration. *arXiv preprint arXiv:2410.02367*, 2024b.

Jintao Zhang, Jia Wei, Pengle Zhang, Xiaoming Xu, Haofeng Huang, Haoxu Wang, Kai Jiang, Jianfei Chen, and Jun Zhu. Sageattention3: Microscaling fp4 attention for inference and an exploration of 8-bit training. *arXiv preprint arXiv:2505.11594*, 2025b.

Jintao Zhang, Chendong Xiang, Haofeng Huang, Jia Wei, Haocheng Xi, Jun Zhu, and Jianfei Chen. Spargeattn: Accurate sparse attention accelerating any model inference. *arXiv e-prints*, pages arXiv–2502, 2025c.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025d.

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, 2024c.

Canyu Zhao, Mingyu Liu, Wen Wang, Weihua Chen, Fan Wang, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024.

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.

Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*, 3(6), 2024.

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 Results

Here we presents more results on 4K 161 frames, 2K 321 frames and 8K 29 frames in Figure 6, Figure 7, Figure 8, Figure 9 and Figure 10. We will include the videos in the supplementary materials.

B Extended Implementation Details

B.1 Training Configuration

We build **AtlasVid** on top of Wan2.1-T2V-1.3B Wan et al. [2025] and train both stages with LoRA rank 16 using the AdamW optimizer (learning rate 1×10^{-4}, β_1=0.9, β_2=0.95, weight decay 0.01). All training is conducted on 2× NVIDIA RTX PRO 6000 Blackwell GPUs, batch size 1 per GPU with gradient accumulation 4 (effective batch size 8). Mixed-precision (`bf16`) is used throughout, and we adopt a flow-matching objective consistent with the Wan2.1 base model.

Stage 1 (semantic generator). We finetune the base model with temporal-scale RoPE (r_t=4) on 720P × 81-frame clips sub-sampled at 4 fps, so that an 81-frame proxy spans an effective horizon of $\sim$20 seconds at 16 fps target frame rate.

Stage 2 (detail branch). The latent volume is partitioned into spatiotemporal cubes of pixel size $(256, 256, 32)$, equivalent to $(8, 8, 4)$ in the 4D latent grid produced by the Wan2.1 3D-VAE. Border cubes that cannot be evenly tiled adopt a smaller block size automatically. We implement the asymmetric mask M (§3.3) on top of PyTorch `flex_attention` so that the kernel supports dynamic resolution and length without re-compilation per shape. This stage is trained for 15K iterations.

Optional 4K refinement. The optional FFN refinement stage finetunes only the feed-forward layers of the detail branch on 13-frame 3840×2160 clips drawn from UltraVideo Xue et al. [2025], for 2K iterations at the same learning rate.

B.2 Inference Configuration

Unless stated otherwise, all reported numbers and figures use Euler sampler and *disable* classifier-free guidance (CFG), which is justified by the ablation in §5. We perform inference on a single NVIDIA RTX PRO 6000 Blackwell GPU. End-to-end runtime includes text encoding, both denoising stages, and 3D-VAE decoding.

C Detailed Metric Definitions

This section formalises the metrics referenced in Table 2 and the ablation tables of the main paper. For all metrics, $v \in \mathbb{R}^{T\times H\times W\times 3}$ denotes a video. Higher-is-better metrics are marked ↑ and lower-is-better ↓.

C.1 High-Definition Metrics

The HD-FVD, HD-LPIPS, and HD-MSE metrics are designed for ultra-high-resolution video evaluation, where standard FVD/LPIPS are constrained by the 224×224 input size of their backbones. We follow the formulation introduced by UltraGen Hu et al. [2026].

HD-FVD ↓ Standard FVD computes the Fréchet distance between I3D Carreira and Zisserman [2017] feature distributions of generated and reference videos. Because I3D operates on 224×224 inputs, naively resizing a 4K clip to 224^2 destroys all high-frequency content. HD-FVD instead splits each frame into a regular grid of non-overlapping $H_l\times W_l$ patches with $H_l\approx W_l\approx224$, extracts I3D features per spatiotemporal patch tube, and reports

$$\text{HD-FVD} = \text{FD}\Big(\mathcal{N}(\mu_g,\Sigma_g)\,\Big\|\,\mathcal{N}(\mu_r,\Sigma_r)\Big),$$

where (μ_g,Σ_g) and (μ_r,Σ_r) are the means and covariances of generated and reference patch-feature distributions, and $\text{FD}(\cdot\|\cdot)$ is the Fréchet distance. We use the public Kinetics-pretrained I3D checkpoint and patch each $T=16$-frame clip into 3×3 spatial patches at 4K (covering the full frame).



Figure 6: 4K 161 Frames results: Each results spans one row. Examples show no quality degradation or color drift. The frame indices are 0, 53, 106, and 160.



Figure 7: 4K 161 Frames results: Each results spans one row. Examples show no quality degradation or color drift. The frame indices are 0, 53, 106, and 160.

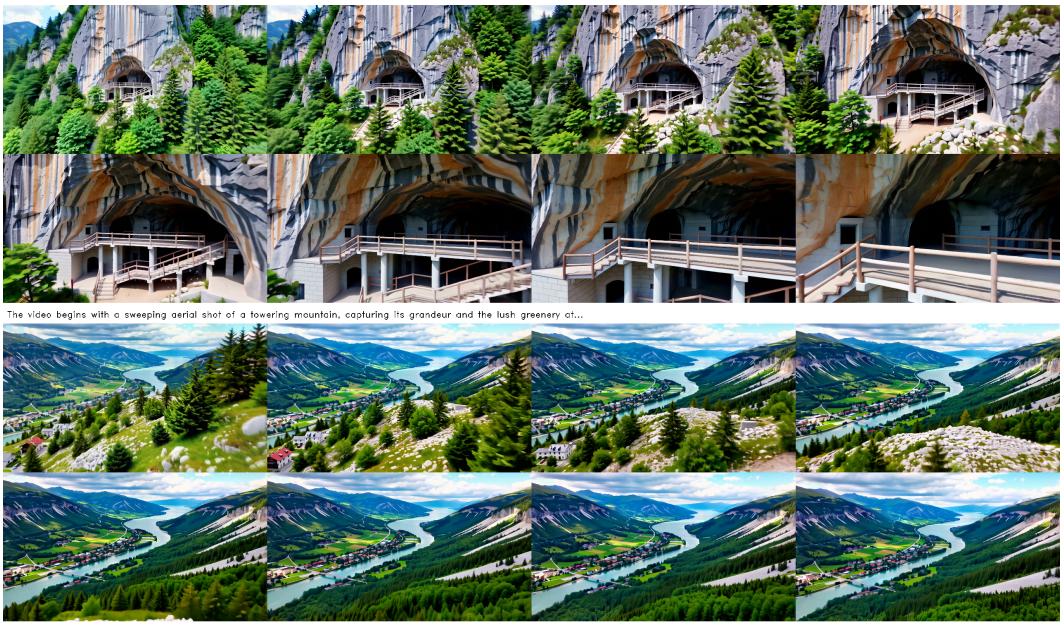

Figure 8: 2K 321 Frames results: Each results spans for two rows. The examples here shows large continuous camera movements. The frame indices are 0, 45, 91, 137, 182, 228, 274 and 320.



Figure 9: 2K 321 Frames results: Each results spans for two rows. The examples here shows large continuous camera movements. The frame indices are 0, 45, 91, 137, 182, 228, 274 and 320.

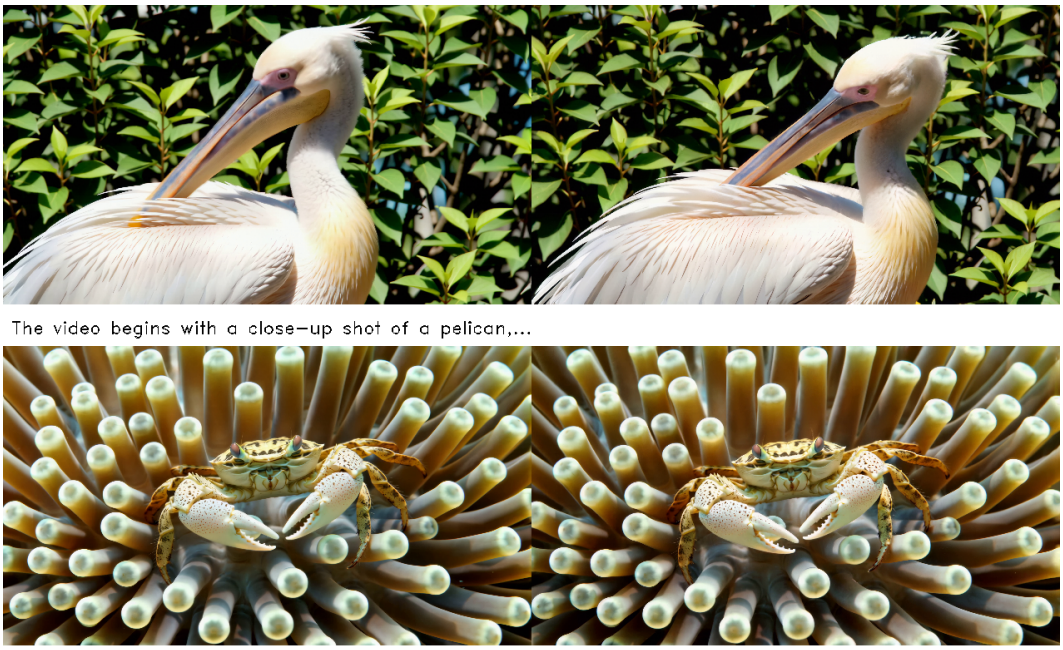

Figure 10: 8K 29 Frames results: Each results spans for two rows. The frame indices are 0 and 28.

Specifically, we apply the same evaluation algorithm to both Wan2.1 (720p) and UltraGen (1080p), which results in relatively high HD-FVD scores. This is because the patch distributions of low-resolution and high-resolution videos differ significantly.

HD-LPIPS ↑ This *no-reference* metric quantifies how much high-frequency content is preserved in a generated video by measuring its perceptual distance from progressively low-pass-filtered copies of itself. Let v_t denote the t-th frame of a video V of length T, and let $v_{t,D,2^k}$ denote v_t bilinearly down-sampled by a factor of 2^k and then bilinearly up-sampled back to the original resolution. We define

$$\text{HD-LPIPS}(V) \;=\; \frac{1}{T}\sum_{t=1}^{T}\sum_{k\in\mathcal{K}} \text{LPIPS}\big(v_t,\, v_{t,D,2^k}\big)\,,$$

where $\mathcal{K} = \{3, 4, 5\}$ corresponds to down-sampling factors of 8, 16, and 32, chosen to capture detail across multiple spatial scales while avoiding both noise-dominated (small factors) and structure-dominated (large factors) regimes. A higher HD-LPIPS indicates that the generated video differs more from its low-pass-filtered versions, suggesting richer high-frequency content. We note that this metric reflects high-frequency energy in general, and should therefore be interpreted alongside HD-FVD and perceptual quality metrics to disambiguate genuine fine detail from high-frequency artifacts.

C.2 Text Video Alignment

CLIP ↑**.** We report frame-averaged cosine similarity between the prompt embedding and per-frame visual embeddings. For a video with T frames and prompt p,

$$\text{CLIP}(v, p) \;=\; \tfrac{1}{T}\sum_{t=1}^{T} \frac{\langle f_{\text{img}}(v_t),\, f_{\text{txt}}(p)\rangle}{\|f_{\text{img}}(v_t)\|\,\|f_{\text{txt}}(p)\|}.$$

C.3 VBench Dimensions

We follow the VBench Huang et al. [2024b] protocol and report the eight dimensions used in Table 2. All scores lie in $[0, 1]$ and higher is better.

- **Subject Consistency (SC)** — frame-to-frame DINO Caron et al. [2021] cosine similarity of the foreground subject; measures whether the subject identity is preserved across the clip.
- **Background Consistency (BC)** — frame-to-frame CLIP image cosine similarity of the background region; measures scene stability.
- **Temporal Flickering (TF)** — mean absolute pixel difference between consecutive frames in static regions; rewards low flicker.
- **Motion Smoothness (MS)** — drops every other frame, interpolates it back with the AMT Li et al. [2023] video frame interpolation model, and reports the inverse reconstruction error; rewards physically plausible motion.
- **Dynamic Degree (DD)** — average RAFT Teed and Deng [2020] optical-flow magnitude; rewards non-static videos. Reported in $[0, 100]$ following the official VBench convention.
- **Aesthetic Quality (AQ)** — per-frame LAION aesthetic predictor score, averaged temporally.
- **Imaging Quality (IQ)** — per-frame MUSIQ Ke et al. [2021] SPAQ-trained image-quality predictor, averaged temporally.
- **Color (Clr.)** — GRiT Wu et al. [2024] caption-based color attribute alignment between the rendered subject and the color word in the prompt.

D Sensitivity to the Global Proxy

Controlled proxy perturbation. To test whether deviations in the global proxy are amplified by the detail branch, we construct proxies by downsampling high-resolution videos, inject Gaussian noise of relative magnitude σ into the proxy while fixing all other sources of randomness, and measure the RMSE between the resulting 4K output and the output obtained with the clean proxy. As shown in Table 5, the output RMSE grows sub-linearly with σ, and the ratio RMSE$/\sigma$ remains between 0.15 and 0.22. Within this controlled setting, we therefore observe no super-linear amplification of proxy deviations. The CLIP score of the final output gradually decreases as the perturbation becomes stronger, as expected; nevertheless, the detail branch consistently improves prompt alignment over the corresponding perturbed proxy, suggesting that text conditioning and the pretrained visual prior can partially compensate for degraded low-level proxy information. We emphasize that this experiment only evaluates continuous perturbations and does not establish recovery from discrete semantic failures such as missing objects or incorrect layouts (see the limitations in the main paper).

Table 5: Sensitivity of the final 4K output to Gaussian perturbations of the global proxy. ΔCLIP denotes the CLIP gain of the final output over the corresponding perturbed proxy.

σ	0.10	0.25	0.50	1.00
RMSE	0.019	0.056	0.100	0.146
RMSE$/\sigma$	0.19	0.22	0.20	0.15
CLIP (ΔCLIP)	0.3012 (+0.005)	0.2991 (+0.013)	0.2903 (+0.016)	0.2878 (+0.034)

Training with perturbed proxies. As a proof of concept towards improving robustness to proxy errors, we fine-tune a variant of the detail branch with Gaussian proxy-noise augmentation. Evaluated with a strongly perturbed proxy (σ=1.00), this variant remains close to the default model evaluated with the clean proxy (Table 6). This suggests that explicitly training under corrupted proxies is a feasible direction, although semantic proxy errors remain outside the scope of this experiment.

Table 6: Training with Gaussian proxy-noise augmentation.

Model	**Inference proxy**	**HD-FVD**$\downarrow$	**HD-LPIPS**$\uparrow$	**CLIP**$\uparrow$
Ours (default)	clean (σ=0)	284.71	0.6695	0.3019
Ours + noise-aug. training	perturbed (σ=1.00)	298.12	0.6601	0.2982

E Additional Analyses

Frame averaging versus stride sub-sampling. The ablation on r_t in Section 5 constructs proxy frames by stride sub-sampling. As an alternative at the same r_t=4, we construct each proxy frame by averaging 4 consecutive frames (Table 7). Frame averaging modestly improves HD-FVD and CLIP, while MS decreases slightly. Both constructions capture the same coarse motion trend and reach similar quality, which is consistent with the proxy mainly needing to convey low-frequency scene evolution. All other results in this paper use stride sub-sampling.

Table 7: Proxy construction at r_t=4. Stride: proxy frames are sub-sampled with stride r_t; 4-frame average: each proxy frame averages 4 consecutive frames.

Proxy construction	**HD-FVD↓**	**MS↑**	**CLIP↑**
Stride (**ours**)	284.71	**99.31**	0.3019
4-frame average	**275.49**	99.16	**0.3077**

Quantitative seam analysis. To complement the qualitative evidence in Section 5, we test whether cube boundaries leave visible seams in the generated videos. For a regular spatial grid with period p, we compute the ratio of the mean image gradient magnitude on the grid lines to that in the grid interior, where a ratio of 1 indicates no excess edge energy on the grid. We evaluate p=128 px, whose grid contains every boundary of the 256 px cubes (Appendix B), together with the off-grid periods p=96 px and p=112 px. Seams at cube boundaries would raise the ratio at p=128 px above the off-grid periods. As shown in Table 8, the ratio is essentially identical across all three periods. The residual of about 1.16 is therefore not specific to cube boundaries, and we observe no cube-aligned seams by this measure. This measure targets grid-aligned discontinuities and does not by itself establish the long-range consistency of lighting or large objects.

Table 8: Ratio of the mean gradient magnitude on the grid lines to that in the grid interior for different grid periods.

Grid period p	96 px	112 px	128 px
Boundary/interior gradient ratio	1.165	1.167	1.164

F Comparison with T3-Video

Table 9: Comparison with T3-Video, whose statistics are taken from [Zhang et al., 2025a]. Iterations: optimizer iterations × global batch size.

	T3-Video	**AtlasVid (Ours)**
Global semantic modeling	implicit, within HR window attention	explicit LR, low-FPS proxy
Long-duration generation	degrades beyond training length (81f)	coherent up to 321f (trained at 81f)
Training data	native 4K, 81f	720P (+ optional 4K, 13f)
Fine-tuning	full-parameter, 720P → 4K	LoRA at 720P (+ optional FFN at 4K)
GPUs	64	2×RTX PRO 6000
Iterations	5K × 64	15K × 8 for Stage 2 (+ optional 2K)
Training memory	179.4G	39.8G (77.9G w/ 4K refinement)

T3-Video [Zhang et al., 2025a] also adopts local-window attention; the key distinction lies in how global semantics are modeled (Table 9). T3-Video applies window attention directly to the high-resolution tokens without a decoupled semantic proxy, which requires full-parameter fine-tuning on native 4K videos and, as shown in Figure 3, leads to structural artifacts and quality degradation beyond its training length. In contrast, **AtlasVid** delegates global semantics to an explicit low-resolution, low-frame-rate proxy, so that the detail branch only needs locality-preserving attention. This asymmetric global-local decoupling allows training mainly with LoRA at 720P on two GPUs without native 4K data, while generating coherent videos of up to 321 frames. Note that the 500-iteration 4K fine-tuning reported by T3-Video starts from a 720P model already trained for 5K iterations with a global batch size of 64, and thus does not reflect its full training budget.

G Broader Impact

Positive societal impacts. AtlasVid lowers the barrier to ultra-high-resolution long-video generation by training on as few as two workstation-class GPUs (NVIDIA RTX PRO 6000 Blackwell) at 720P resolution and transferring directly to 4K and beyond. This has several positive implications. (i) It democratises high-resolution video synthesis, making it practical for academic groups, independent creators, and non-profit educational projects that cannot afford clusters of 32–64 H-class GPUs. (ii) It reduces the energy and carbon footprint of training and serving UHR video models by an order of magnitude, since the dominant cost — quadratic spatiotemporal attention — is replaced with a locality-preserving variant whose attention FLOPs grow linearly in the number of spatiotemporal cubes. (iii) It can support accessibility applications such as automatic generation of high-quality educational visuals, sign-language interpretation videos, or visualisations for the visually-impaired community where high spatial detail matters.

Potential negative societal impacts. Like all powerful generative video systems, **AtlasVid** could in principle be misused for (i) producing photorealistic disinformation or non-consensual synthetic media ("deepfakes"), (ii) generating misleading long-form footage of public figures or events, or (iii) bypassing moderation tools that were trained on lower-resolution synthetic content. Because our framework is specifically designed to scale long-horizon and ultra-high-resolution synthesis, the resulting outputs are harder to distinguish from real footage than those of short low-resolution generators, which sharpens the dual-use concern.

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 abstract and introduction clearly state the main claims. These claims are supported by experiments in the paper.

   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: Limitations are discussed at the end of Section 5, with a sensitivity analysis of the global proxy in Appendix D.

   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: We don't include any theoretical proofs.

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 experiment section clearly describes datasets, APIs, evaluation metrics, and hardware setup. The methodology also describes the module design in detail.

   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: [No]

Justification: The paper builds on publicly available pretrained models and datasets. Code will be released at `https://github.com/ziyang1106/AtlasVid`.

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: All training and evaluation details are provided in the experiment section and Appendix B.

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: [No]

Justification: Error bars are not reported because repeatedly generating 4K long videos with multiple random seeds is computationally prohibitive.

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 experiment section and Appendix F describe the hardware (NVIDIA RTX PRO 6000 Blackwell GPUs), training iterations, and GPU memory usage.

   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 presented in this paper adheres to the NeurIPS Code of Ethics. All datasets used are properly cited or created from scratch, no human subjects or sensitive personal data are involved, and the work does not pose foreseeable risks related to safety, fairness, privacy, or misuse.

   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: Appendix G discusses positive impacts (e.g., lower compute barriers and accessibility applications) and negative impacts (e.g., disinformation and non-consensual synthetic media).

    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: [Yes]

    Justification: Safeguards will be implemented and detailed upon the release of the trained model. These include measures such as gated access, usage restrictions, and output watermarking to mitigate risks of misuse. The paper acknowledges potential concerns around video generation, and commits to responsible deployment practices aligned with community standards.

    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 paper uses the UltraVideo dataset and the Wan2.1-T2V-1.3B model, both under open-source licenses. These are cited in the references and licensed appropriately.

    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: [No]

    Justification: Code and model weights will be released with documentation at `https://github.com/ziyang1106/AtlasVid`.

    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: Not applicable; no human subjects research is involved.

    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: We use LLM to polish our paper writing, i.e. grammar, wording.

    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.