



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

EndoGen: Conditional Autoregressive Endoscopic Video Generation

Xinyu Liu¹, Hengyu Liu¹, Cheng Wang¹, Tianming Liu², Yixuan Yuan^{1,✉}

¹ The Chinese University of Hong Kong, Hong Kong SAR

² University of Georgia, GA, USA
yxyuan@ee.cuhk.edu.hk

Abstract. Endoscopic video generation is crucial for advancing medical imaging and enhancing diagnostic capabilities. However, prior efforts in this field have either focused on static images, lacking the dynamic context required for practical applications, or have relied on unconditional generation that fails to provide meaningful references for clinicians. Therefore, in this paper, we propose the first conditional endoscopic video generation framework, namely EndoGen. Specifically, we build an autoregressive model with a tailored Spatiotemporal Grid-Frame Patterning (SGP) strategy. It reformulates the learning of generating multiple frames as a grid-based image generation pattern, which effectively capitalizes the inherent global dependency modeling capabilities of autoregressive architectures. Furthermore, we propose a Semantic-Aware Token Masking (SAT) mechanism, which enhances the model’s ability to produce rich and diverse content by selectively focusing on semantically meaningful regions during the generation process. Through extensive experiments, we demonstrate the effectiveness of our framework in generating high-quality, conditionally guided endoscopic content, and improves the performance of downstream task of polyp segmentation. Code released at <https://www.github.com/CUHK-AIM-Group/EndoGen>.

Keywords: Endoscopy · Autoregressive Models · Token Masking · Conditional Video Generation.

1 Introduction

Endoscopy video generation is a critical task with far-reaching implications for medical applications, including surgical training, diagnostic system development, and patient education [11, 13, 15, 25]. Realistic and controllable video synthesis can simulate rare pathological conditions, enable personalized surgical planning, and provide high-quality datasets for training AI models. However, existing generation methods primarily focus on static image synthesis [4, 18] or unconditional video generation [11]. Static images lack the temporal dynamics essential for simulating endoscopic procedures [19]. For unconditional video models [11], they produce arbitrary sequences that are not aligned with specific anatomical

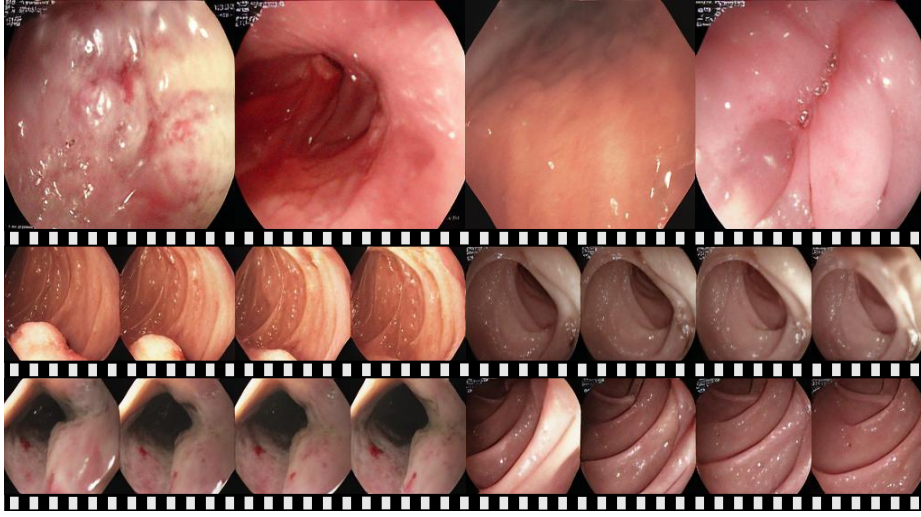


Fig. 1. Endoscopic frames and videos with different resolutions generated by EndoGen.

or pathological conditions when needed by doctors [29]. These limitations hinder their practical utility in scenarios requiring targeted outputs, such as generating videos of specific pathologies or tailoring simulations for surgical training. Thus, there is an urgent need for a conditional endoscopy video generation framework that can produce high-quality videos tailored to specific anatomical or procedural constraints.

Recent advances in autoregressive (AR) models [10, 23] have demonstrated superior conditional modeling capabilities compared to diffusion-based methods, particularly in tasks requiring long-range dependencies, such as text and image generation [12, 21, 22, 28]. With a condition token, AR models operate by predicting the next token based on all previously generated tokens, enabling them to capture complex hierarchical relationships and generate highly coherent outputs. However, despite their strengths, AR models are typically data-hungry [3, 23] and have been largely confined to static image generation. Extending these models to endoscopy video generation poses significant challenges, as naive approaches often result in temporal inconsistencies and fail to leverage the inherent long-range dependencies of video data [28]. This raises an important question: *Can we adapt the long-range conditional modeling capabilities of AR models to generate temporally coherent and contextually relevant endoscopic videos?*

To address this challenge, we initially construct a framework for conditional endoscopic video generation, named EndoGen. Specially, we develop a Spatiotemporal Grid-Frame Patterning (SGP) strategy to effectively train the AR model to learn spatial and temporal dependencies simultaneously. SGP redefines multi-frame generation as a synthesis task of a grid of interconnected images, which leverages the inherent capability of AR in modeling long-range relationships while preserving inter-frame continuity. This approach allows the gener-

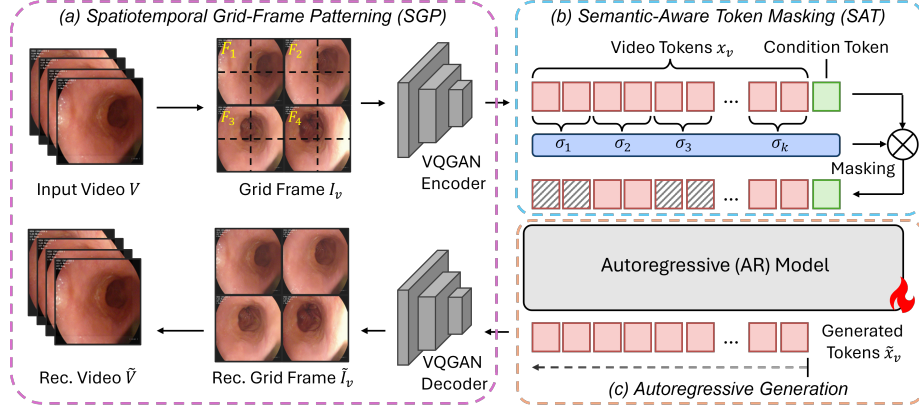


Fig. 2. Illustration of the EndoGen framework. During training, each input video undergoes (a) Spatiotemporal Grid-Frame Patterning (SGP), (b) Semantic-Aware Token Masking (SAT), and (c) Autoregressive Generation. During inference, video tokens are generated autoregressively based on the provided condition token and then reconstructed into video format.

ation of temporally consistent and detail preserved endoscopic sequences. Furthermore, to enhance the diversity and clinical relevance of the generated videos, we introduce a Semantic-Aware Token Masking (SAT) mechanism. SAT dynamically masks video tokens with less or redundant information, while preserving those with rich semantic content based on their intrinsic feature variance. This design encourages the model to focus on informative features that align closely with clinical objectives. With the proposed learning strategies, our framework is capable to generate highly realistic endoscopic videos across various conditions. We display generated frames and videos with different resolutions in Fig. 1.

We extensively evaluate our method on video generation and downstream task. Experimental results demonstrate that EndoGen generates temporally coherent and clinically relevant endoscopic videos, outperforming existing methods in terms of both visual fidelity and utility for downstream application. Our work not only advances the state of the art in medical video generation but also opens new avenues for leveraging AR models in dynamic medical imaging tasks.

2 Methodology

The overview of EndoGen is presented in Fig. 2. During training, it reformulates the input video as grid frames with SGP (Sec. 2.1) and generates video tokens. Then, the video tokens are adaptively masked with SAT (Sec. 2.2) to learn more diverse content. Specially, a conditional token is indexed from a set of learnable embeddings [21], and serves as the starting prefilling token. Starting from it, the model generates a sequence of video tokens autoregressively. Without conditional token, the model can only generate random class samples and fails to produce

desired class videos when needed by doctors. After concatenating the masked video and condition tokens, we feed them into the AR model to generate tokens autoregressively, and a standard cross-entropy loss [22] is utilized for supervision of the generated token. At inference time, only a condition token is provided to the AR model and the generated tokens are decoded and reconstructed into the original video format.

2.1 Spatiotemporal Grid-Frame Patterning (SGP)

To bridge the gap between text/image and video generation in autoregressive models, we propose SGP, an effective strategy to encode both spatial and temporal information into a unified representation, which is shown in Fig. 2(a). Traditional video generation approaches either process videos with 3D blocks [8] or with interleaved spatial and temporal modules [2, 16]. However, the 3D block-based methods suffer from high computational complexity and memory requirements during training [17], while the interleaved spatial-temporal methods introduce architectural complexity and could struggle in maintaining temporal consistency [27].

Different from them, our method maps the temporal sequence into a spatial representation, enabling simultaneous modeling of spatial and temporal dependencies via attention computation. Specifically, for each input video sequence V with frames $\{F_1, F_2, \dots, F_N\}$, we arrange them in a specific grid-based pattern I_v , and the I_v is fed into a VQGAN [5] encoder to obtain the latent feature $x_v = E(I_v)$. Specifically, SGP arranges video frames in a sequential, row-by-row format within a large image, which ensures the frames maintain temporal dynamics when processed by the AR model. Afterwards, x_v is processed with the proposed SAT (described in Sec. 2.2) and reconstructed with the AR model in an autoregressive manner. The reconstructed latent representation $\tilde{x}_v$ is subsequently processed through the VQGAN decoder, which generates a reconstructed grid frame pattern $\tilde{I}_v = D(\tilde{x}_v)$. Finally, the framework decomposes the $\tilde{I}_v$ back into individual frames and reorders them to form the output video sequence $\tilde{V}$, ensuring temporal coherence throughout the generation process. SGP efficiently compresses temporal information into spatial patterns while preserving frame-to-frame relationships, which ensures the consistency of generation results.

2.2 Semantic-Aware Token Masking (SAT)

To enhance the diversity and clinical relevance of the generated videos, we introduce a SAT mechanism, as shown in Fig. 2(b). SAT dynamically prioritizes tokens with rich semantic content based on their intrinsic feature variance during training, while masking those with less informative or redundant features. This selective masking operation ensures that the model focuses on capturing informative features that are more aligned with clinical objectives, such as lesion areas or surgical tools.

Specifically, we are given a tokenized video feature x_v of shape $(B, T \times L, C)$, where B is the batch size, T is the number of frames, L is the token length of a

single frame, and C is the feature dimension. Specially, we first split the feature with $(T \times L)/H$ segments, with each has a token length of H . For each segment, the variance across the channel dimension is computed, and a masking ratio is adaptively determined based on the variance:

$$\sigma_i^2 = \frac{1}{H} \sum_{h=1}^H (s_{i,h} - \mu_i)^2, \quad p_i = \text{Clamp} \left(\left(1 - \frac{\sigma_i^2}{\max(\sigma_i^2)} \right) \cdot p_{\max}, 0, p_{\max} \right), \quad (1)$$

where μ_i and σ_i^2 are the mean and variance values for the i -th segment s_i , and $p_{\max}$ is the maximum threshold for the masking ratio. During training, a binary mask M_i is applied to each segment based on the computed ratio, ensuring that only the most informative tokens are retained:

$$s'_i = s_i \odot M_i, \quad \text{where } M_i \sim \text{Bernoulli}(1 - p_i). \quad (2)$$

With SAT, the model is encouraged to generate videos that are not only temporally coherent but also semantic meaningful, addressing a critical limitation of existing video generation methods.

3 Experiments

3.1 Datasets and Implementation Details

We conduct experiments on two endoscopic video datasets. HyperKvasir [1] contains videos with 8 different pathological findings: {barretts, cancer, esophagitis, gastric-antral-vascular-ectasia, gastric-banding-perforated, polyps, ulcer, varices}. SurgVisdom [31] contains surgical videos on porcine model with 3 surgical tasks: {dissection, knot-tying, needle-driving}. The AR model is trained for 300 epoch using AdamW optimizer with learning rate 1e-4. H is set to 8. In the main comparison experiments, we use 16-frame video clips from the datasets with a specific sampling interval, and resize each frame to the 128×128 resolution for training. We also present results for videos with 64 frames or a spatial resolution of 256×256 in the supplementary material. We apply a frozen VQGAN pretrained on general domain data [5] to reduce training cost, and use the ImageNet pretrained class conditional image generation model [21] as the AR model weight initialization. We compare with diffusion based methods SimDA [26] and VDM [8], as well as the autoregressive method VideoGPT [28]. Per-class Fréchet Video Distance (FVD) [24], Content-Debiased Fréchet Video Distance (CD-FVD) [6], Fréchet Inception Distance (FID) [7], and Learned Perceptual Image Patch Similarity (LPIPS) [30] are used as the evaluation metrics. For all these metrics, lower values indicate better performance.

3.2 Video Generation Performance

Comparison with State-of-the-arts. As shown in Table 1, EndoGen achieves state-of-the-art performance across all eight pathological findings in the conditional generation on HyperKvasir, outperforming existing methods by significant

Table 1. Conditional video generation FVD results on HyperKvasir [1] with different pathological findings, where lower values are better. Bold denotes best performance.

Method	Bar.	Cancer	Eso.	Ecta.	Perf.	Polyps	Ulcer	Varices	Avg.
SimDA [26]	3479.1	5065.1	2041.4	3643.6	1641.7	3656.2	3688.0	3919.3	3391.8
VDM [8]	1758.8	4635.1	1366.9	2057.6	897.0	2348.3	2172.6	1766.5	2125.4
VidGPT [28]	1433.1	2965.7	955.7	1649.4	636.3	1705.1	1616.1	1427.7	1548.6
EndoGen	402.1	908.0	286.3	628.1	300.6	423.2	496.6	612.9	507.2

Table 2. Conditional video generation FVD results on SurgVisdom [31] with different surgical tasks, lower is better.

Method	Dis.	Knot.	Dri.	Avg.
SimDA [26]	3682.8	5889.2	3342.7	4304.9
VDM [8]	1948.2	2716.3	2365.4	2343.3
VidGPT [28]	3394.5	2397.5	2197.6	2663.2
EndoGen	1324.9	1606.5	1249.5	1393.6

Table 3. Results comparison on the HyperKvasir [1] dataset with different evaluation metrics, lower values are better.

Method	CD-FVD	FID	LPIPS
SimDA [26]	1319.9	288.4	0.565
VDM [8]	851.4	246.8	0.652
VidGPT [28]	980.6	235.8	0.563
EndoGen	765.3	76.56	0.528

margins. It is observed that diffusion-based approaches like VDM [8] could struggle with fine-grained anatomical consistency (e.g., 2172.6 FVD for ulcers), while our method shows a significantly reduced FVD value of 496.6. Compared to the autoregressive models like VidGPT [28], EndoGen demonstrates a better ability in generating complex pathologies such as varices, with 612.9 vs 1427.7 FVD. Notably, EndoGen shows particular strength in capturing subtle variations in Barrett’s esophagus, which is attributed to our SAT mechanism that prioritizes diagnostically relevant features. In the qualitative comparison in Fig. 3, EndoGen demonstrates more anatomically accurate and temporally coherent endoscopic videos under different conditions. Meanwhile, the generated videos show clearer textures and smoother transitions.

In Table 2, EndoGen also shows superior performance on the SurgVisdom dataset, achieving 40.6% lower FVD compared to the diffusion-based VDM. This demonstrates its robustness to diverse procedural dynamics. From Fig. 3, compared to other methods [8, 28] that show distorted or blurry content in the challenging task, EndoGen offers superior visual representation of tissues and equipment, meanwhile effectively capturing the characteristics of the corresponding surgical task. Furthermore, Table 3 reveals that EndoGen achieves

Table 4. Ablation of the components and maximum threshold for the masking ratio on the HyperKvasir dataset. Lower values denote better performance.

Metric	w/o SGP	w/o SAT	$p_{\max}=0.1$	$p_{\max}=0.2$	$p_{\max}=0.3$	$p_{\max}=0.4$
FVD	2617.5	562.0	533.8	514.8	507.2	521.2

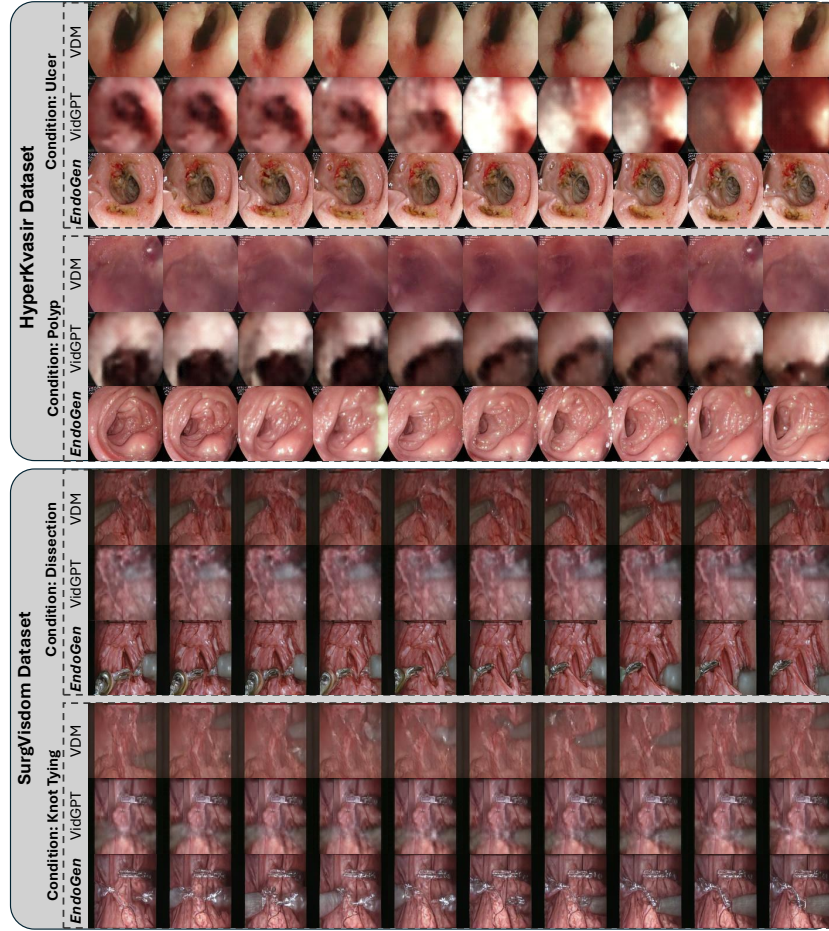


Fig. 3. Qualitative comparison on the HyperKvasir [1] and SurgVisdom [31] datasets with different conditions.

state-of-the-art results across various key metrics. These results validate that EndoGen could effectively leverage the long-range dependency modeling of autoregressive models in diverse scenarios in endoscopic video generation.

Ablation Studies. In Table 4, we ablate the components of EndoGen on the HyperKvasir dataset. Replacing SGP with a simple 2D reshaping of the video sequence results in a significant decline in performance, demonstrating the effectiveness of the proposed grid frame in capturing spatial and temporal information. Removing SAT also leads to reduced diversity and fidelity in the generated videos. Additionally, we explored various maximum thresholds $p_{\max}$ for masking, and setting it to 0.3 yields optimal performance, striking a balance between model learning complexity and capability enhancement.

Table 5. Performance comparison on the semi-supervised polyp segmentation task. Blue subscript denotes the improvement over the supervised baseline. The fg and bg denotes the foreground polyp and the background regions, respectively. Lab. denotes labeled real data. Unl.-Real denotes unlabeled real data. Unl.-Syn denotes unlabeled synthetic data by EndoGen. Bold refers to the best result.

Method	Lab.	Unl.-Real	Unl.-Syn	Dice (%)	IoU _{fg} (%)	IoU _{bg} (%)
Supervised	✓			69.75	61.72	90.58
FixMatch [20]	✓	✓		70.80	62.66	91.19
	✓		✓	70.96 _{↑1.21}	62.79 _{↑1.07}	91.49 _{↑0.91}
	✓	✓	✓	71.03 _{↑1.28}	63.14 _{↑1.42}	91.66 _{↑1.10}
PolypMix [9]	✓	✓		87.13	82.59	95.55
	✓		✓	87.84 _{↑18.09}	82.40 _{↑20.68}	95.47 _{↑4.89}
	✓	✓	✓	87.92 _{↑18.17}	82.41 _{↑20.69}	95.77 _{↑5.19}

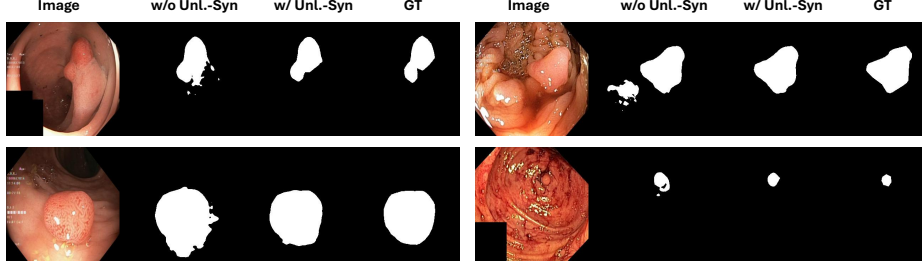


Fig. 4. Qualitative results of semi-supervised polyp segmentation.

3.3 Downstream Task: Semi-supervised Polyp Segmentation

Semi-supervised medical image segmentation is an essential approach that reduces the labeling cost for improved performance [14]. To evaluate the fidelity of EndoGen synthetic videos, we generate polyp frames as the unlabeled data for the semi-supervised polyp segmentation task, and train the segmentation model different semi-supervised methods [9,20]. We compare three training settings: using real unlabeled data (Unl.-Real); using synthetic unlabeled data (Unl.-Syn); and using both real and synthetic unlabeled data. We utilize 1,000 images from the HyperKvasir polyp segmentation dataset [1], splitting it into an 8:2 train-test ratio. In the training set, 10% of the images are labeled, while the remaining are unlabeled. Additionally, we randomly sample the same number of synthetic frames generated by EndoGen to create the unlabeled synthetic set. According to the results in Tab. 5, replacing real with synthetic data could even yield higher Dice scores of 70.96% with FixMatch [20] and 87.84% with PolypMix [9], demonstrating that EndoGen-generated data could effectively serves as a substitution of real data. Moreover, combining real and synthetic further improves performance, which indicates that the synthetic data complements real data and enhances overall segmentation quality. Fig. 4 gives a qualitative comparison be-

tween the segmentation results without and with our synthetic data. From the two cases in the left column, segmented results with our synthetic data capture better polyp structure and demonstrate more accurate boundary. In the right column, model trained with our additional unlabeled data performs better on small objects and effectively reduces false positives.

4 Conclusion

In this paper, we introduce EndoGen, an innovative framework for conditional autoregressive endoscopic video generation. EndoGen reformulates video sequence learning as a grid-frame pattern using SGP, and we propose an SAT strategy to enhance the diversity and clinical relevance of the generated results. Extensive validation has shown its superiority in both generation performance and downstream application. We hope that EndoGen will effectively support clinicians and advance research in medical generative models.

Acknowledgments. This work was supported by Innovation and Technology Commission Innovation and Technology Fund ITS/229/22, PRP/082/24FX and Hong Kong Research Grants Council (RGC) General Research Fund 14204321.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. 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)
2. 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)
3. Deng, H., Pan, T., Diao, H., Luo, Z., Cui, Y., Lu, H., Shan, S., Qi, Y., Wang, X.: Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169* (2024)
4. Diamantis, D.E., Gatoula, P., Iakovidis, D.K.: Endovae: Generating endoscopic images with a variational autoencoder. In: *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. pp. 1–5. IEEE (2022)
5. Esser, P., Rombach, R., Ommer, B.: Taming transformers for high-resolution image synthesis. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 12873–12883 (2021)
6. 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)
7. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)

8. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. *Advances in Neural Information Processing Systems* **35**, 8633–8646 (2022)
9. Jia, X., Shen, Y., Yang, J., Song, R., Zhang, W., Meng, M.Q.H., Liao, J.C., Xing, L.: Polypmixnet: Enhancing semi-supervised polyp segmentation with polyp-aware augmentation. *Computers in Biology and Medicine* **170**, 108006 (2024)
10. Lee, D., Kim, C., Kim, S., Cho, M., Han, W.S.: Autoregressive image generation using residual quantization. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 11523–11532 (2022)
11. 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. *arXiv preprint arXiv:2403.11050* (2024)
12. Li, T., Tian, Y., Li, H., Deng, M., He, K.: Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems* **37**, 56424–56445 (2025)
13. Li, W., Liu, X., Yang, Q., Yuan, Y.: From static to dynamic diagnostics: Boosting medical image analysis via motion-informed generative videos. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 195–205. Springer (2024)
14. Liu, X., Li, W., Yuan, Y.: Diffrect: Latent diffusion label rectification for semi-supervised medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 56–66. Springer (2024)
15. Liu, X., Yuan, Y.: A source-free domain adaptive polyp detection framework with style diversification flow. *IEEE Transactions on Medical Imaging* **41**(7), 1897–1908 (2022)
16. Ma, X., Wang, Y., Jia, G., Chen, X., Liu, Z., Li, Y.F., Chen, C., Qiao, Y.: Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048* (2024)
17. Mittal, S., et al.: A survey of accelerator architectures for 3d convolution neural networks. *Journal of Systems Architecture* **115**, 102041 (2021)
18. Sharma, V., Kumar, A., Jha, D., Bhuyan, M.K., Das, P.K., Bagci, U.: Controlpolypnet: towards controlled colon polyp synthesis for improved polyp segmentation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2325–2334 (2024)
19. 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)
20. 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** (2020)
21. Sun, P., Jiang, Y., Chen, S., Zhang, S., Peng, B., Luo, P., Yuan, Z.: Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525* (2024)
22. Tian, K., Jiang, Y., Yuan, Z., Peng, B., Wang, L.: Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems* **37**, 84839–84865 (2025)
23. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023)

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. Wang, Z., Liu, C., Zhang, S., Dou, Q.: Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
26. Xing, Z., Dai, Q., Hu, H., Wu, Z., Jiang, Y.G.: Simda: Simple diffusion adapter for efficient video generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7827–7839 (2024)
27. Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., Dai, Q.: Stat: Spatial-temporal attention mechanism for video captioning. IEEE transactions on multimedia **22**(1), 229–241 (2019)
28. Yan, W., Zhang, Y., Abbeel, P., Srinivas, A.: Videogpt: Video generation using vq-vae and transformers. arXiv preprint arXiv:2104.10157 (2021)
29. Yellapragada, S., Graikos, A., Prasanna, P., Kurc, T., Saltz, J., Samaras, D.: Pathldm: Text conditioned latent diffusion model for histopathology. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 5182–5191 (2024)
30. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
31. Zia, A., Bhattacharyya, K., Liu, X., Wang, Z., Kondo, S., Colleoni, E., van Amsterdam, B., Hussain, R., Hussain, R., Maier-Hein, L., et al.: Surgical visual domain adaptation: Results from the miccai 2020 surgvisdom challenge. arXiv preprint arXiv:2102.13644 (2021)