Rethinking Video Generation: Overcoming the Limits of Pretrained Models

Ugur Demir*, Charlotte Lynn Chen†, Anthony Zhao‡, Nicholas X. Wang§, Michelle Wang¶, Derek Li§, Aggelos K. Katsaggelos*
*Northwestern University, Evanston, USA
ugurdemir2023@u.northwestern.edu, a-katsaggelos@northwestern.edu
†University of Washington, Seattle, USA
charc2006cv@gmail.com
‡University of Toronto, Toronto, Canada
anthonyz.zhao@mail.utoronto.ca
§BASIS Independent Silicon Valley, San Jose, USA
nicholas.x.wang@gmail.com, 2007derekli@gmail.com
¶Robert Louis Stevenson, Pebble Beach, USA
michellewang375@gmail.com

***Abstract*—Video generation has emerged as a promising research direction following the success of Stable Diffusion-based models in photo-realistic image synthesis. These models excel at generating fixed-size content but struggle with arbitrary-size video generation due to architectural constraints and training biases. Pretrained Stable Diffusion models have democratized access to powerful generative capabilities. However, they limit researchers to specific architectures and training configurations. These models primarily use U-Net architectures trained on square-format content, resulting in quality degradation when generating videos at non-standard dimensions. In this work, we revisit the video generation topics and propose new solutions to overcome the limits of pre-trained models, for arbitrary-size video generation. A novel approach was proposed that uses existing models without expensive retraining. Our method builds upon the publicly available AnimateDiff model, a video outpainting technique. This combination enables content generation at any desired dimension while maintaining visual quality. We also employ a frame interpolation technique that increases the temporal resolution without altering the context, allowing for smooth motion and enhanced temporal coherence. The approach eliminates the need for large-scale computing clusters typically required for training foundation models. It provides a practical solution for researchers working with pre-trained models who need flexible control over both spatial and temporal dimensions of video generation. This work presents preliminary results showing that our modular approach achieves better quality than direct arbitrary-size generation, demonstrating promise for future integration into end-to-end video generation models.**

***Index Terms*—stable diffusion, video generation, text-to-video generation.**

I. Introduction

Recent advances in deep learning have revolutionized content generation, particularly in the domains of image and video synthesis. Deep generative models have democratized content creation, enabling non-artists to produce professional-quality content while helping artists streamline their production workflows. The evolution of these technologies began with two-dimensional image generation, where Variational Autoencoders (VAEs) [1] and Generative Adversarial Networks (GANs) [2] demonstrated the capability to synthesize high-resolution, photorealistic images. While these early successes in image generation were promising, extending similar approaches to video generation proved challenging due to training instabilities, dataset limitations, and computational constraints.

The introduction of Stable Diffusion models [3] marked a significant breakthrough, addressing many of the training stability issues that plagued earlier approaches. Coupled with advances in hardware capabilities and the availability of large-scale datasets, these models enabled more sophisticated conditioning mechanisms, particularly in text-to-video generation. Users can now generate high-fidelity video content from textual descriptions, producing results with impressive texture quality and visual coherence [4], [5].

However, a significant limitation persists in current state-of-the-art models: they are predominantly trained on fixed-size, square-format content to sim-

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

plify the training process. This architectural choice introduces an inherent bias, where model performance degrades significantly when generating content at arbitrary dimensions. This limitation poses a practical challenge, as contemporary media consumption spans diverse aspect ratios and resolutions across different devices and platforms. While data augmentation during training can partially address arbitrary-size generation, the fundamental constraint of fixed-size input noise in Stable Diffusion models remains unresolved.

The scarcity of publicly available video generation models compounds this challenge, as training these models from scratch requires weeks of computation on large-scale clusters. Recent approaches like AnimateDiff have shown promise by fine-tuning existing image generation models with motion modules, but they inherit the same dimensional constraints of their base architectures.

To address these limitations, we propose a novel approach that enables arbitrary-size video generation without requiring expensive model retraining. Our method extends the capabilities of the publicly available AnimateDiff model through a video outpainting technique that enables content generation at any desired dimension while maintaining visual quality. We complement this with a frame interpolation method that enhances temporal resolution and motion smoothness without altering the video context. Our approach eliminates the need for large-scale computing clusters while maintaining compatibility with existing pre-trained models. This solution offers flexible control over both spatial and temporal dimensions while preserving visual quality across different aspect ratios and resolutions.

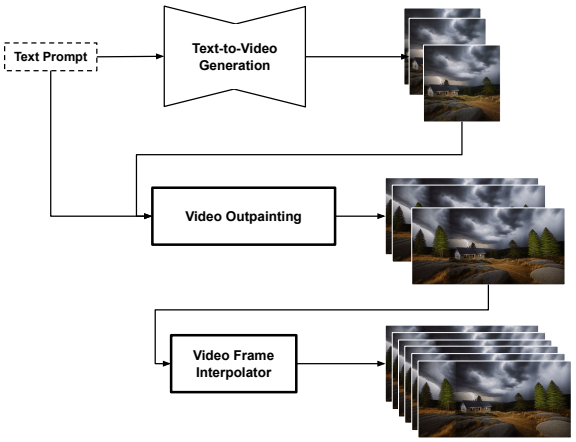

Fig. 1. Overview of the proposed arbitrary-size video generation pipeline. The system first generates a base video at fixed training dimensions using text-guided AnimateDiff [5], then extends spatial dimensions through video outpainting while preserving context via text conditioning, and finally increases temporal resolution through frame interpolation.

II. Related Work

A. Stable Diffusion

Diffusion models have emerged as a powerful class of generative models in computer vision, offering a novel approach to high-fidelity data synthesis. These models are grounded in the theoretical framework of Markov chains and stochastic processes. The fundamental principle underlying diffusion models is the gradual application of Gaussian noise to data points, followed by learning an iterative denoising process to reverse this diffusion [3].

Stable Diffusion models operate in a latent space to efficiently process high-dimensional data. Given an image $x \in \mathbb{R}^{H\times W\times 3}$ in RGB space, the encoder E maps x to a latent representation $z = E(x)$, where $z \in \mathbb{R}^{h\times w\times ch}$. The decoder D reconstructs the image from this latent representation, resulting in $\tilde{x} = D(z)$. This latent space approach significantly reduces the computational complexity compared to pixel-space diffusion models.

The core of Stable Diffusion is a conditional diffusion model $\epsilon_\theta : \mathcal{Z} \times \mathcal{C} \to \mathcal{Z}$ trained to predict a less noisy version of the latent representation z, conditioned on some variable $c \in \mathbb{R}^D$. The conditioning variable can represent various forms of guidance, such as text embeddings, scribbles, or edge maps. Starting with an initial noisy latent variable $z_T \sim \mathcal{N}(0, \mathbf{I})$, the reverse diffusion process iteratively denoises z_T over T steps to generate a realistic image latent z_0 that conforms to the conditioning c:

$$z_{t-1} = \epsilon_\theta(z_t, c) \quad \text{for } t = T, T-1, \ldots, 1, \qquad (1)$$

where ϵ_θ represents the denoising network, typically implemented as a modified U-Net architecture [6]. Unlike standard diffusion models that operate directly in pixel space, Stable Diffusion's latent approach reduces memory requirements and computational costs while maintaining high fidelity in the generated images. This efficiency has made it particularly suitable for practical applications and further development of specialized generation tasks.

B. Arbitrary Size Content Generation

The widespread adoption of Stable Diffusion models has highlighted the challenge of generating high-quality content at arbitrary dimensions without retraining. This limitation stems from the models' train-

ing process, which typically uses fixed-size images to optimize computational efficiency and stability. Several approaches have emerged to address this constraint while leveraging pre-trained models.

ElasticDiffusion [7] introduced a novel two-stage denoising process that enables arbitrary-size generation while maintaining the quality advantages of fixed-size training. In the first stage, the approach decomposes the input noise into patches matching the model's training dimensions. This decomposition leverages the model's optimal performance at its native resolution, as the denoising process operates on each patch independently. These individually denoised patches are then reconstructed to preserve the desired output dimensions. The second stage addresses global coherence through a complementary denoising process. The input noise is downscaled to the model's native dimensions, with additional noise padding applied for non-square aspects. After denoising, the padding is removed to obtain the final global structure. The method combines the locally detailed patches from the first stage with the globally coherent structure from the second stage to produce the final output.

However, extending these arbitrary-size generation techniques to video remains challenging due to the additional complexity of maintaining temporal consistency. Our work builds upon these foundations while specifically addressing the temporal aspects of arbitrary-size video generation.

C. Outpainting

Outpainting extends visual content beyond its original boundaries, presenting a fundamentally ill-posed problem where the model must synthesize plausible content for unobserved regions. While image outpainting has seen significant progress with Stable Diffusion-based approaches, video outpainting introduces additional challenges of maintaining temporal consistency alongside spatial coherence.

Early approaches to video outpainting applied image-based methods to individual frames independently. However, this frame-by-frame processing often resulted in temporal inconsistencies and flickering artifacts, as the ill-posed nature of outpainting led to divergent solutions across consecutive frames. Recent works have addressed these limitations by explicitly modeling temporal relationships [8], [9].

These advances in video outpainting provide essential foundations for arbitrary-size video generation, though challenges remain in scaling to significantly larger spatial dimensions while maintaining temporal consistency. Our work builds upon these developments, particularly in handling the interplay between spatial extension and temporal coherence.

III. Method

Our approach addresses the challenge of arbitrary-sized video generation by extending the capabilities of existing Stable Diffusion-based models. While current text-to-video generation models like AnimateDiff demonstrate impressive results, they are constrained to fixed dimensions due to their U-Net architecture and training configurations. We propose a three-stage pipeline that overcomes these limitations without requiring model retraining. Figure 1 shows the overall architecture.

A. Base Video Generation

The initial stage employs AnimateDiff, which uses a U-Net architecture for denoising. Given a text prompt p, it is first encoded into embedding $c \in \mathbb{R}^D$ using a pre-trained text encoder. A random noise tensor $z_T \in \mathbb{R}^{h \times w \times t \times ch}$ is sampled from a Gaussian distribution, where h, w, and t represent height, width, and temporal dimensions respectively, typically fixed at 512×512×16 during training. The denoising network iteratively processes this noise with the condition embedding:

$$z_{t-1} = \epsilon_\theta(z_t, c) \quad \text{for } t = T, T-1, \ldots, 1 \tag{2}$$

B. Spatial Dimension Extension

The second stage employs a video-aware outpainting model to extend the spatial dimensions of the generated video. Given the base video $V_b \in \mathbb{R}^{512 \times 512 \times 16 \times 3}$, we extend it to the target dimensions $V_t \in \mathbb{R}^{H \times W \times 16 \times 3}$, where H and W are arbitrary height and width dimensions. The outpainting process maintains temporal consistency by considering the entire video sequence rather than processing frames independently:

$$V_t = O(V_b, H, W) \tag{3}$$

where O represents the outpainting operation. This approach preserves both spatial coherence across the extended regions and temporal consistency across frames.

C. Temporal Extension

The final stage increases temporal resolution through video frame interpolation [10]. Given the spatially extended video V_t with 16 frames, we generate intermediate frames to reach the desired temporal length L. The interpolation process can be expressed as:

$$V_f = I(V_t, L) \tag{4}$$

Fig. 2. Visual comparison of video generation results. First rows: Base results from AnimateDiff [5] model at fixed training size. Second rows: AnimateDiff is forced to generate at an arbitrary size showing context deviation. Our method achieves the desired dimensions while preserving the original video context and quality.

where I represents the frame interpolation operation and $V_f \in \mathbb{R}^{H \times W \times L \times 3}$ is the final video.

D. Pipeline

This three-stage pipeline offers several practical advantages. By decomposing the generation process, we reduce the computational complexity compared to generating high-resolution, long videos directly. The approach utilizes existing pre-trained models for each stage, eliminating the need for additional training or modifications to the original AnimateDiff architecture. This modularity allows for independent improvements in each stage and makes the solution more accessible to researchers working with limited computational resources.

The temporal consistency is maintained throughout the pipeline; the base generation provides initial motion coherence, the video-aware outpainting preserves this coherence while extending spatial dimensions, and the frame interpolation enhances temporal resolution while respecting the existing motion patterns. This hierarchical approach ensures that the final output maintains both spatial and temporal consistency at arbitrary dimensions.

IV. Discussion

Our proposed approach to arbitrary-size video generation demonstrates several key advantages while also revealing important considerations for future research. The three-stage pipeline successfully addresses the dimensional constraints of pre-trained models without requiring expensive retraining processes, making it particularly valuable for researchers and practitioners with limited computational resources.

The decomposition of the generation process into base generation, spatial extension, and temporal interpolation provides flexibility in handling different aspect ratios and temporal requirements. Our experiments show that this approach maintains visual quality across a wide range of dimensions, though we observe some limitations in extreme cases. When the target dimensions significantly exceed the base model's training size, we notice a gradual degradation in fine detail preservation, particularly in regions far from the original boundaries.

One notable advantage of our method is its modularity. Each stage can be independently improved or replaced as better models become available. For instance, the base generation stage currently uses AnimateDiff, but it could be replaced with any future text-to-video model that provides higher quality or better control. Similarly, advances in video outpainting or frame interpolation techniques can be readily incorporated into our pipeline without requiring modifications to other stages.

Figure 2 demonstrates the visual comparison between baseline video generation and our proposed method. The baseline AnimateDiff model produces high-quality results at its native training dimensions of 512x512 with 16 frames. However, when attempting to generate videos at 1024x512x28 or 512x1024x28 dimensions, AnimateDiff significantly alters the video context, producing either disproportionate subjects or degraded textures. Our method successfully extends the video dimensions while preserving both the original context and texture quality. The results show that our approach maintains visual consistency with the base generated video while achieving the desired arbitrary dimensions, demonstrating effective spatial and temporal extension without compromising content fidelity.

V. Conclusion

In this paper, we presented a novel approach to arbitrary-size video generation that addresses a significant limitation of current Stable Diffusion-based models. Our method leverages existing pre-trained models through a three-stage pipeline, eliminating the need for expensive retraining while enabling flexible control over both spatial and temporal dimensions. The preliminary results demonstrate that our approach successfully generates videos at arbitrary dimensions while preserving visual quality and temporal consistency. Our method's modular design allows for independent improvements in each stage as better models become available, making it adaptable to future advances in video generation technology while maintaining efficient resource utilization. In future work, we plan to integrate these modular components directly into the base video generator architecture, working toward an end-to-end solution for arbitrary-size video generation.

References

[1] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.

[4] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach, "Stable video diffusion: Scaling latent video diffusion models to large datasets," 2023. [Online]. Available: https://arxiv.org/abs/2311.15127

[5] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," 2023.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[7] M. Haji-Ali, G. Balakrishnan, and V. Ordonez, "Elasticdiffusion: Training-free arbitrary size image generation," 2023.

[8] H. W. J. Y. W. Z. Q. T. H. W. S. M. Q. C. W. L. Qihua Chen, Yue Ma, "Follow-your-canvas: Higher-resolution video outpainting with extensive content generation," *arXiv preprint arXiv:XXXX.XXX0*, 2024.

[9] F.-Y. Wang, X. Wu, Z. Huang, X. Shi, D. Shen, G. Song, Y. Liu, and H. Li, "Be-your-outpainter: Mastering video outpainting through input-specific adaptation," 2024. [Online]. Available: https://arxiv.org/abs/2403.13745

[10] F. Reda, J. Kontkanen, E. Tabellion, D. Sun, C. Pantofaru, and B. Curless, "Film: Frame interpolation for large motion," in *European Conference on Computer Vision (ECCV)*, 2022.