ScrollScape: Unlocking 32K Image Generation With Video Diffusion Priors

Haodong Yu[1], Yabo Zhang[1], Donglin Di[2], Ruyi Zhang[1], and Wangmeng Zuo[1] *

[1] Harbin Institute of Technology, Harbin, China
24S103334@stu.hit.edu.cn,wmzuo@hit.edu.cn
[2] Li Auto, Beijing, China

Fig. 1: Imagery of ScrollScape. ScrollScape reformulates high resolution synthesis at extreme aspect ratios such as 8 : 1 as a sequential video panning task. Leveraging robust video diffusion priors, it achieves exceptional 32K resolution across canvases ranging from traditional scrolls to photorealistic panoramas.

Abstract. While diffusion models excel at generating images with conventional dimensions, pushing them to synthesize ultra-high-resolution imagery at extreme aspect ratios (EAR) often triggers catastrophic structural failures, such as object repetition and spatial fragmentation. This limitation fundamentally stems from a lack of robust spatial priors, as static text-to-image models are primarily trained on image distributions with conventional dimensions. To overcome this bottleneck, we present ScrollScape, a novel framework that reformulates EAR image synthesis into a continuous video generation process through two core innovations. By mapping the spatial expansion of a massive canvas to the temporal evolution of video frames, ScrollScape leverages the inherent

* Corresponding author.

temporal consistency of video models as a powerful global constraint to ensure long-range structural integrity. Specifically, Scanning Positional Encoding (ScanPE) distributes global coordinates across frames to act as a flexible moving camera, while Scrolling Super-Resolution (ScrollSR) leverages video super-resolution priors to circumvent memory bottlenecks, efficiently scaling outputs to an unprecedented 32K resolution. Fine-tuned on a curated 3K multi-ratio image dataset, ScrollScape effectively aligns pre-trained video priors with the EAR generation task. Extensive evaluations demonstrate that it significantly outperforms existing image-diffusion baselines by eliminating severe localized artifacts. Consequently, our method overcomes inherent structural bottlenecks to ensure exceptional global coherence and visual fidelity across diverse domains at extreme scales.

Keywords: Ultra Image Generation · Video Diffusion Priors

1 Introduction

Diffusion models [9, 10, 16, 20, 23–25, 28, 31, 40, 48–50, 55–57, 62, 64] have achieved remarkable success in synthesizing imagery with high quality and photorealistic detail. However, even the most preeminent models, such as Nano Banana Pro [52] and GPT-Image [36], frequently suffer from catastrophic structural failure and fidelity loss when the target output deviates from standard dimensions. This limitation is particularly evident when these models are pushed to generate ultra-high-resolution imagery with extreme aspect ratios (EAR), such as landscape panoramas or traditional long scroll paintings. We attribute this bottleneck to the fact that these foundational models are predominantly optimized on datasets with modest resolutions and conventional proportions. Consequently, they lack the robust spatial priors and long-range dependency modeling necessary to maintain global coherence across the massive, non-standard canvases required for such tasks.

Existing methods [2,4,8,17,19,26,34,40] attempt to modify the inference process or architecture of text-to-image diffusion models to synthesize EAR imagery at ultra high resolution. For the generation of extreme aspect ratios (*e.g.*, $8:1$), Sync Diffusion [29] partitions the target space into overlapping patches, which are subsequently processed piece-by-piece by the base model. However, this localized generation paradigm inherently struggles to guarantee global structural coherence, often leading to fragmented compositions. Furthermore, to achieve higher spatial resolutions, techniques like ScaleCrafter [19] and DyPE [26] manipulate the internal representations by dilating convolution modules and interpolating position embeddings. Although these adjustments expand the capacity for large-scale synthesis, the generation process remains highly unstable, frequently introducing severe artifacts such as object repetition. In essence, these training-free modifications demonstrate that relying solely on spatial priors derived from conventional image distributions is fundamentally insufficient to guarantee structural coherence at either ultra-high resolutions or extreme aspect ratios.

To break free from the inherent bottlenecks of purely image diffusion priors, we reformulate ultra-high-resolution EAR synthesis from a static image generation problem into a temporal video generation task. The core intuition is straightforward: we conceptualize the generation of a massive image as a continuous scanning process, translating the spatial layout of the canvas into a sequence of video frames. This paradigm shift enables us to harness the robust spatiotemporal priors of pre-trained video diffusion models [3,9,25,28,55–57,62,64]. Unlike static text-to-image approaches that inevitably suffer from localized fragmentation and object repetition, the inherent temporal consistency of video models acts as a powerful global constraint, naturally ensuring long-range structural integrity and narrative continuity across massive canvases.

To seamlessly adapt video diffusion priors for EAR synthesis, we introduce ScrollScape, powered by Scanning Positional Encoding (ScanPE) and Scrolling Super-Resolution (ScrollSR). First, standard video models assign identical spatial coordinates to every temporal frame. This rigid configuration anchors the generation process to a stationary viewpoint, rendering the model inherently incapable of panning across a massive canvas. ScanPE overcomes this by distributing the global spatial coordinates of the target image across successive video frames. This re-engineering allows ScrollScape to act as a moving camera, synthesizing arbitrary aspect ratios through flexible scanning strategies. Second, to tackle the immense computational cost of ultra-high-resolution generation, we propose ScrollSR. Rather than directly synthesizing massive pixels, ScrollSR leverages video super-resolution diffusion priors [65, 66] to refine the content frame-by-frame. This efficient approach circumvents memory bottlenecks, successfully scaling the final output to an unprecedented 32K. Finally, to train ScrollScape for EAR generation, we curated a lightweight dataset of approximately $3,000$ high-resolution, multi-ratio images. Empirical evidence confirms that this highly curated data is sufficient to perfectly align the pre-trained video model with the EAR synthesis objective.

Extensive evaluations demonstrate that ScrollScape achieves unprecedented high-fidelity EAR synthesis at up to 32K resolution, significantly outperforming existing approaches based on image diffusion models. Most notably, our method fundamentally overcomes the severe artifacts plaguing these baselines, including localized content repetition and structural fragmentation. Consequently, it ensures exceptional global coherence across diverse visual domains, ranging from photorealistic landscapes to stylized artistic scrolls.

To summarize, our key contributions are three-fold:

- We introduce ScrollScape, a novel framework that reformulates EAR image synthesis as a sequential video panning process, effectively harnessing the video diffusion priors to ensure long-range structural coherence.
- We propose Scanning Positional Encoding (ScanPE) to re-engineer coordinate distributions for flexible, continuous generation across arbitrary aspect ratios, alongside a Scrolling Super-Resolution (ScrollSR) module that leverages video super-resolution diffusion priors frame-by-frame to achieve an unprecedented 32K resolution.

– Extensive experiments demonstrate that ScrollScape successfully eliminates severe artifacts like object repetition and fragmentation. Furthermore, we contribute a curated dataset of 3,000 high-resolution, multi-ratio imagery to facilitate model adaptation and future research.

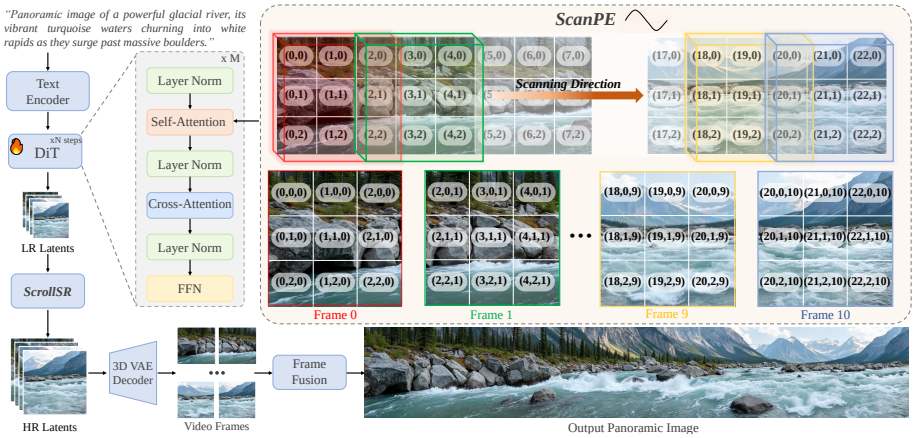

Fig. 2: Overview of ScrollScape Framework. ScrollScape reformulates high resolution EAR synthesis as a sequential video panning task. ScanPE re engineers coordinate distributions by mapping global spatial indices (x, y) onto a temporal sequence to ensure structural coherence across massive scales without the repetition typical of standard models. After generating low resolution latents via a hierarchical DiT, the ScrollSR module utilizes video super-resolution priors to refine the latent sequence while preserving sequential consistency. Finally, a 3D VAE decoder and frame fusion stage produce seamless, photorealistic panoramas and traditional scrolls at an exceptional 32K resolution.

2 Related Work

2.1 High Resolution EAR Synthesis

The synthesis of high resolution imagery with extreme aspect ratios (*e.g.*, $1:8$) remains a significant challenge for text-to-image models [2, 11, 14, 27, 35, 41, 46, 58, 59]. For extreme aspect ratio generation, existing methods [4, 29, 34] partition the target space into overlapping patches. However, this localized generation paradigm inherently lacks a unified global structural prior, leading to the spatial fragmentation and disjointed compositions typical of window based methods.

Furthermore, to achieve higher spatial resolutions, techniques like ScaleCrafter [19] and DyPE [26] manipulate internal representations through dilated convolutions or positional interpolation. While these adjustments expand the synthesis capacity, they fail to resolve the stationary bias of static architectures. Consequently, the generation remains unstable, frequently triggering catastrophic structural failures such as object repetition when pushed beyond the model's original training distribution. In essence, these modifications demonstrate that relying solely on stationary spatial priors of static text-to-image

models is fundamentally insufficient to guarantee global structural coherence at extreme scales. This technological bottleneck necessitates our paradigm shift toward leveraging spatiotemporal video diffusion priors for high resolution EAR synthesis at an unprecedented 32K scale.

2.2 Video Diffusion Models as Generative Priors

Video diffusion models [9, 10, 16, 20, 25, 28, 42, 43, 45, 55–57, 62, 64] have demonstrated a significantly stronger capacity for simulating real world dynamics compared to image diffusion models. While primarily designed for video generation, recent studies have begun to harness these robust priors for image centric applications, such as multi view synthesis [30, 54] and image editing [1, 44, 63].Specifically, Frame2Frame [44], Magic Fixup [1], and Framepainter [63] typify this trend by reformulating image editing as a temporal trajectory or a video-based propagation process. Frame2Frame [44], for instance, leverages a generative video model to create a continuous path from an initial state to a target state, thereby ensuring zero-shot coherence throughout the editing process.

However, existing explorations predominantly focus on leveraging temporal consistency for frame-to-frame transitions. The potential of these priors to regulate long distance spatial consistency as well as the utilization of video super-resolution priors for massive scaling remains largely underexplored. ScrollScape diverges from these approaches by uniquely dual purposing video knowledge. We first utilize video diffusion priors to generate structurally coherent extreme aspect ratio (EAR) imagery, effectively transforming temporal continuity into a spatial regulator. Building on this foundation, we are the first to leverage video super-resolution (VSR) diffusion priors to scale the results to an unprecedented 32K resolution.

2.3 Positional Encoding in Transformers

The structural coherence of Diffusion Transformers (DiT) [18, 33, 38, 53] is fundamentally governed by their positional encoding [47] schemes. In video DiTs, Rotary Positional Embedding (RoPE) [12, 13, 39, 51] has become the de facto standard due to its ability to model relative spatiotemporal dependencies. By representing positions as rotations in the embedding space, RoPE ensures that attention weights depend strictly on relative displacement. The efficacy of RoPE is highly susceptible to sequence length, particularly in high resolution EAR synthesis where inference scales often exceed the original training distribution.

As positional indices shift into out of distribution phases, the model's global structural awareness is compromised. This distributional shift frequently manifests as object repetition and spatial fragmentation, mirroring the structural failures seen in static architectures when pushed beyond their design limits. Standard 3D-RoPE is typically restricted to a static viewport, anchoring spatial indices to fixed frame centric grids. This invariant mapping fails to guide the global spatial evolution required for expansive canvases, particularly within a sliding window paradigm that must track global trajectories while maintaining

stable biases. To bridge this gap, we introduce ScanPE, a flexible coordinate system that re-engineers positional distributions to enable the moving camera vision for high resolution EAR synthesis.

3 Method

We propose ScrollScape, a framework that reformulates ultra-high-resolution EAR image synthesis as sequential video panning. By mapping the spatial extension of long-scroll imagery to the temporal domain, ScrollScape leverages video diffusion priors for coherent generation (Sec. 3.1). We further introduce Scanning Positional Encoding (ScanPE) to maintain global spatial consistency (Sec. 3.2), and ScrollSR to exploit video super-resolution priors for training-free resolution enhancement (Sec. 3.3).

3.1 Ultra Image Generation as Video Scanning

High resolution EAR synthesis aims to generate content across extreme aspect ratios such as digital scrolls and panoramic landscapes. The fundamental challenge lies in maintaining global coherence across an exceptionally large spatial span. We observe that the spatial extension of such imagery can be effectively modeled as a temporal progression. Since video sequences naturally capture the structural continuity of the real world, video diffusion priors provide the ideal foundation for maintaining consistency across extended canvases. To this end, we reformulate EAR synthesis as a sequential generation task by projecting spatial coordinates onto the temporal domain. Unlike traditional text-to-image frameworks, ScrollScape leverages video diffusion priors to model spatial extension as a continuous temporal progression. Specifically, given a pretrained VAE, we perform synthesis in the latent space. Let $z \in \mathbb{R}^{H\times W\times C}$ denote the latent representation of the target panoramic canvas. Instead of generating z at once, we decompose it into a sequence of T overlapping local windows, denoted as $\mathcal{S} = \{z_t\}_{t=1}^{T}$, where each $z_t \in \mathbb{R}^{h\times l\times C}$ corresponds to the content observed at the t-th scanning step. To place these local windows on a shared canvas, we assign each window a global anchor $\mathbf{O}_t$. For a standard linear scan, the anchor of the t-th window is defined as

$$\mathbf{O}_t = \mathbf{P}_{\text{init}} + (t-1)\Delta\mathbf{d}, \quad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{P}_{\text{init}}$ is the initial position, Δ is the scanning stride, and $\mathbf{d}$ is the scanning direction. Horizontal panorama generation is a special case with $\mathbf{d} = (0, 1)$, where the horizontal offset reduces to $\phi_t = (t-1)\Delta$. The t-th local window therefore covers the latent-space region

$$\Omega_t = [O_t^h, O_t^h + h) \times [O_t^w, O_t^w + l). \tag{2}$$

By binding each temporal step to a specific spatial anchor, this formulation turns what would otherwise be a repetitive tiling process into a coherent scanning trajectory across the global canvas.

3.2 Scanning Positional Encoding

Although sequential video panning enables high resolution EAR synthesis, standard 3D RoPE lacks the precise global coordinate control required to track spatial evolution across successive frames. This deficiency forces the model to operate within a frame centric coordinate system, limiting its ability to maintain long range structural integrity across massive canvases. To resolve this, we introduce Scanning Positional Encoding (ScanPE), which generalizes the static position $\mathbf{p}$ into a dynamic scanning trajectory. By mapping tokens into a unified global coordinate system through spatial displacements, ScanPE re engineers the model to act as a moving camera. The linear scan in Eq. (1) can be further generalized to flexible scanning trajectories by allowing the scanning direction to vary across steps. The global anchor position is formulated as

$$\mathbf{O}_t = \sum_{k=1}^{t-1} \Delta \cdot \mathbf{d}_k + \mathbf{P}_{\text{init}}, \quad \text{with } \mathbf{O}_1 = \mathbf{P}_{\text{init}}, \tag{3}$$

where Δ denotes the scanning stride and $\mathbf{d}_k \in \{(\pm 1, 0), (0, \pm 1)\}$ denotes the unit scanning direction at step k. When $\mathbf{d}_k$ is fixed for all steps, Eq. (3) reduces to the linear scan in Eq. (1). When $\mathbf{d}_k$ switches periodically at row or column boundaries, it forms a snake-like trajectory for covering two-dimensional or non-standard canvases. For a token located at the local coordinate $\mathbf{p}_{\text{loc}} = (h_{\text{loc}}, w_{\text{loc}})$ within the t-th window, ScanPE maps it to the global canvas coordinate by

$$\mathbf{P}_g(t, \mathbf{p}_{\text{loc}}) = \mathbf{p}_{\text{loc}} + \mathbf{O}_t. \tag{4}$$

Here $\mathbf{P}_g = (H_g, W_g)$ denotes the global latent coordinate. In this way, tokens from different scanning steps no longer share the same frame-centric spatial indices, but are placed according to their actual locations on the global canvas. We then apply RoPE separately along the temporal, vertical, and horizontal axes. Specifically, each attention head is divided into three subspaces for t, H_g, and W_g, and the final positional embedding is obtained by concatenating the axis-wise rotary embeddings:

$$\mathcal{R}(t, H_g, W_g) = \Theta_t(t; \theta^t) \oplus \Theta_h(H_g; \theta^h) \oplus \Theta_w(W_g; \theta^w), \tag{5}$$

where $\oplus$ denotes channel-wise concatenation, Θ_t, Θ_h, and Θ_w denote the RoPE rotation operators for the temporal, vertical, and horizontal axes, respectively. While ScanPE provides global coordinate control, directly applying video diffusion priors to EAR synthesis still introduces a distributional mismatch. To mitigate this issue, we perform a lightweight alignment stage using conditional flow matching. Given a data latent $z^{\text{data}} \sim p_{\text{data}}$ and Gaussian noise $\epsilon \sim \mathcal{N}(0, I)$, we define the interpolation path as

$$z_\tau = (1 - \tau)\epsilon + \tau z^{\text{data}}, \quad \tau \in [0, 1]. \tag{6}$$

The model is trained to predict the velocity from noise to data:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{z^{\text{data}}, \epsilon, \tau} \left[\left\| v_\theta(z_\tau, \tau, c, \mathcal{R}) - (z^{\text{data}} - \epsilon) \right\|_2^2 \right], \tag{7}$$

where c denotes the text prompt and $\mathcal{R}$ denotes the ScanPE embedding. This alignment stage adapts the pretrained video prior to EAR distributions while preserving its spatiotemporal regularization.

3.3 Scrolling Super Resolution

To elevate visual fidelity, we introduce Scrolling Super-Resolution (ScrollSR), which performs high-resolution refinement directly in the latent space. Rather than directly synthesizing the full 32K canvas in pixel space, ScrollSR first operates on the low-resolution latent sequence produced by the video panning model and then refines it into high-resolution latent representations. This latent-space design avoids the memory overhead of direct ultra-high-resolution pixel generation while preserving the temporal and spatial continuity inherited from the video prior. While latent-space refinement ensures computational efficiency, the subsequent translation into the pixel domain can still be affected by the 3D VAE, which may introduce flickering or boundary artifacts when decoding adjacent latent windows with significant appearance variations. To reduce these artifacts, we use a coordinate-aligned latent decoding strategy that is consistent with the global scanning coordinates defined by ScanPE. Instead of processing the high-resolution latent sequence as a uniform temporal stream, we group latent blocks according to their global anchors $\{\mathbf{O}_t\}_{t=1}^{T}$ before 3D VAE decoding. For each group, an anchor block provides a stable decoding context, while neighboring blocks are decoded under the same coordinate reference established by ScanPE. This coordinate-aligned latent decoding strategy keeps the decoder context consistent across adjacent windows without requiring additional training. As a result, ScrollSR produces high-fidelity outputs with fewer boundary artifacts and geometric distortions while preserving global structural alignment.

3.4 Panoramic Frame Fusion

Following coordinate-aligned latent-to-pixel reconstruction, we integrate the generated sequential blocks into a unified 32K panorama via a weighted fusion strategy. This stage resolves residual inconsistencies across overlapping sliding windows while suppressing flickering and transient disturbances introduced during sequential decoding. For each spatial window t, suppose it produces F candidate decoded tiles $\{I_{t,i}\}_{i=1}^{F}$ under different local decoding contexts. We select a representative tile based on median feature consensus. Let $f(I_{t,i})$ denote the feature statistic of the i-th candidate in block t. We first compute the channel-wise median feature

$$\mathbf{m}_t = \mathrm{Median}_{k=1}^{F}\, f(I_{t,k}), \tag{8}$$

and then select the representative tile by

$$i_t^* = \underset{i\in\{1,\dots,F\}}{\arg\min} \left\| f(I_{t,i}) - \mathbf{m}_t \right\|_2, \quad \bar{I}_t = I_{t,i_t^*}. \tag{9}$$

This selection filters out transient outliers while preserving high-frequency structural details. Before fusion, we map the latent-space anchor $\mathbf{O}_t$ to the pixel domain as $\tilde{\mathbf{O}}_t = s\mathbf{O}_t$, where s denotes the overall upsampling factor from latent coordinates to the final pixel coordinates. For a global pixel coordinate $\mathbf{P}$, the final panorama is reconstructed by weighted blending:

$$\mathcal{I}_{\text{pano}}(\mathbf{P}) = \frac{\sum_{t=1}^{T} M_t(\mathbf{P} - \tilde{\mathbf{O}}_t)\bar{I}_t(\mathbf{P} - \tilde{\mathbf{O}}_t)}{\sum_{t=1}^{T} M_t(\mathbf{P} - \tilde{\mathbf{O}}_t) + \epsilon}, \tag{10}$$

where M_t is a distance-based ramp mask that is zero outside the support of the t-th tile, and ϵ avoids division by zero.

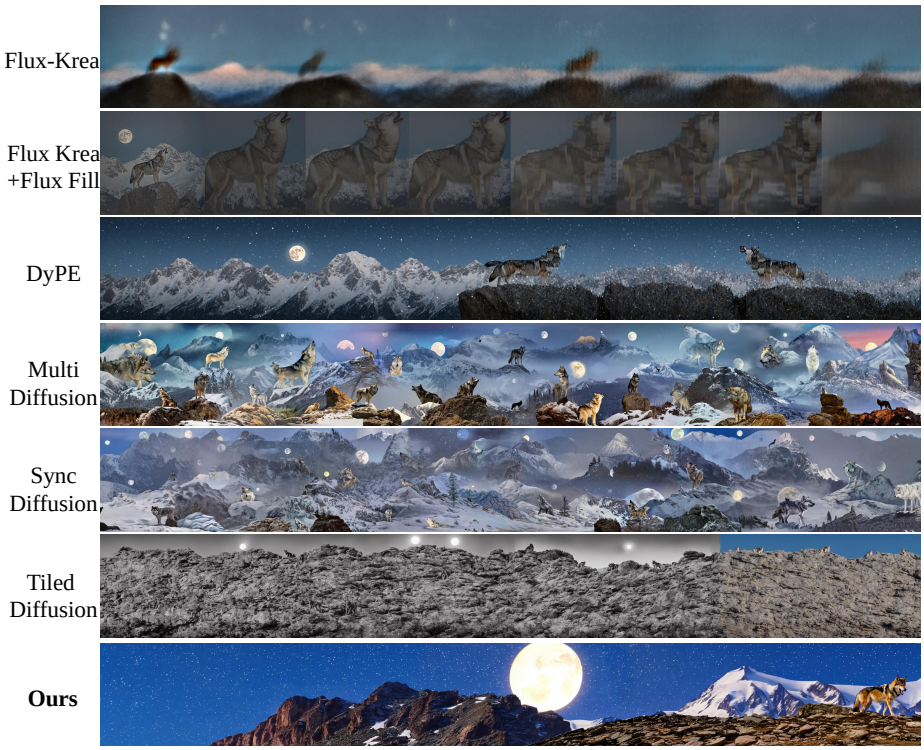

Fig. 3: Qualitative comparison on 8:1 horizontal panoramic scroll imagery. ScrollScape achieves rigorous structural coherence and expansive global diversity, whereas baselines suffer from the semantic repetition and boundary artifacts common in tiled synthesis.

4 Experiments

4.1 Experimental Setup

Dataset Construction. To support the synthesis of EAR content, we curate a specialized training dataset consisting of 3,000 high-quality panoramic samples. Specifically, this includes 2,000 natural landscape images with resolution

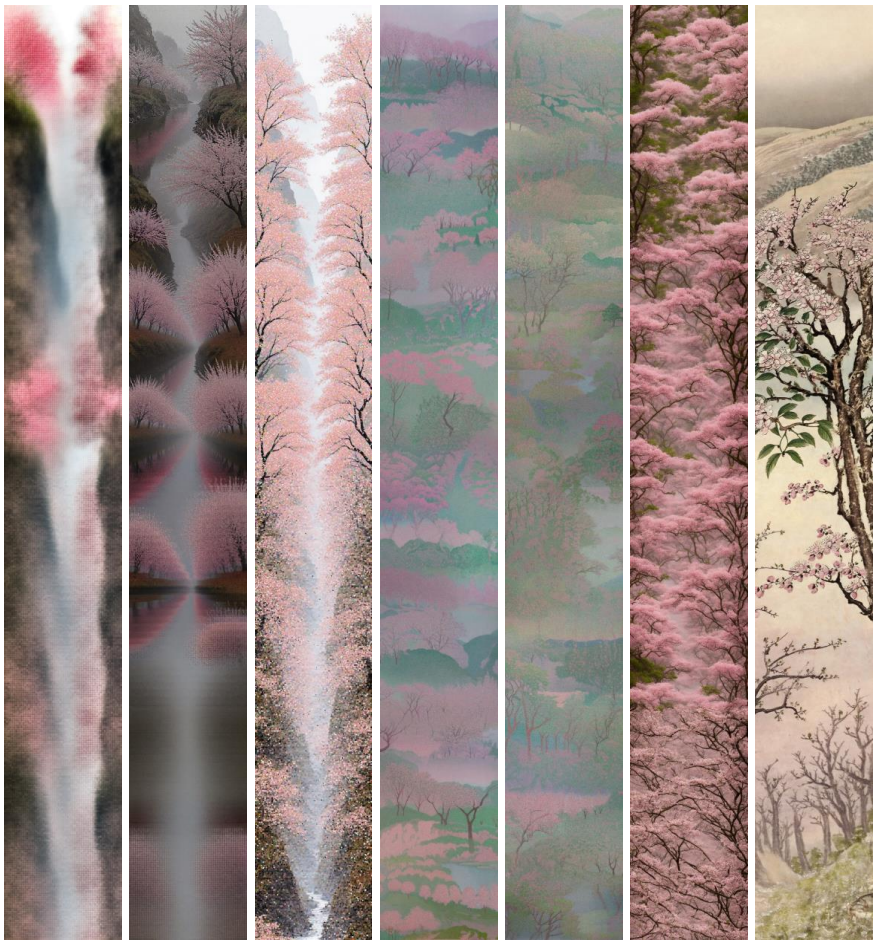

Fig. 4: Qualitative results on 8:1 vertical scroll paintings. ScrollScape maintains superior global structural coherence and expansive diversity, effectively eliminating the boundary artifacts and semantic repetition found in existing frameworks.

ratios of 6:1 or higher and 1,000 traditional Chinese landscape paintings maintained at a consistent 6:1 ratio. These samples provide the model with essential compositional priors for continuous, horizontal-flowing structures, ranging from geological vistas to traditional ink-wash scroll aesthetics.

Implementation Details. We initialize ScrollScape with Wan2.1-T2V-1.3B [55] and finetune it on our collected training samples. We train the model on two A100 GPUs with a total batch size of 4. The model is optimized for $10,000$ iterations with a learning rate of 1×10^{-5} using the AdamW [32] algorithm. During inference, the model first generates a low-resolution latent sequence to maintain

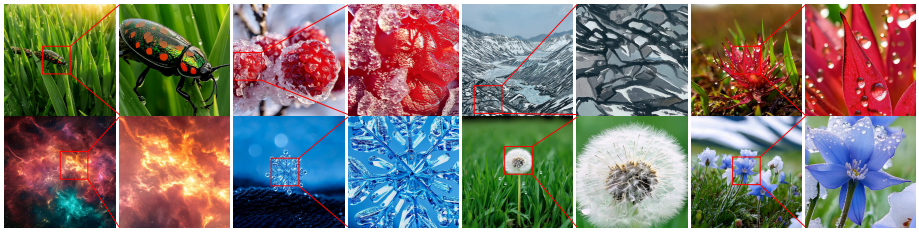

Fig. 5: Demonstration of high fidelity 8K imagery generated by ScrollScape. The results showcase the versatility of **ScrollScape** in producing high fidelity imagery with structural richness and local clarity across various subject matters, ranging from microscopic textures to macroscopic landscapes. *Please zoom in for high resolution details.*

global coherence across the scanning trajectory. The modified FlashVSR module is then applied in the latent space to refine the low-resolution latents into high-resolution latent representations. These refined latents are decoded by the 3D VAE and fused at the pixel level to reach the final 32K resolution. We use a linear ramp function for the pixel-level fusion of overlapping tiles to ensure a seamless transition across the expansive panorama.

4.2 Comparison with Baselines

Qualitative Results. Figures 3 and 4 present a qualitative comparison of panoramas synthesized at an extreme aspect ratio of 8:1, with the short side resolution maintained at 1024 pixels. While existing methods struggle to maintain quality under such constraints, ScrollScape demonstrates a significant leap in both structural integrity and content richness. Specifically, diffusion based approaches such as DyPE [26], FLUX.1-Krea-dev [7], and FLUX.1-Fill-dev [6] exhibit severe visual artifacts and geometric distortions once the aspect ratio exceeds 4:1. Meanwhile, tiling based frameworks, including MultiDiffusion [4], SyncDiffusion [29], and Tiled Diffusion [34], suffer from conspicuous content repetition and irrational global structures, failing to establish a coherent narrative flow. In contrast, ScrollScape maintains robust structural coherence and diverse, non redundant details across various themes. This performance demonstrates a superior balance between local fidelity and global architectural logic, significantly outperforming competitive baselines.

As shown in Fig. 5, ScrollScape maintains global structural integrity and high fidelity details at 8K resolution. Beyond panoramic synthesis, the hybrid mode ensures standard ratio generation with consistent structural layout and clarity. Zoomed in examples, such as beetle carapaces and ice crystals, demonstrate sharp symmetry and intricate textures, proving that ScrollScape preserves global architectural logic alongside fine grained details within a single unified architecture.

Table 1: Quantitative comparison. ScrollScape surpasses alternative approaches across all metrics. The best results are **bolded**.

Method	FID ↓	CLIP ↑	KID ↓ ($\times 10^{-2}$)	Style-L ↓ ($\times 10^{-3}$)	GSD ↑ (LPIPS)	GSD ↓ (DINOv2)
FLUX-Krea	333.7	20.9	20.4	6.0	0.174	0.822
FLUX-Krea+Fill	281.4	19.1	4.9	13.5	0.293	0.841
DyPE	248.1	24.6	4.7	5.5	0.569	0.682
MultiDiffusion	261.7	29.7	3.7	5.0	0.658	0.902
SyncDiffusion	245.2	26.5	3.2	4.7	0.618	0.895
Tiled Diffusion	241.2	27.3	3.0	4.5	0.591	0.901
ScrollScape (Ours)	**214.7**	**30.0**	**2.0**	**4.0**	**0.674**	**0.670**

Quantitative Results. As demonstrated in Tab. 1, we adopt a patch based assessment strategy to mitigate geometric distortion and aspect ratio bias during evaluation. Specifically, we partition the 8 : 1 panoramas into non-overlapping spatial grids with a 1:1 aspect ratio. Local fidelity and distribution similarity are quantified by averaging FID [22] and KID [5] scores across these patches. Specifically, both metrics are computed by comparing the patches against images from Aesthetic Eval [60] after resizing them to match the corresponding dimensions. Semantic alignment is evaluated via CLIP scores [21]. To ensure seamless junctions, we assess stylistic uniformity and visual continuity using Intra Style Loss [15] between adjacent grids. We introduce Global Structural Diversity (GSD) to quantify long range architectural variety and detect mode repetition across the panoramic manifold. Rather than focusing on local fidelity, GSD evaluates the correlation between distant spatial grids via two complementary perspectives. Specifically, we utilize LPIPS [61] distances to capture perceptual texture variation and DINOv2 [37] feature cosine similarity to assess semantic non redundancy. By identifying high feature similarity between spatially separated regions, GSD effectively exposes the semantic looping and object duplication inherent in tiling based frameworks. Ultimately, this metric confirms that ScrollScape preserves a consistent artistic style while continuously synthesizing unique and varied structural elements throughout the entire expansive sequence. The experimental results show that ScrollScape maintains high quality details while effectively avoiding the content repetition typical of traditional generative methods.

User Study. To evaluate the perceptual quality of ScrollScape, we conducted a user study utilizing 100 diverse prompts spanning multiple themes. For each prompt, we generated panoramas with an 8:1 aspect ratio in both horizontal and vertical orientations to compare our method against existing baselines. Twenty independent raters performed anonymized side by side evaluations across three primary dimensions: (i) structural coherence for logical and seamless global structure, (ii) content richness for varied details without redundancy, and (iii)

Table 2: User Preference Study. The percentages denote the fraction of raters who prefer our results over each baseline.

Method Comparison	Structural Coh.	Content Rich.	Image Qual.
Ours vs DyPE	92%	89%	87%
Ours vs MultiDiffusion	83%	76%	82%
Ours vs SyncDiffusion	85%	74%	78%
Ours vs Tiled Diffusion	76%	85%	74%

Table 3: Quantitative ablation study. While Wan2.1 achieves an outstanding Style-L score due to mode repetition, ScrollScape demonstrates superior performance across most metrics, ensuring both high fidelity and global structural diversity.

Method	FID ↓	CLIP ↑	KID ↓ ($\times 10^{-2}$)	Style-L ↓ ($\times 10^{-3}$)	GSD ↑ (LPIPS)	GSD ↓ (DINOv2)
Wan2.1	246.5	25.9	4.7	**2.0**	0.303	0.975
w/o Training	272.0	24.6	6.3	7.6	0.569	0.858
Vanilla latents	231.6	25.6	3.2	6.0	0.469	0.730
w/o ScrollSR	218.8	26.5	2.4	4.9	0.668	0.671
ScrollScape (Ours)	**214.7**	**30.0**	**2.0**	4.0	**0.674**	**0.670**

Fig. 6: Qualitative ablation study on the effectiveness of ScanPE. From top to bottom: (a) the vanilla Wan2.1 exhibiting severe content repetition; (b) w/o training, where ScanPE successfully breaks repetitive patterns but yields chaotic textures; (c) vanilla latents showing significant blurring due to decoding limitations; and (d) ScrollScape producing seamless panoramas with structural coherence and rich details.

image quality for visual fidelity and texture clarity. As summarized in Tab. 2, ScrollScape is significantly preferred over baselines across all three perspectives.

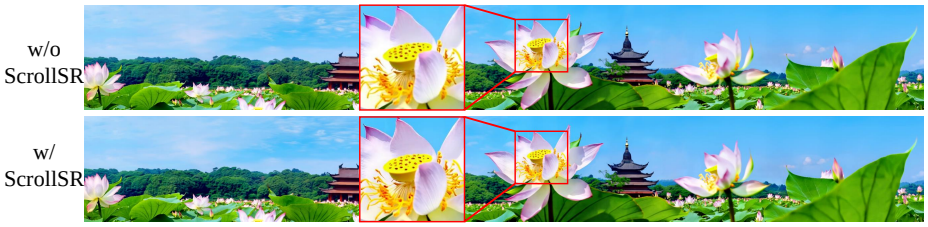

Fig. 7: Qualitative ablation study on the effectiveness of ScrollSR. The baseline w/o ScrollSR (top) suffers from perceptual blurriness in the lotus petals. In contrast, our full framework w/ ScrollSR (bottom) successfully restores fine grained textures and structural definition, as evidenced by the sharp details and high resolution clarity in the zoomed in patches.

4.3 Ablation Study

As illustrated in Fig. 6, the vanilla Wan2.1 model suffers from severe content repetition when generating sequences at extreme aspect ratios. Each segment appears as a near identical replica of the preceding one, failing to establish a meaningful spatial progression. The variant w/o Training underscores the vital role of ScanPE in maintaining structural diversity. By successfully breaking the repetitive patterns observed in the baseline, this variant proves that ScanPE provides a superior structural foundation for EAR synthesis. However, without the proposed training phase, the model produces chaotic textures and significant visual noise, demonstrating that training is essential to harmonize video diffusion priors with the high resolution static domain.

Similarly, the results of vanilla latents show noticeable blurring and structural degradation. This indicates that directly decoding extended latent sequences with a standard 3D VAE is insufficient to preserve fine-grained details across adjacent windows. In contrast, our full model achieves exceptional structural coherence and content richness. As evidenced in Fig. 7, the w/o ScrollSR variant maintains global structural integrity but fails to achieve perceptual sharpness, which results in the visibly soft foreground textures. Our full model combines ScanPE, coordinate-aligned latent decoding, and ScrollSR to reduce spatial misalignment and decoding artifacts. This combination yields non-redundant panoramas with the high-resolution clarity required for photorealistic synthesis.

Tab. 3 also highlights the effectiveness of each proposed module by providing a detailed numerical breakdown of their contributions. The quantitative results consistently align with our qualitative observations, reinforcing the necessity of our integrated approach for high-resolution EAR synthesis. For instance, consistent with previous observations, Wan's lower Style-L score is primarily due to the generation of a singular, repeated pattern.

5 Conclusion

In this paper, we reframe ultra-high-resolution EAR synthesis as a sequential video panning task and introduce ScrollScape to leverage the robust spatiotemporal priors of video diffusion models. Built upon Wan2.1-T2V-1.3B, ScrollScape effectively eliminates object repetition and spatial fragmentation while ensuring seamless global coherence. To address the limitations of static spatial indexing, we propose ScanPE to act as a moving camera and ScrollSR to efficiently scale outputs to an unprecedented 32K resolution. Our experiments highlight the effectiveness of this approach, which significantly outperforms existing image diffusion baselines across diverse domains. Furthermore, our framework demonstrates strong generalization by maintaining high fidelity details from microscopic textures to macroscopic landscapes. We hope our work will inspire further exploration of video priors as spatial regulators for high resolution generative tasks.

Acknowledgements

This work was supported by the National Cyber Security-National Science and Technology Major Project under Grant No. 2026ZD1500500.

References

1. Alzayer, H., Xia, Z., Zhang, X., Shechtman, E., Huang, J.B., Gharbi, M.: Magic fixup: Streamlining photo editing by watching dynamic videos. ACM Transactions on Graphics **44**(5), 1–25 (2025)
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., et al.: ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
3. Bar-Tal, O., Chefer, H., Tov, O., Herrmann, C., Paiss, R., Zada, S., Ephrat, A., Hur, J., Liu, G., Raj, A., et al.: Lumiere: A space-time diffusion model for video generation. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
4. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation
5. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. arXiv preprint arXiv:1801.01401 (2018)
6. Black Forest Labs: FLUX Fill. `https://docs.bfl.ml/flux_tools/flux_1_fill/` (2024), accessed: 2026-02-27
7. Black Forest Labs: Flux Krea. `https://bfl.ai/blog/flux-1-krea-dev` (2024), accessed: 2026-02-27
8. Black Forest Labs: FLUX.1. `https://blackforestlabs.ai/` (2024), accessed: 2026-02-27
9. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)
10. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., et al.: Video generation models as world simulators. OpenAI Blog **1**(8), 1 (2024)

11. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M.H., Murphy, K., Freeman, W.T., Rubinstein, M., et al.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint arXiv:2301.00704 (2023)
12. Chen, S., Wong, S., Chen, L., Tian, Y.: Extending context window of large language models via positional interpolation. arXiv preprint arXiv:2306.15595 (2023)
13. Ding, Y., Zhang, L.L., Zhang, C., Xu, Y., Shang, N., Xu, J., Yang, F., Yang, M.: Longrope: Extending llm context window beyond 2 million tokens. arXiv preprint arXiv:2402.13753 (2024)
14. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
15. Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. arXiv preprint arXiv:1508.06576 (2015)
16. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. arXiv preprint arXiv:2307.04725 (2023)
17. Haji-Ali, M., Balakrishnan, G., Ordonez, V.: 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. pp. 6603–6612 (2024)
18. Hatamizadeh, A., Song, J., Liu, G., Kautz, J., Vahdat, A.: Diffit: Diffusion vision transformers for image generation. In: European Conference on Computer Vision. pp. 37–55. Springer (2024)
19. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In: The Twelfth International Conference on Learning Representations (2023)
20. Henschel, R., Khachatryan, L., Poghosyan, H., Hayrapetyan, D., Tadevosyan, V., Wang, Z., Navasardyan, S., Shi, H.: Streamingt2v: Consistent, dynamic, and extendable long video generation from text. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 2568–2577 (2025)
21. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 7514–7528 (2021)
22. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
23. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models (2020)
24. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
25. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in neural information processing systems **35**, 8633–8646 (2022)
26. Issachar, N., Yariv, G., Benaim, S., Adi, Y., Lischinski, D., Fattal, R.: Dype: Dynamic position extrapolation for ultra high resolution diffusion. arXiv preprint arXiv:2510.20766 (2025)
27. Kang, M., Zhu, J.Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: CVPR (2023)

28. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
29. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. Advances in Neural Information Processing Systems **36**, 50648–50660 (2023)
30. Li, P., Liu, Y., Long, X., Zhang, F., Lin, C., Li, M., Qi, X., Zhang, S., Xue, W., Luo, W., et al.: Era3d: High-resolution multiview diffusion using efficient row-wise attention. Advances in Neural Information Processing Systems **37**, 55975–56000 (2024)
31. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
32. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
33. Ma, X., Wang, Y., Chen, X., Jia, G., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. arXiv preprint arXiv:2401.03048 (2024)
34. Madar, O., Fried, O.: Tiled diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7795–7804 (2025)
35. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: ICML (2022)
36. OpenAI: GPT-Image (gpt-image-1) model documentation. `https://chatgpt.com/images/` (2025), accessed: 2026-02-27
37. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
38. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
39. Peng, B., Quesnelle, J., Fan, H., Shippole, E.: Yarn: Efficient context window extension of large language models. arXiv preprint arXiv:2309.00071 (2023)
40. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
41. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
42. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International conference on machine learning. pp. 8821–8831. Pmlr (2021)
43. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
44. Rotstein, N., Yona, G., Silver, D., Velich, R., Bensaïd, D., Kimmel, R.: Pathways on the image manifold: Image editing via video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7857–7866 (2025)
45. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)

46. Sauer, A., Karras, T., Laine, S., Geiger, A., Aila, T.: Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. arXiv preprint arXiv:2301.09515 (2023)
47. Shaw, P., Uszkoreit, J., Vaswani, A.: Self-attention with relative position representations. arXiv preprint arXiv:1803.02155 (2018)
48. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: ICML (2015)
49. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
50. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
51. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
52. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
53. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
54. Voleti, V., Yao, C.H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., Jampani, V.: Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In: European Conference on Computer Vision. pp. 439–457. Springer (2024)
55. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
56. Wang, X., Yuan, H., Zhang, S., Chen, D., Wang, J., Zhang, Y., Shen, Y., Zhao, D., Zhou, J.: Videocomposer: Compositional video synthesis with motion controllability. Advances in Neural Information Processing Systems **36**, 7594–7611 (2023)
57. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
58. Yu, J., Xu, Y., Koh, J.Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B.K., et al.: Scaling autoregressive models for content-rich text-to-image generation. arXiv preprint arXiv:2206.10789 (2022)
59. Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.N.: Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 5907–5915 (2017)
60. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23464–23473 (2025)
61. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
62. Zhang, Y., Wei, Y., Jiang, D., Zhang, X., Zuo, W., Tian, Q.: Controlvideo: Training-free controllable text-to-video generation. arxiv 2023. arXiv preprint arXiv:2305.13077
63. Zhang, Y., Zhou, X., Zeng, Y., Xu, H., Li, H., Zuo, W.: Framepainter: Endowing interactive image editing with video diffusion priors. arXiv preprint arXiv:2501.08225 (2025)

64. Zheng, Z., Peng, X., Yang, T., Shen, C., Li, S., Liu, H., Zhou, Y., Li, T., You, Y.: Open-sora: Democratizing efficient video production for all. arXiv preprint arXiv:2412.20404 (2024)
65. Zhou, S., Yang, P., Wang, J., Luo, Y., Loy, C.C.: Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2535–2545 (2024)
66. Zhuang, J., Guo, S., Cai, X., Li, X., Liu, Y., Yuan, C., Xue, T.: Flashvsr: Towards real-time diffusion-based streaming video super-resolution. arXiv preprint arXiv:2510.12747 (2025)