

# ExVideo: Extending Video Diffusion Models via Parameter-Efficient Post-Tuning

Zhongjie Duan<sup>1,2</sup>, Hong Zhang<sup>2,3</sup>, Wenmeng Zhou<sup>2</sup>, Cen Chen<sup>1</sup>, Yaliang Li<sup>2</sup>,  
Yu Zhang<sup>3</sup>, Yingda Chen<sup>2</sup>

<sup>1</sup>East China Normal University

<sup>2</sup>Alibaba Group

<sup>3</sup>Zhejiang University

zjduan@stu.ecnu.edu.cn, {hongzhang99, zhangyu80}@zju.edu.cn,  
{wenmeng.zwm, yaliang.li, yingda.chen}@alibaba-inc.com, cenchen@dase.ecnu.edu.cn

## Abstract

Recently, advancements in video synthesis have attracted significant attention. Video synthesis models have demonstrated the practical applicability of diffusion models in creating dynamic visual content. Despite these advancements, the extension of video lengths remains constrained by computational resources. Most existing video synthesis models are limited to generating short video clips. In this paper, we propose a novel post-tuning methodology for video synthesis models, called ExVideo. This approach is designed to enhance the capability of current video synthesis models, allowing them to produce content over extended temporal durations while incurring lower training expenditures. In particular, we design extension strategies across common temporal model architectures respectively, including 3D convolution, temporal attention, and positional embedding. To evaluate the efficacy of our proposed post-tuning approach, we trained ExSVD, an extended model based on Stable Video Diffusion model. Our approach enhances the model’s capacity to generate up to  $5\times$  its original number of frames, requiring only 1.5k GPU hours of training on a dataset comprising 40k videos. Importantly, the substantial increase in video length doesn’t compromise the model’s innate generalization capabilities, and the model showcases its advantages in generating videos of diverse styles and resolutions. We have released the source code and the enhanced model publicly<sup>1</sup>.

## 1 Introduction

In recent years, diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020] have achieved outstanding results in image synthesis, significantly surpassing previous GANs [Dhariwal and Nichol, 2021]. These achievements have subsequently fostered a burgeoning interest in the adaptation of diffusion models for video synthesis. The diffusion models such as Stable Video Diffusion [Blattmann *et al.*, 2023], Animatediff [Guo *et al.*, 2023], and VideoCrafter2 [Chen *et al.*, 2024a]

epitomize this research trajectory, showcasing the ability to produce frames that are not only coherent but also of high visual quality. These achievements underscore the practicality and potential of employing diffusion models in the field of video synthesis. With the groundbreaking results of SORA [Liu *et al.*, 2024] reported at the beginning of 2024, the research direction of video synthesis has once again attracted widespread attention.

Although current video synthesis models are capable of producing video clips of satisfactory quality, the generated videos are generally short, and extending their duration remains a challenge. Current methodologies can be categorized into three types to generate longer videos. 1) Pre-training using long video datasets [Chen *et al.*, 2024b; Wang *et al.*, 2023b]. Through extensive training with long video samples, it is foreseeable that models can improve their ability to generate longer videos. However, training with such datasets would result in prohibitively escalated costs. Consequently, given the computational constraints, current video generation models are primarily trained on short video clips. 2) Generating videos in a streaming [Kodaira *et al.*, 2023] or sliding window [Duan *et al.*, 2024] manner. Without further training, longer videos can be generated by stitching together several short video segments. However, this approach leads to lower video coherence. In addition, existing video generation models lack the capability for long-term video understanding, making the accumulation of errors inevitable. As a result, during the generation of long videos, the visual quality is prone to deterioration, manifesting as a breakdown in the imagery. 3) Frame interpolation [Huang *et al.*, 2022; Wu *et al.*, 2024]. Video frame interpolation models offer a method to augment the frame count of generated videos. However, this approach is inadequate for extending the narrative timeframe of the video. While it increases the number of frames, maintaining the original frame rate would result in an unnatural slow-motion effect, thereby failing to extend the narrative span of the video content. These challenges underscore the necessity for innovative solutions capable of overcoming the existing hurdles associated with video duration extension, without compromising video quality or coherence.

Recent breakthroughs in the development of LLMs (large language models) [Xiong *et al.*, 2023; Xiao *et al.*, 2023; Chen *et al.*, 2023b] have inspired us. Notably, LLMs, despite

<sup>1</sup><https://ecnu-cilab.github.io/ExVideoProjectPage/>.

being trained on fixed-length data, exhibit remarkable proficiency in understanding contexts of variable lengths. This flexibility is further enhanced through the integration of supplementary components and the application of lightweight training procedures, enabling the processing of exceptionally lengthy texts. Such innovations have motivated us to explore analogous methodologies within video synthesis models. In this paper, we introduce a novel post-tuning strategy, called ExVideo, specifically designed to empower existing video synthesis models to produce extended-duration videos within the constraints of limited computational resources. We have designed an extension structure for mainstream video synthesis model architectures. This framework incorporates adapter components, meticulously engineered to preserve the intrinsic generalization capabilities of the base model. Through post-tuning, we enhance the temporal modules of the model, thereby facilitating the processing of content across longer temporal spans.

In theory, ExVideo is designed to be compatible with the majority of existing video synthesis models. To empirically validate the efficacy of our post-tuning methodology, we applied it to the Stable Video Diffusion model [Blattmann *et al.*, 2023], a popular open-source image-to-video model. Through ExVideo, we can extend the original frame synthesis capacity from a limit of 25 frames to 128 frames. Importantly, this expansion was achieved without compromising the model’s distinguished generative capabilities. Additionally, the enhanced model exhibits the versatility to be seamlessly integrated with text-to-image models [Rombach *et al.*, 2022; Li *et al.*, 2024b; Chen *et al.*, 2023a]. This synergistic amalgamation establishes robust and versatile text-to-video pipelines. This adaptability underscores the potential of our post-training technique, the source code and the extended model will be released publicly. In summary, the contributions of this paper include:

- We present ExVideo, a post-tuning technique for video synthesis models that can extend the temporal scale of existing models to facilitate the generation of long videos.
- Based on Stable Video Diffusion (SVD), we have trained an extended video synthesis model named ExSVD. This model is capable of generating coherent videos of up to 128 frames while preserving the generative capabilities of the original model.
- Through comprehensive empirical experiments, we demonstrate the feasibility of enhancing video synthesis models via post-tuning, thereby presenting an innovative approach to the training of large-scale models for extended video synthesis.

## 2 Related Work

### 2.1 Diffusion Models

Diffusion models [Sohl-Dickstein *et al.*, 2015; Ho *et al.*, 2020] are a category of generative models that characterize the content generation as a Markov random process. Unlike GANs [Goodfellow *et al.*, 2014], diffusion models do not require adversarial training, thus making their training process

more stable. Moreover, through an iterative generation process, diffusion models are capable of producing images with exceptionally high quality. In recent years, image synthesis models based on diffusion, including Pixart [Chen *et al.*, 2023a], Imagen [Saharia *et al.*, 2022], Hunyuan-DiT [Li *et al.*, 2024b], and the Stable Diffusion series [Rombach *et al.*, 2022], have achieved impressive success. Diffusion models have given rise to a vast open-source technology ecosystem. Technologies such as LoRA [Hu *et al.*, 2021], ControlNet [Zhang *et al.*, 2023] and IP-Adapter [Ye *et al.*, 2023] have endowed the generation process of diffusion models with a high degree of controllability, thereby meeting the needs of various application scenarios.

### 2.2 Video Synthesis

Given the remarkable success of diffusion models in image synthesis, video synthesis approaches based on diffusion have also been proposed in recent years. For example, by adding additional motion modules to the UNet model [Ronneberger *et al.*, 2015] in Stable Diffusion [Rombach *et al.*, 2022], AnimateDiff [Guo *et al.*, 2023] transfers the capabilities of image synthesis to video synthesis. Stable Video Diffusion [Blattmann *et al.*, 2023] is an image-to-video model architecture and can synthesize video clips after end-to-end video synthesis training. Unlike image synthesis models, video synthesis models require substantial computational resources since the model needs to process multiple frames simultaneously. As a result, most existing video generation models [Guo *et al.*, 2023; Wang *et al.*, 2023a] can only produce very short video clips. For instance, AnimateDiff can generate up to 32 frames, while Stable Video Diffusion can generate a maximum of 25 frames. This limitation prompts us to explore the methodology to extend video synthesis models over longer temporal scales.

### 2.3 Extending Generative Models

Although the existing diffusion models are trained with a fixed scale, such as Stable Diffusion being trained at a fixed resolution of  $512 \times 512$ , some approaches can extend them to larger scales. For instance, in image synthesis, approaches like Mixture of Diffusers [Jiménez, 2023], MultiDiffusion [Bar-Tal *et al.*, 2023], and ScaleCrafter [He *et al.*, 2023] can increase the resolution of generated images by altering the inference process of the UNet model in Stable Diffusion. Similar techniques have also emerged in the field of large language models. With the help of positional encoding technologies such as RoPE [Su *et al.*, 2024] and ALiBi [Press *et al.*, 2021], large language models can extrapolate to longer text processing tasks under the premise of training with limited-length texts. Post-tuning can further help language models achieve super-long text comprehension and generation [Xiong *et al.*, 2023; Chen *et al.*, 2023b]. These research findings have inspired and motivate us to explore the extension of video synthesis models. We aim to endow existing video synthesis models with the capability to generate longer videos.

## 3 Methodology

In this section, we first review the architectures of mainstream video diffusion models, then discuss the methodologies we

have adopted to extend the temporal modules for long video synthesis, and finally introduce the post-tuning strategy.

### 3.1 Preliminaries

The huge demands of computational resources for training video synthesis models lead to a prevalent practice of adapting existing image synthesis models for video generation. This adaptation is typically achieved by incorporating temporal modules into the model for generating dynamic content. We provide a comprehensive overview of temporal module architectures as follows:

- **3D convolution** [Li *et al.*, 2021]: Convolution layers form the foundational blocks in computer vision. 2D convolution layers have been employed in the UNet [Ronneberger *et al.*, 2015] architecture, which is widely used in diffusion models. By extending 2D convolutions into the third dimension, these layers are seamlessly adapted in video synthesis models. Research indicates that convolution layers in diffusion models exhibit a high degree of adaptability across various resolutions [Bar-Tal *et al.*, 2023], which is a testament to their capacity for generalization.
- **Temporal attention** [Vaswani *et al.*, 2017]: In image synthesis, the importance of attention mechanisms is underscored by their contribution to the generation of images with remarkable fidelity, as evidenced by the ablation studies in latent diffusion [Rombach *et al.*, 2022]. Transferring spatial attention mechanisms to the video domain raises concerns regarding computational efficiency due to the quadratic time complexity of the attention operators. To circumvent this computational bottleneck, advanced video synthesis models typically adopt temporal attention layers [Guo *et al.*, 2023; Blattmann *et al.*, 2023] that optimize efficiency by curtailing the volume of embeddings processed by each attention operator.
- **Positional embedding** [Su *et al.*, 2024]: The native attention layers cannot model the positional information in videos. Therefore, video synthesis models typically incorporate positional embeddings to enrich the embedding space with positional information. Positional embeddings can be instantiated through diverse methodologies. For example, AnimateDiff [Guo *et al.*, 2023] opts for learnable parameters to establish positional embeddings, whereas Stable Video Diffusion [Blattmann *et al.*, 2023] utilizes trigonometric functions to generate static positional embeddings.

### 3.2 Extending Temporal Modules

Most video synthesis models are pre-trained on videos comprising only a constrained number of frames due to limited computational resources. For instance, Stable Video Diffusion [Blattmann *et al.*, 2023] is capable of generating a maximum of 25 frames, while AnimateDiff [Guo *et al.*, 2023] is limited to synthesizing videos of up to 32 frames. To augment these models to produce extended videos, we propose enhancements to the temporal modules within these models.

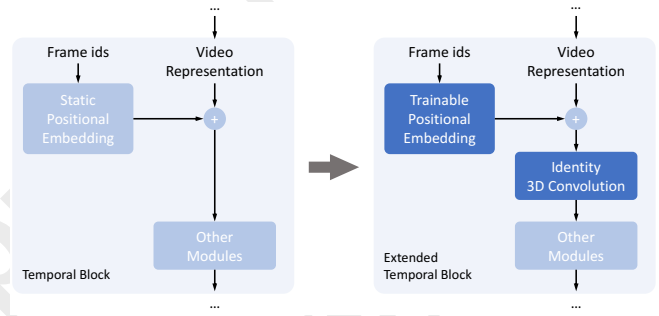


Figure 1: The architecture of extended temporal blocks in Stable Video Diffusion. We replace the static positional embedding with a trainable positional embedding and add an adaptive identity 3D convolution layer to learn long-term video features. The modifications are adaptive, preserving the original generalization abilities of the pre-trained model. All parameters outside the temporal block are fixed while training for lower memory usage.

Firstly, the inherent functionality of 3D convolution layers to adaptively accommodate various scales has been previously validated through empirical studies [Jiménez, 2023; Bar-Tal *et al.*, 2023; He *et al.*, 2023], even without necessitating fine-tuning. Consequently, we opt to retain the 3D convolution layers in their original form to preserve these capabilities. Secondly, regarding the temporal attention modules, research on large language models has demonstrated the potential for scaling existing models to accommodate longer contextual sequences [Xiong *et al.*, 2023; Chen *et al.*, 2023b]. Inspired by these findings, we fine-tune the parameters within the temporal attention layers during the training process to enhance their efficacy over extended frame sequences. Thirdly, for the positional embedding layers, either static or trainable embeddings cannot be directly applied to longer videos. To circumvent this pitfall while ensuring compatibility with a wide array of existing video models, we use extended trainable parameters to replace the original positional embeddings. These extended trainable positional embeddings are initialized in a cyclic pattern, drawing upon the configurations of the pre-existing embeddings. Further, drawing inspiration from various adapter models [Hu *et al.*, 2021; Zhang *et al.*, 2023], we incorporate an additional identity 3D convolution layer after the positional embedding layer, aimed at learning long-term information. The central unit of this 3D convolution kernel is initialized as an identity matrix, and the remaining parameters are initialized to zero. The identity 3D convolution layer ensures that, before training, there is no alteration to the video representation, thereby maintaining consistency with the original computational process.

We apply our devised extending approach to Stable Video Diffusion [Blattmann *et al.*, 2023], which is a popular model within open-source communities for video synthesis. The comparative architectures, both pre and post-extension, are illustrated in Figure 1. Because of the fundamental similarities that underpin the construction of temporal blocks within video synthesis models, our extending approach can also be applied to various video synthesis models.

### 3.3 Post-Tuning

After extending the temporal blocks in the video synthesis models, we enhance the model’s abilities to generate extended videos via post-tuning. To circumvent potential copyright concerns with video content, we employed a publicly available dataset OpenSoraPlan<sup>2</sup>, which comprises 40,258 videos. These videos were sourced from copyright-free platforms. The videos in this dataset maintain a resolution of  $512 \times 512$ . ExVideo expands its capacity to 128 frames. Over such extended sequences, full training is deemed impractical because of the substantial computational requirements. Instead, we employed several engineering optimizations aimed at optimizing GPU memory usage. These optimizations are crucial for managing the increased computational load and facilitating efficient training within limited hardware resources:

- **Parameter freezing:** All parameters except the temporal blocks are frozen.
- **Mixed precision training:** We deploy a mixed precision training program by converting some parameters to 16-bit floating-point format.
- **Gradient checkpointing** [Feng and Huang, 2021]: Gradient checkpointing is enabled in the model. By storing intermediate states during forward passes and recomputing gradients on-demand during the backward pass, this technique effectively decreases memory usage.
- **Flash Attention** [Dao, 2023]: We integrate Flash Attention to enhance the computational efficiency of attention mechanisms.
- **Shard optimizer states and gradients:** We leverage DeepSpeed [Rasley *et al.*, 2020], a library optimized for distributed training, to enable shard optimizer states and gradients across multiple GPUs.

The loss function and the noise scheduler are consistent with the original model. The learning rate is  $1 \times 10^{-5}$  and the batch size on each GPU is 1. The training was conducted using only 8 NVIDIA A100 GPUs over one week. In order to ensure the stability of the training process, exponential moving averages were employed for the update of weights.

## 4 Experiments

By integrating extended temporal modules into the original Stable Video Diffusion (SVD) model and performing post-tuning, we have developed the Extended Stable Video Diffusion (ExSVD) model. This enhanced model is capable of generating coherent videos with lengths of up to 128 frames. To validate its capabilities, we have conducted comprehensive experiments comprising three components. First, we perform a comparative analysis between our ExSVD and the original SVD model to elucidate the enhancements achieved through post-tuning. Second, we evaluate the performance of ExSVD in comparison to other publicly accessible models. Finally, we present illustrative examples to provide a tangible understanding of the model’s performance.

<sup>2</sup><https://hf.co/datasets/LanguageBind/Open-Sora-Plan-v1.0.0>

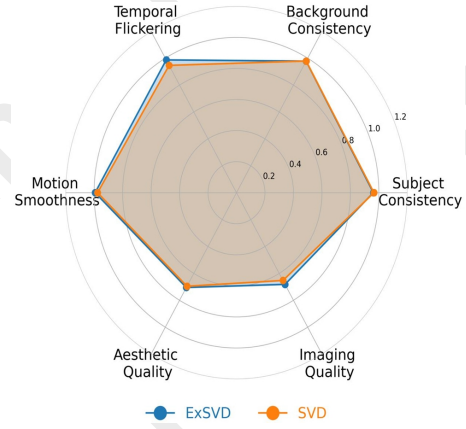


Figure 2: Automatic metrics computed based on the videos generated by ExSVD and the original SVD model. When the model is extended to a larger time scale, our approach effectively preserves the original capabilities of the model.

### 4.1 Evaluation on Extended Model Performance

To assess the performance of ExVideo, we present the results of our ExSVD model and compare them with those of the original SVD model across two primary dimensions: automatic metrics and human evaluation.

**Parameter Settings:** The comparative experiments are conducted based on VBench [Huang *et al.*, 2024], which is a comprehensive suite of tools designed to automatically assess the quality of generated videos. Following VBench’s prompt sampling methodology, we assigned 5 random seeds to each prompt for the text-to-video generation process. Initially, we employed the Hunyuan-DiT [Li *et al.*, 2024b] text-to-image model to generate 5 images for each prompt. Subsequently, for each image, both our ExSVD and the SVD models were used to generate corresponding videos, resulting in a total of 4720 ( $944 \times 5$ ) videos for each model. We used the DDIM [Song *et al.*, 2020] sampler with 50 steps for sampling.

**Evaluation Metrics:** VBench assesses video performance from two broad perspectives: video quality and video-condition consistency. Video quality focuses on the perceptual quality of the synthesized video, including temporal quality and frame-wise quality. Video-condition consistency focuses on whether the synthesized video aligns with the user-provided guiding condition (text prompt) and includes metrics from semantic and style dimensions. It is important to note that the dynamic degrees in the original implementation of VBench are sensitive to frames per second (FPS) and the total number of frames. Our findings indicate that this metric can significantly impact the overall score, potentially resulting in biased comparisons between different models. We will address this issue in future research, as there currently are no valid metrics available for evaluation.

**Quantitive Results:** Based on the generated videos from our ExSVD and the original SVD model, we have calculated all the automatic metrics in VBench. The visualized results of video quality are shown in Figure 2, and the detailed numerical results are presented in Table 1. From the perspective of video quality, most metrics of the ExSVD model are on par



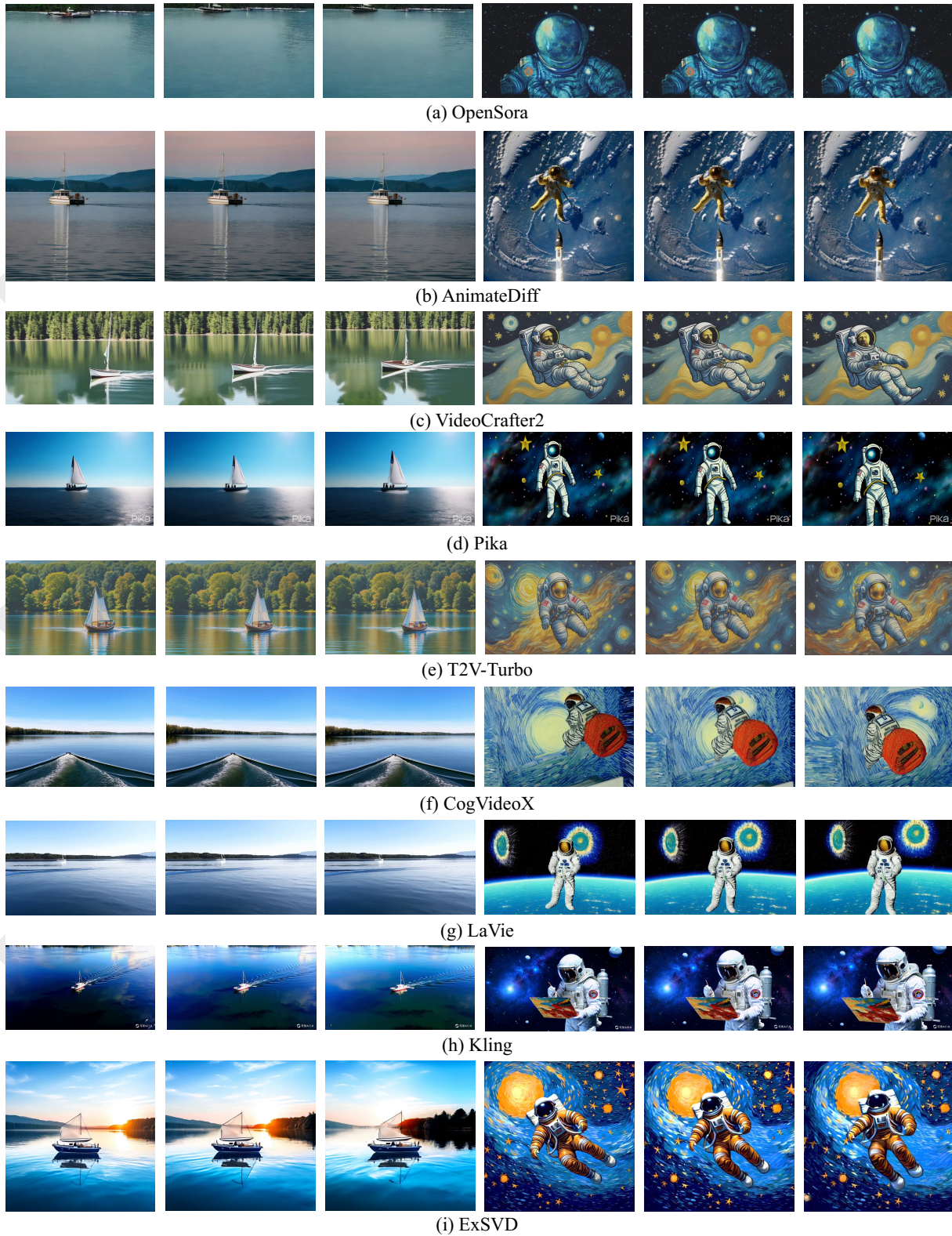


Figure 3: Visual comparisons of text-to-video results from several existing video synthesis models and our Extended model. The prompts are “a boat sailing smoothly on a calm lake” and “an astronaut flying in space, Van Gogh style”. In our pipeline, the first frame is generated by Hunyuan-DiT, and ExSVD generates the video according to the first frame. We highly recommend readers watch the videos on our project page: <https://ecnu-cilab.github.io/ExVideoProjectPage/>.

Model	Total Score	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Aesthetic Quality	Imaging Quality	Object Class
OpenSora [Zheng <i>et al.</i> , 2024]	79.23%	94.45%	97.90%	99.47%	98.20%	56.18%	60.94%	83.37%
AnimateDiff [Guo <i>et al.</i> , 2023]	80.27%	95.30%	97.68%	98.75%	97.76%	67.16%	70.10%	90.90%
VideoCrafter2 [Chen <i>et al.</i> , 2024a]	80.44%	96.85%	98.22%	98.41%	97.73%	63.13%	67.22%	92.55%
Pika [Pika, 2024]	80.69%	96.94%	97.36%	<b>99.74%</b>	<b>99.50%</b>	62.04%	61.87%	88.72%
T2V-Turbo [Li <i>et al.</i> , 2024a]	81.01%	96.28%	97.02%	97.48%	97.34%	63.04%	<b>72.49%</b>	93.96%
CogVideoX [Yang <i>et al.</i> , 2024]	81.61%	96.23%	96.52%	98.66%	96.92%	61.98%	62.90%	85.23%
LaVie [Wang <i>et al.</i> , 2023a]	81.75%	97.90%	<b>98.45%</b>	98.76%	98.42%	67.62%	70.39%	<b>97.52%</b>
Kling [Team, 2024]	81.85%	<b>98.33%</b>	97.60%	99.30%	99.40%	61.21%	65.62%	87.24%
SVD [Blattmann <i>et al.</i> , 2023]	81.73%	96.52%	97.89%	94.66%	97.58%	69.56%	65.25%	88.10%
ExSVD (ours)	<b>81.91%</b>	96.11%	97.79%	98.71%	99.31%	<b>70.47%</b>	68.16%	88.99%

Model	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
OpenSora [Zheng <i>et al.</i> , 2024]	58.41%	85.80%	87.49%	67.51%	42.47%	23.89%	24.55%	27.07%
AnimateDiff [Guo <i>et al.</i> , 2023]	36.88%	92.60%	87.47%	34.60%	50.19%	22.42%	<b>26.03%</b>	27.04%
VideoCrafter2 [Chen <i>et al.</i> , 2024a]	40.66%	95.00%	<b>92.92%</b>	35.86%	55.29%	<b>25.13%</b>	25.84%	28.23%
Pika [Pika, 2024]	43.08%	86.20%	90.57%	61.03%	49.83%	22.26%	24.22%	25.94%
T2V-Turbo [Li <i>et al.</i> , 2024a]	54.65%	95.20%	89.90%	38.67%	<b>55.58%</b>	24.42%	25.51%	28.16%
CogVideoX [Yang <i>et al.</i> , 2024]	62.11%	<b>99.40%</b>	82.81%	66.35%	53.20%	24.91%	25.38%	27.59%
LaVie [Wang <i>et al.</i> , 2023a]	64.88%	96.40%	91.65%	38.68%	49.59%	25.09%	25.24%	27.39%
Kling [Team, 2024]	<b>68.05%</b>	93.40%	89.90%	<b>73.03%</b>	50.86%	19.62%	24.17%	26.42%
SVD [Blattmann <i>et al.</i> , 2023]	61.55%	95.20%	82.81%	37.17%	43.68%	25.12%	25.71%	28.39%
ExSVD (ours)	63.38%	95.60%	84.04%	39.94%	43.14%	24.91%	25.96%	<b>28.55%</b>

Table 1: Comparison with other open-accessible models on VBench.

with those of the SVD model, indicating that ExSVD does not degrade the video quality of the original SVD model. In the temporal flickering dimension, the ExSVD model demonstrates superior performance. This enhancement is primarily attributed to the extended temporal block in the ExSVD model, which bolsters the model’s motion prediction capabilities. After post-tuning, the ExSVD model exhibits enhanced temporal consistency, resulting in fewer flickering phenomena and thus a higher temporal flickering score. From the perspective of video-condition consistency, the performances of the ExSVD and SVD models are generally consistent. In the spatial relationship and multiple objects metrics, ExSVD achieves higher scores. This indicates that ExSVD is capable of fully leveraging the text-to-image model to synthesize realistic videos based on the generated images. Overall, ExVideo enhances the total number of frames and duration of videos without compromising video quality.

**Human Evaluation:** In addition to the automatic metrics evaluation, we conducted a human preference experiment involving 30 participants to facilitate a comparative analysis between the SVD and our ExSVD model. In each evaluation session, we randomly selected two videos that both corresponded to the same prompt and presented them to the participants. Participants were instructed to choose from one of three options: “Left is better”, “Tie”, or “Right is better”, without disclosing the names of the models. Each participant evaluated up to 30 randomly selected video pairs. The results indicate that our ExSVD model outperformed the SVD model in terms of human preference, achieving a win rate of 48.35% compared to SVD’s win rate of 34.77%.

## 4.2 Comparison with Publicly Accessible Models

We further evaluated the performance of ExSVD in comparison to other publicly available models. To facilitate a comprehensive analysis that incorporates both text-to-video and image-to-video models, we designed a text-to-video pipeline

that integrates ExSVD with Hunyuan-DiT. This allows for a uniform assessment of the models across the text-to-video task. The parameter settings of this evaluation were consistent with those outlined in the previous subsection. For a thorough evaluation of our models, we selected top-performing video synthesis models for comparison. The results are summarized in Table 1, where the metrics of the baseline models are collected from the original VBench leaderboard. Due to the variability of the dynamic degree metric with respect to frame and FPS, we have opted not to include it in the table. In comparison to other models, our ExSVD outperforms the competition, achieving the highest overall score and excelling in the aesthetic quality, temporal style, and overall consistency metrics. Additionally, ExSVD exhibits competitive performance in the dimensions of temporal flickering, motion smoothness, and multiple objects.

## 4.3 Case Study

**Visual Comparison:** We present a series of video examples generated by ExSVD and other video synthesis models to facilitate an intuitive comparison of their performance. The illustrative results from these models are displayed in Figure 3. We highly recommend readers watch the videos on our project page. A noteworthy observation from this comparison is that the majority of existing video synthesis models tend to produce videos characterized by limited motion dynamics. In contrast, our extended model, which benefits from post-tuning applied over extended temporal durations, exhibits a markedly improved ability to generate videos with significant movement. In the second example, Kling, a competitive close-source model, demonstrates the capability to generate realistic videos, but it is unable to generate the astronaut in the style of Van Gogh. This disparity in performance highlights the advanced generative capabilities of our model.

**Generalization Abilities:** Although the model is trained at a fixed resolution and exclusively utilizes realistic video





Figure 4: Video examples in various resolutions. The first frame is generated by Stable Diffusion 3 with prompt “bonfire, on the stone”.

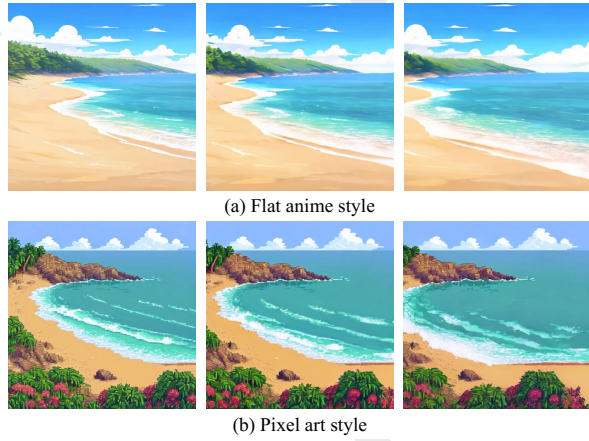


Figure 5: Examples in various styles generated by ExSVD, where the first frame is generated by Stable Diffusion 3. The prompt is “A beautiful coastal beach in spring, waves lapping on sand”, followed by the description of style.

datasets, the extended variant demonstrates remarkable capabilities, allowing for the generation of videos across a spectrum of resolutions and styles. To rigorously assess the performance of ExSVD, we conduct additional evaluations across various resolutions. Figure 4 illustrates several generated video examples that underscore the model’s capacity to generate videos in various resolutions and aspect ratios. This adaptability ensures that the generated videos maintain visual integrity regardless of the resolution parameters. Furthermore, Figure 5 presents an array of stylistic variations, further emphasizing the model’s versatility in accommodating diverse artistic expressions. These examples underscore the robustness and generalizability of ExSVD, offering flexibility in video generation across varying contexts.

**Visualization of Training Process:** We investigated the evolution of the model’s capabilities during the training process. Figure 6 presents the generated videos that exemplify the model’s performance at three distinct phases of training. It is difficult to present the dynamics using still images, thus we present the optical flow, computed by RAFT [Teed and Deng, 2020], to the right of each example for a clearer demon-

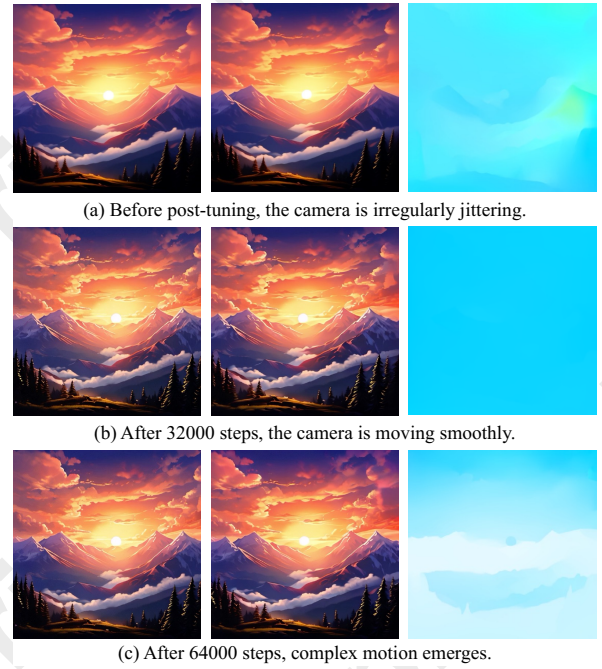


Figure 6: Examples in different phases. The prompt is “sunset, mountains, clouds”. We show the optical flow to visualize motion, with similar colors indicating similar movement directions.

stration of motion. Initially, before training, the extended model architecture was solely capable of guaranteeing the structural integrity of the video frames, which suffered from pronounced jittering artifacts. Progressing through the training, after 32,000 steps, the model began to produce videos displaying smooth camera movements. With continued training up to 64,000 steps, the model further advanced to create complex motions, such as clouds and mountains moving with nuanced, layered speed. The model effectively understands the depth and spatial relationships within the scene. This example intuitively illustrates the process of the model learning long-term information.

## 5 Conclusions and Future Work

In this paper, we delve into the enhancement of video diffusion models through post-tuning. Specifically, we propose a post-tuning approach called ExVideo, which can extend the duration of generated videos and release the potential of video synthesis models. Based on Stable Video Diffusion, our approach achieves a quintupling in the number of frames, while preserving the original generalization abilities. However, the enhanced model remains inherently constrained by the limitations of the base model. Looking ahead, we are committed to furthering our exploration of video synthesis models through post-tuning methodologies.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant Number 62202170 and Alibaba Group through Alibaba Innovation Research Program.

## References

- [Bar-Tal *et al.*, 2023] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *Proceedings of Machine Learning Research*, 202:1737–1752, 2023.
- [Blattmann *et al.*, 2023] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [Chen *et al.*, 2023a] Junsong Chen, YU Jincheng, GE Chongjian, Lewei Yao, Enze Xie, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *The Twelfth International Conference on Learning Representations*, 2023.
- [Chen *et al.*, 2023b] Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. Longlora: Efficient fine-tuning of long-context large language models. In *The Twelfth International Conference on Learning Representations*, 2023.
- [Chen *et al.*, 2024a] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024.
- [Chen *et al.*, 2024b] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *arXiv preprint arXiv:2402.19479*, 2024.
- [Dao, 2023] Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.
- [Dhariwal and Nichol, 2021] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.
- [Duan *et al.*, 2024] Zhongjie Duan, Chengyu Wang, Cen Chen, Weining Qian, and Jun Huang. Diffutoon: High-resolution editable toon shading via diffusion models. *arXiv preprint arXiv:2401.16224*, 2024.
- [Feng and Huang, 2021] Jianwei Feng and Dong Huang. Optimal gradient checkpoint search for arbitrary computation graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11433–11442, 2021.
- [Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.
- [Guo *et al.*, 2023] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2023.
- [He *et al.*, 2023] 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.
- [Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
- [Hu *et al.*, 2021] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.
- [Huang *et al.*, 2022] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *European Conference on Computer Vision*, pages 624–642. Springer, 2022.
- [Huang *et al.*, 2024] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024.
- [Jiménez, 2023] Álvaro Barbero Jiménez. Mixture of diffusers for scene composition and high resolution image generation. *arXiv preprint arXiv:2302.02412*, 2023.
- [Kodaira *et al.*, 2023] Akio Kodaira, Chenfeng Xu, Toshiki Hazama, Takanori Yoshimoto, Kohei Ohno, Shogo Mitsuhori, Soichi Sugano, Hanying Cho, Zhijian Liu, and Kurt Keutzer. Streamdiffusion: A pipeline-level solution for real-time interactive generation. *arXiv preprint arXiv:2312.12491*, 2023.
- [Li *et al.*, 2021] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: analysis, applications, and prospects. *IEEE transactions on neural networks and learning systems*, 33(12):6999–7019, 2021.
- [Li *et al.*, 2024a] Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhui Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024.
- [Li *et al.*, 2024b] Zhimin Li, Jianwei Zhang, Qin Lin, Jiangfeng Xiong, Yanxin Long, Xinchu Deng, Yingfang Zhang, Xingchao Liu, Minbin Huang, Zedong Xiao, et al.



- Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. *arXiv preprint arXiv:2405.08748*, 2024.
- [Liu et al., 2024] Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.
- [Pika, 2024] Pika. Pika, 2024.
- [Press et al., 2021] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.
- [Rasley et al., 2020] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506, 2020.
- [Rombach et al., 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [Ronneberger et al., 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.
- [Saharia et al., 2022] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.
- [Sohl-Dickstein et al., 2015] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
- [Song et al., 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [Su et al., 2024] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.
- [Team, 2024] Kuaishou AI Team. Kling, 2024.
- [Teed and Deng, 2020] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020.
- [Vaswani et al., 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [Wang et al., 2023a] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023.
- [Wang et al., 2023b] Yi Wang, Yinan He, Yizhuo Li, Kun-chang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.
- [Wu et al., 2024] Guangyang Wu, Xin Tao, Changlin Li, Wenyi Wang, Xiaohong Liu, and Qingqing Zheng. Perception-oriented video frame interpolation via asymmetric blending. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2753–2762, 2024.
- [Xiao et al., 2023] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2023.
- [Xiong et al., 2023] Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.
- [Yang et al., 2024] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [Ye et al., 2023] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- [Zhang et al., 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- [Zheng et al., 2024] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024.