FEAT: Full-Dimensional Efficient Attention Transformer for Medical Video Generation

Huihan Wang[1], Zhiwen Yang[1], Hui Zhang[2], Dan Zhao[3], Bingzheng Wei[4], and Yan Xu[1](✉)

[1] School of Biological Science and Medical Engineering, State Key Laboratory of Software Development Environment, Key Laboratory of Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Center for Biomedical Engineering, Beihang University, Beijing 100191, China
xuyan04@gmail.com
[2] Department of Biomedical Engineering, Tsinghua University, Beijing 100084, China
[3] Department of Gynecology Oncology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China
[4] ByteDance Inc., Beijing 100098, China

Abstract. Synthesizing high-quality medical videos remains a significant challenge due to the need for modeling both spatial consistency and temporal dynamics. Existing Transformer-based approaches face critical limitations, including insufficient channel interactions, high computational complexity from self-attention, and coarse denoising guidance from timestep embeddings when handling varying noise levels. In this work, we propose FEAT, a full-dimensional efficient attention Transformer, which addresses these issues through three key innovations: (1) a unified paradigm with sequential spatial-temporal-channel attention mechanisms to capture global dependencies across all dimensions, (2) a linear-complexity design for attention mechanisms in each dimension, utilizing weighted key-value attention and global channel attention, and (3) a residual value guidance module that provides fine-grained pixel-level guidance to adapt to different noise levels. We evaluate FEAT on standard benchmarks and downstream tasks, demonstrating that FEAT-S, with only 23% of the parameters of the state-of-the-art model Endora, achieves comparable or even superior performance. Furthermore, FEAT-L surpasses all comparison methods across multiple datasets, showcasing both superior effectiveness and scalability. Code is available at here.

Keywords: Video Generation · Medical Video · Efficient Transformer.

1 Introduction

Recent advancements in diffusion models have revolutionized artificial intelligence-generated content (AIGC) in medical imaging, enabling transformative applications in image synthesis [1], cross-modal translation [2], and image reconstruction

Equal contribution – H. Wang and Z. Yang.

[3]. While these models demonstrate remarkable capabilities in generating static medical images with spatial information, synthesizing high-fidelity dynamic medical videos—which require modeling additional temporal dynamics and consistency—remains a significant challenge. To this end, researchers have explored various approaches to encoding spatial-temporal dynamics [4,5,6,7], including pseudo-3D convolution [4], serial 2D+1D (spatial + temporal) convolutions [7], and spatial-temporal self-attention [6,5]. Given the ability of self-attention to capture long-range dependencies and the scalability of Transformers, most recent studies have largely embraced the Transformer architecture, employing cascading spatial and temporal self-attention mechanisms [6].

However, the current Transformer incorporating both spatial and temporal self-attention still faces three critical limitations: **(1) Inadequate Channel-Wise Interaction.** Despite their sophisticated handling of spatial and temporal dimensions, existing architectures neglect building channel dependencies crucial for modeling feature compositions. Additionally, the impressive generation performance of diffusion models relies heavily on the denoising process while channel attention [8] has been widely proven to be effective for denoising. Omitting building interactions over such an important dimension hinders the model performance. **(2) Prohibitive Computational Complexity.** The self-attention mechanisms used to model both spatial and temporal dependencies suffer from quadratic computational complexity, which severely limits their practical applicability in medical videos with high resolution and many frames. **(3) Coarse Denoising Guidance.** In diffusion models, the model needs to adapt to inputs affected by different noise levels across various timesteps. Existing methods rely on timestep embeddings as global-level guidance, using adaptive layer normalization (adaLN) [9] to adapt to specific noise levels. However, this approach is too coarse and fails to account for dynamic interactions between noise patterns and video content. While recent work [6] has utilized attention maps from DINO [10] to account for content information for finer-grained guidance, this method introduces additional substantial computational overhead during training. Therefore, existing methods have drawbacks in achieving efficient and effective medical video generation.

To address the aforementioned challenges, we present FEAT, a full-dimension efficient attention Transformer for medical video generation through three key innovations: (1) Full-Dimensional Dependency Modeling. FEAT introduces a unified paradigm with sequential spatial-temporal-channel attention, establishing global dependencies across all dimensions and enabling holistic feature modeling of medical videos. (2) Linear Complexity Design. FEAT replaces conventional self-attention with two computationally efficient components: (a) weighted key-value (WKV) attention [11,12,13] inspired by RWKV [11] for modeling spatial and temporal dependencies, and (b) global channel attention [8] for modeling channel dependencies. Both components achieve global dependencies within their respective dimensions while maintaining linear computational complexity [14]. (3) Residual Value Guidance. FEAT introduces a novel residual value guidance module (ResVGM) that leverages input embeddings—encoding both video con-

tent and specific noise patterns—as fine-grained pixel-level guidance to adapt the model for processing input of different timesteps. The ResVGM is parameter-efficient with negligible computational overhead while significantly improving generation performance. With these three innovations, FEAT achieves both efficient and effective medical video generation. Experiments show that a small version of FEAT (denoted as FEAT-S), with only 23% of the parameters of the state-of-the-art model Endora [6], delivers comparable or even superior performance. Furthermore, the larger version, FEAT-L, outperforms all comparison methods across different datasets.

Our contributions are three-fold:

- We propose FEAT, a novel full-dimensional efficient attention Transformer for medical video generation. FEAT establishes global dependencies across all dimensions, including spatial, temporal, and channel, thereby enhancing the model's ability to capture holistic relationships in medical videos.
- We replace the original self-attention mechanism, which suffers from quadratic computational complexity, with attention mechanisms that establish global dependencies with linear complexity, thereby enhancing model efficiency.
- We propose a novel residual value guidance module (ResVGM) that leverages input embeddings with both video content and specific noise patterns to provide fine-grained pixel-level guidance. This allows FEAT to effectively adapt to different timesteps with minimal computational overhead, significantly improving generation performance.

2 Method

We first introduce the preliminaries of the diffusion model for video generation in Section. 2.1. In Section 2.2, we then present the details of the proposed full-dimensional efficient attention Transformer (FEAT), including its overall architecture and the specific efficient attention mechanism tailored to each dimension. Finally, in Section 2.3, we introduce the novel residual value guidance module (ResVGM), which provides fine-grained pixel-level guidance for adapting to different denoising timesteps.

2.1 Preliminaries

Diffusion probabilistic models have emerged as a groundbreaking paradigm in generative modeling, demonstrating remarkable potential for image and video synthesis. These models operate by learning to transform random noise sampled from a standard normal distribution $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ into high-fidelity data samples through an iterative denoising procedure. The forward diffusion process gradually corrupts input data x_0 by adding Gaussian noise across T timesteps. This is defined by the transition $q(\mathbf{x}_t|\mathbf{x}_{t-1})$, with the marginal distribution at timestep t expressed as: $q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\alpha_t\mathbf{x}_0, \sigma_t^2\mathbf{I})$, where the coefficients of α_t and σ_t are designed such that x_T convergence to $\mathcal{N}(\mathbf{0}, \mathbf{I})$ as $t \to T$ [15]. In the

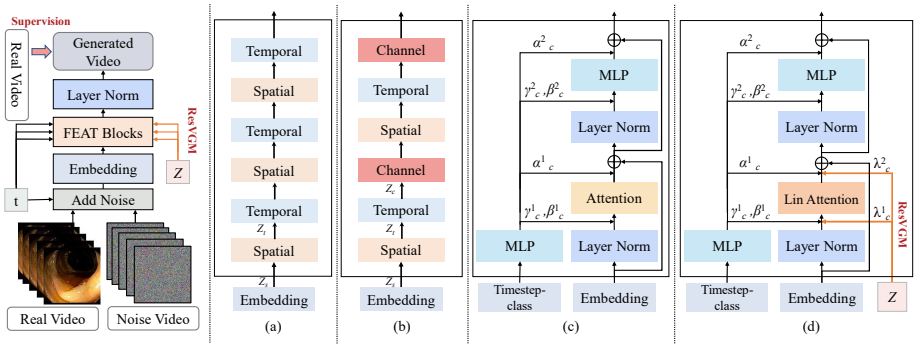

Fig. 1. The pipeline of FEAT for medical video generation. (a) Architecture of conventional models using cascaded spatial-temporal Transformer blocks. (b) Architecture of FEAT, which incorporates cascaded spatial-temporal-channel Transformer blocks. (c) Details of the conventional Transformer block, featuring quadratic computational complexity for self-attention and global timestep guidance. (d) Details of the Transformer block in FEAT, utilizing attention with linear computational complexity and guidance from both global timestep and pixel-level residual value Z.

reverse diffusion process, a noise prediction network $\epsilon_\theta(\mathbf{x}_t, t)$ parameterizes the transition $p(\mathbf{x}_{t-1}|\mathbf{x}_t)$, iteratively denoising x_t to recover the data distribution. The training process involves optimizing the evidence lower bound (ELBO) optimization [15]:

$$\text{ELBO} = \mathbb{E}\left[\left\|\epsilon_\theta\left(\alpha_t \mathbf{x}_0 + \sigma_t \epsilon;\, t\right) - \epsilon\right\|_2^2\right], \tag{1}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and t follows a uniform sampling.

Since training diffusion models directly in high-resolution pixel space can be computationally expensive, we adopt the widely used approach of latent diffusion models [16,17], performing the diffusion process in an encoded latent space with the help of a pretrained autoencoder [16] for both encoding and decoding.

2.2 Full-Dimensional Efficient Attention Transformer

Existing Transformer architectures for medical video generation often face three major drawbacks: insufficient channel-wise interaction, excessive computational complexity due to self-attention, and coarse denoising guidance from the timestep. To overcome these challenges, we propose the full-dimensional efficient attention Transformer (FEAT), as illustrated in Figure. 1, which introduces three key innovations: (1) Unlike conventional architectures that primarily establish spatial-temporal dependencies (as shown in Figure. 1 (a)), FEAT builds global dependencies across all dimensions, including spatial, temporal, and channel dimensions (as shown in Figure. 1 (b)). (2) To mitigate the computational burden imposed by self-attention in traditional Transformer blocks (as seen in Figure. 1 (c)), FEAT leverages attention mechanisms that achieve global attention with linear computational complexity across all dimensions (as shown in Figure. 1

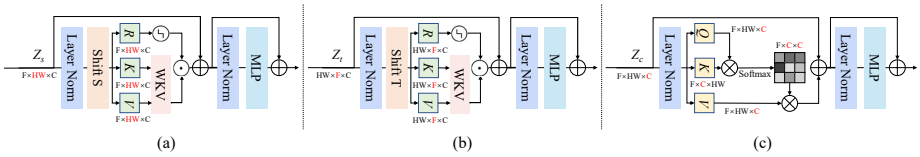

Fig. 2. Three distinct Transformer blocks in FEAT. (a) Spatial Transformer block with WKV attention [12]. (b) Temporal Transformer block with WKV attention [12]. (c) Channel Transformer block with global channel attention [8]. F, H, W, and C represent the frame number, height, width, and channel of the input feature, respectively.

(d)). (3) To address the coarse, global-level guidance that struggles to adapt to varying noise levels at different timesteps, FEAT introduces a residual value guidance module (ResVGM) for fine-grained, pixel-level denoising (as shown in Figure. 1 (d)). In the following subsection, we will describe the architecture of the Transformer blocks that establish spatial, temporal, and channel dependencies in detail. The details of the ResVGM are elaborated in Subsection 2.3.

To ensure efficient modeling across all dimensions, we design different Transformer blocks with efficient attention for each dimension. Given the exceptional performance of weighted key-value (WKV) attention [12,13] and global channel attention [8] in denoising, coupled with their ability to achieve global attention with linear computational complexity, we choose to apply them to denoising diffusion video generation. Specifically, **for the spatial and temporal Transformer blocks**, we adopt the WKV attention mechanism as described in [12], as illustrated in Figure 2 (a) and (b). To better accommodate the spatial and temporal dimensions, we modify the original token-shift mechanism from [12], which is designed to enhance locality. For the spatial Transformer block, we introduce 2D depth-wise convolution [18] (denoted 'Shift S') to strengthen locality in the spatial dimension. Similarly, for the temporal Transformer block, we apply 1D depth-wise convolution (denoted 'Shift T') to enhance locality in the temporal dimension. **For the channel Transformer block**, we directly employ the global channel attention mechanism proposed by [8], as depicted in Figure 2 (c). With these three Transformer blocks sequentially cascaded, FEAT can efficiently establish global dependencies across spatial, temporal, and channel dimensions, enabling holistic feature modeling for medical videos.

2.3 Residual Value Guidance Module

Most existing video diffusion models employ the timestep t as global guidance to adapt to specific noise levels in the denoising process. However, this method is relatively coarse and insufficient for content-dependent denoising. To overcome this limitation, we propose integrating the input embedding (denoted as Z) as an additional, fine-grained guidance. During the denoising process, the input embedding Z—obtained via convolution of the input (or the denoising output at the previous timestep)—encodes both the generated video content and the associated noise patterns. These components provide crucial guidance for achieving

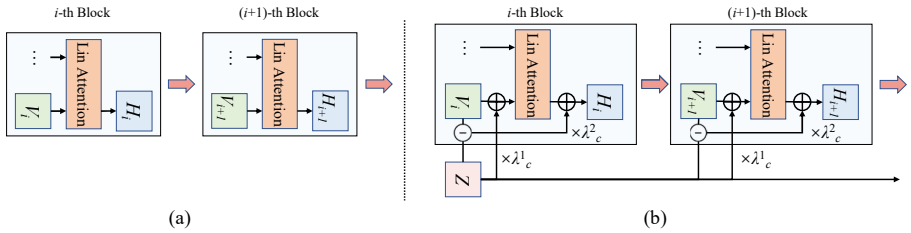

Fig. 3. The schematic diagram of the proposed ResVGM, where (a) represents the original framework, and (b) denotes the framework with ResVGM incorporated to different Transformer blocks. The frameworks primarily illustrate operations surrounding the attention mechanism in Transformer blocks, where ResVGM is integrated, while other modules are omitted for simplicity.

content-dependent denoising at specific noise levels. As illustrated in Figure 3, we incorporate the input embedding Z to all the Transformer blocks as fine-grained guidance. Specifically, for the i-th Transformer block, Z is added as a residual value [19] to interact with the input value V_i in the attention and the output hidden H_i as follows:

$$H_i = \text{LinAttention}(Q_i, K_i, V_i + \lambda_c^1 Z) + \lambda_c^2 (Z - V_i), \tag{2}$$

where LinAttention$(\cdot)$ denotes the two attention mechanisms—WKV attention and global channel attention—which both exhibit linear computational complexity. Q_i, K_i, and V_i denote the query, key, and value, respectively. Note that Q_i can be omitted in WKV attention. λ_c^1, $\lambda_c^2 \in \mathbb{R}^C$ are two learnable weighting parameters. This process ensures that feature extraction across all Transformer blocks in the model is gradually refined based on the input video content and noise level. The ResVGM introduces negligible additional parameters and computational overhead, while significantly improving performance.

3 Experiments

3.1 Experiment Settings

Datasets and Evaluation. Our experimental evaluation is conducted on two publicly available medical video datasets: Colonoscopic [20] and Kvasir-Capsule [21]. Adhering to standardized video processing protocols [6], we preprocess the data by uniformly extracting 16-frame sequences from continuous videos through fixed-interval sampling. All frames are resized to 128×128 pixel resolution during model training to ensure dimensional consistency. For quantitative assessment, we employ four established evaluation metrics: Fréchet Inception Distance (FID) [22], Inception Score (IS) [23], Fréchet Video Distance (FVD) [24], and its content-debiased variant CD-FVD [25]. Following the evaluation framework of StyleGAN-V [26], we compute FVD scores through statistical analysis of 2048

Table 1. Quantitative Comparisons on Medical Video Datasets.

Method	Colonoscopic [20]				Kvasir-Capsule [21]				Parameters(M)↓	FLOPs(G)↓
	FVD ↓	CD-FVD ↓	FID ↓	IS ↑	FVD ↓	CD-FVD ↓	FID ↓	IS ↑		
StyleGAN-V [26] (CVPR'22)	2110.7	1032.8	226.14	2.12	183.5	898.4	31.61	2.77	\	\
LVDM [17] (Arxiv'23)	1036.7	792.9	96.85	1.93	1027.8	615.4	200.90	1.46	\	\
MoStGAN-V [27] (CVPR'23)	468.5	592.0	53.17	3.37	82.8	168.3	17.34	2.53	\	\
Endora [6] (MICCAI'24)	460.7	545.3	13.41	3.90	72.3	152.3	10.61	2.54	673.7	465.8
FEAT-S(Ours)	415.4	444.0	13.34	3.96	72.2	138.2	9.97	2.65	158.0	118.7
FEAT-L(Ours)	351.1	397.0	12.31	4.01	59.2	116.2	8.65	2.70	629.0	472.1

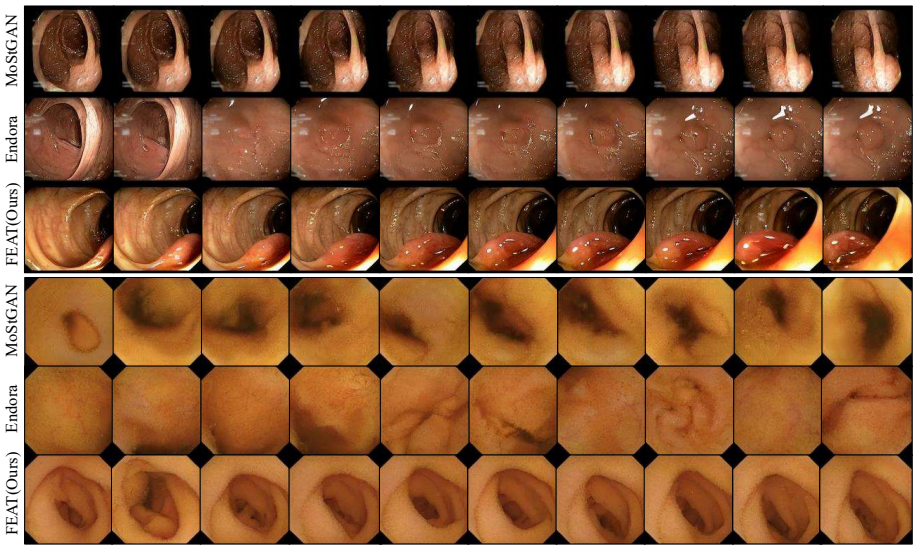

Fig. 4. Qualitative Comparison on Colonoscopic and Kvasir-Capsule Datasets.

video samples, with each sample maintaining the complete 16-frame temporal structure to preserve motion dynamics and temporal coherence.

Implementation Details. Our implementation employs the AdamW optimizer with a fixed learning rate of 1×10^{-4} across all architectural configurations. Data preprocessing incorporates basic horizontal flipping as the sole augmentation strategy to preserve feature authenticity. The model architecture integrates a pretrained variational autoencoder from the Stable Diffusion framework [28] as its foundational component, enhanced by 27 specialized neural modules organized in an interleaved configuration: 9 spatial processors for geometric feature extraction, 9 temporal analyzers for motion pattern modeling, and 9 channel operators for cross-dimensional feature interaction. Hidden dimensions are configured as d=512 for small (S) models and d=1024 for large (L) model variants to accommodate computational constraints. Following established GAN training protocols [9], we implement exponential moving average (EMA) stabilization with all final outputs generated from converged EMA parameters, ensuring training stability and output consistency.

Table 2. Semi-supervised Classification Result (F1 Score) on PolyDiag [29] .

Method	Colonoscopic [20]
Supervised-only	74.5
LVDM	76.2 (+1.7)
Endora	87.0 (+12.5)
FEAT-S(Ours)	89.9 (+15.4)
FEAT-L(Ours)	**91.3 (+16.8)**

Table 3. Ablation Studies of Proposed Components on Colonoscopic [20] Dataset.

WKV	Channel	ResVGM	FVD↓	FID↓
✗	✗	✗	990.0	23.45
✓	✗	✗	788.4	20.16
✓	✓	✗	583.6	16.98
✓	✓	✓	**415.4**	**13.34**

3.2 Comparison with State-of-the-arts

We conduct performance comparison by replicating several advanced video generation models designed for general scenarios on the medical video datasets, including StyleGAN-V [26], MoStGAN-V [27], LVDM [17], and Endora [6]. As shown in Table. 1, FEAT-S achieves comparable performance to Endora while requiring significantly fewer parameters and lower computational costs. Meanwhile, FEAT-L outperforms state-of-the-art methods. The visual qualitative comparison results in Figure. 4 demonstrate that FEAT can generate videos with higher quality and consistency.

3.3 Further Empirical Studies

In this section, we demonstrate the data augmentation effects of leveraging the videos generated by our FEAT for downstream tasks, and conduct rigorous ablation experiments on the proposed improvements.

Downstream Task. We explore the use of generated videos as unlabeled data for semi-supervised learning, specifically leveraging the FixMatch framework [30] on video-based disease diagnosis benchmarks, such as PolyDiag [29]. For this experiment, we randomly select 40 labeled videos ($n_l = 40$) from the PolyDiag training set and use 200 generated videos ($n_u = 200$) from Colonoscopic [20] as unlabeled data. The F1 scores for disease diagnosis, along with the performance improvements over the supervised-only baseline, are presented in Table. 2. The results clearly demonstrate that data generated by FEAT significantly boosts the performance of downstream tasks compared to both the supervised learning baseline and other video generation techniques, thereby confirming FEAT's effectiveness as a reliable video data augmenter for video-based analysis tasks.

Ablation Studies. Table. 3 presents an ablation study to evaluate the key components of the proposed FEAT-S model. We begin with a baseline that employs a simple spatial-temporal Transformer diffusion model, without incorporating any of the proposed strategies. Next, we incrementally add the three proposed design strategies: WKV attention, channel attention and ResVGM. The results clearly show that each strategy contributes to a progressive improvement in the model's performance, highlighting the essential role of these design choices in enhancing the effectiveness of the medical video generation model.

4 Conclusion

This paper introduces FEAT, a novel full-dimensional efficient attention Transformer that significantly advances medical video generation. FEAT addresses three key challenges—limited channel-wise interaction, prohibitive computational cost, and coarse denoising guidance—through three core innovations. First, a sequential spatial-temporal-channel attention paradigm enables holistic feature modeling across all dimensions. Second, a linear-complexity attention design makes it scale efficiently to high-resolution videos. Third, a lightweight residual-value guidance module adaptively refines denoising, optimizing generation performance at negligible extra computational cost. Experimental results demonstrate that FEAT outperforms existing methods in terms of both efficiency and effectiveness, marking a substantial step forward in the field of medical video generation. Future work will extend FEAT to additional imaging modalities and conduct more comprehensive evaluations.

Acknowledgments. This work is supported by the National Natural Science Foundation in China under Grant 62371016 and U23B2063, the Bejing Natural Science Foundation Haidian District Joint Fund in China under Grant L222032, the Fundamental Research Funds for the Central University of China from the State Key Laboratory of Software Development Environment in Beihang University in China, the 111 Proiect in China under Grant B13003, the SinoUnion Healthcare Inc. under the eHealth program, and the high performance computing (HPC) resources at Beihang University.

Disclosure of Interests. We have no conflicts of interest to disclose.

References

1. Dorjsembe, Z., Odonchimed, S., Xiao, F.: Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In: Medical imaging with deep learning (2022)
2. Wang, Z., Zhang, L., Wang, L., Zhang, Z.: Soft masked mamba diffusion model for ct to mri conversion. arXiv preprint arXiv:2406.15910 (2024)
3. Liu, J., Anirudh, R., Thiagarajan, J.J., He, S., Mohan, K.A., Kamilov, U.S., Kim, H.: Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In: Proceedings of the IEEE/CVF International conference on computer vision. pp. 10498–10508 (2023)
4. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. arXiv preprint arXiv:2209.14792 (2022)
5. Chen, H., Zhang, Y., Cun, X., Xia, M., Wang, X., Weng, C., Shan, Y.: Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7310–7320 (2024)
6. Li, C., Liu, H., Liu, Y., Feng, B.Y., Li, W., Liu, X., Chen, Z., Shao, J., Yuan, Y.: Endora: Video generation models as endoscopy simulators. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 230–240. Springer (2024)

7. Xing, J., Xia, M., Liu, Y., Zhang, Y., Zhang, Y., He, Y., Liu, H., Chen, H., Cun, X., Wang, X., et al.: Make-your-video: Customized video generation using textual and structural guidance. IEEE Transactions on Visualization and Computer Graphics (2024)
8. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5728–5739 (2022)
9. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
11. Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Grella, M., et al.: Rwkv: Reinventing rnns for the transformer era. arXiv preprint arXiv:2305.13048 (2023)
12. Duan, Y., Wang, W., Chen, Z., Zhu, X., Lu, L., Lu, T., Qiao, Y., Li, H., Dai, J., Wang, W.: Vision-rwkv: Efficient and scalable visual perception with rwkv-like architectures. arXiv preprint arXiv:2403.02308 (2024)
13. Yang, Z., Li, J., Zhang, H., Zhao, D., Wei, B., Xu, Y.: Restore-rwkv: Efficient and effective medical image restoration with rwkv. arXiv preprint arXiv:2407.11087 (2024)
14. Shen, Z., Zhang, M., Zhao, H., Yi, S., Li, H.: Efficient attention: Attention with linear complexities. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3531–3539 (2021)
15. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. Advances in neural information processing systems **34**, 21696–21707 (2021)
16. 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)
17. He, Y., Yang, T., Zhang, Y., Shan, Y., Chen, Q.: Latent video diffusion models for high-fidelity long video generation. arXiv preprint arXiv:2211.13221 (2022)
18. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017)
19. Zhou, Z., Wu, T., Jiang, Z., Lan, Z.: Value residual learning for alleviating attention concentration in transformers. arXiv preprint arXiv:2410.17897 (2024)
20. Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE transactions on medical imaging **35**(9), 2051–2063 (2016)
21. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)
22. Parmar, G., Zhang, R., Zhu, J.Y.: On buggy resizing libraries and surprising subtleties in fid calculation. arXiv preprint arXiv:2104.11222 **5**(14), 6 (2021)
23. Saito, M., Matsumoto, E., Saito, S.: Temporal generative adversarial nets with singular value clipping. In: Proceedings of the IEEE international conference on computer vision. pp. 2830–2839 (2017)

24. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018)
25. Ge, S., Mahapatra, A., Parmar, G., Zhu, J.Y., Huang, J.B.: On the content bias in fréchet video distance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7277–7288 (2024)
26. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3626–3636 (2022)
27. Shen, X., Li, X., Elhoseiny, M.: Mostgan-v: Video generation with temporal motion styles. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5652–5661 (2023)
28. 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)
29. Tian, Y., Pang, G., Liu, F., Liu, Y., Wang, C., Chen, Y., Verjans, J., Carneiro, G.: Contrastive transformer-based multiple instance learning for weakly supervised polyp frame detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 88–98. Springer (2022)
30. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)