Endora: Video Generation Models as Endoscopy Simulators

Chenxin Li[1⋆], Hengyu Liu[1⋆], Yifan Liu[1⋆], Brandon Y. Feng[2], Wuyang Li[1], Xinyu Liu[1], Zhen Chen[3], Jing Shao[4], Yixuan Yuan[1(✉)]

[1] The Chinese University of Hong Kong, [2] Massachusetts Institute of Technology, [3] Centre for Artificial Intelligence and Robotics, Hong Kong , [4]Shanghai AI Lab

Abstract. Generative models hold promise for revolutionizing medical education, robot-assisted surgery, and data augmentation for machine learning. Despite progress in generating 2D medical images, the complex domain of clinical video generation has largely remained untapped. This paper introduces *Endora*, an innovative approach to generate medical videos that simulate clinical endoscopy scenes. We present a novel generative model design that integrates a meticulously crafted spatial-temporal video transformer with advanced 2D vision foundation model priors, explicitly modeling spatial-temporal dynamics during video generation. We also pioneer the first public benchmark for endoscopy simulation with video generation models, adapting existing state-of-the-art methods for this endeavor. *Endora* demonstrates exceptional visual quality in generating endoscopy videos, surpassing state-of-the-art methods in extensive testing. Moreover, we explore how this endoscopy simulator can empower downstream video analysis tasks and even generate 3D medical scenes with multi-view consistency. In a nutshell, *Endora* marks a notable breakthrough in the deployment of generative AI for clinical endoscopy research, setting a substantial stage for further advances in medical content generation. Project page: https://endora-medvidgen.github.io/.

Keywords: Medical Generative AI · Video Generation · Endoscopy

1 Introduction

Recent strides in generative AI have sparked significant interest across medical disciplines [20,17], pushing the frontiers of computer-aided diagnostics to new heights [38]. Amidst a broad array of endeavors in medical AI—ranging from visual question answering [5] and text summarizing [30] to image reconstruction [40] and translation [4], and even mixed reality for surgical assistance [27,26,19,41,15]—we venture into uncharted territories and ask: Can we generate dynamic, realistic, and complex content like clinical endoscopy videos?

Endoscopy is a field at the forefront for advances in gastrointestinal disease diagnosis, minimally invasive procedures, and robotic surgeries. Despite its critical

⋆ Equal contribution

role, endoscopic research and training are hindered by the scarcity of visual data, as capturing images inside the body with small endoscopes is inherently difficult. The dire need for a diverse and quality-rich collection of clinical endoscopy videos underscores the urgency for breakthroughs in medical generative AI. We aim to build a powerful endoscopy video simulator and create an extensive array of high-quality endoscopy videos, thus enriching the resources available for medical professionals and improving training data for surgical robots and AI algorithms. This exciting venture prompts us to probe deeper into several research questions: ❶ **Establishing Video Benchmarks**: Medical imaging and text have established benchmark applications, such as automated text report generation and image reconstruction [30]. Can we extend this success to medical videos and properly benchmark endoscopy simulation quality? ❷ **Spatial-Temporal Modeling**: While current methods are effective in generating realistic 2D clinical images by generative adversarial networks (GANs) and diffusion models [12], the dynamic nature of endoscopy videos, rich with spatial-temporal correlation, poses a significant challenge. Can our models effectively simulate the intricacies of real-life surgical procedures?

Driven by these questions, we formulate a framework to generate spatially and temporally coherent and plausible endoscopy videos to synthesize realistic clinical scenes. Marking a departure from traditional approaches to medical content generation that primarily deal with textual and 2D image data, we aspire to set **a holistic benchmark** for future explorations in video generation models within the medical domain. In particular, by meticulously crafting our model, *Endora*, for dynamic medical videos, we explore the initial experience baseline for pipeline design towards endoscopy video simulation (Fig. 1). We further pioneer the exploration of experimental baseline in endoscopic video generation, characterized by the comprehensive collection of clinical videos, and adapt existing generic video generation models for this purpose (Sec. 3.1 and Sec. 3.2). Simultaneously, we thoroughly investigate the extensive dimensions of evaluation baseline in video generation, including the fidelity of generated content, the improvement in performance for downstream video analysis through data augmentation, and the geometrical quality assessed by multi-view consistency in generation (Sec. 3.2 and Sec. 3.3).

To address the unique challenges of capturing the **spatio-temporal complexity and fluidity** of real-life medical procedures, we integrate an advanced video transformer architecture with a latent diffusion model, facilitating the extraction of long-range correlations in terms of both spatial and temporal dimension from video data. Specially, the training process involves using a pre-trained variational autoencoder [37,22] to encode video inputs into a latent space. These encoded features are then processed through a sequence of transformer blocks. Furthermore, to ensure consistency across video frames, we employ a feature prior from a 2D foundation model, DINO [7], which helps in regulating key features from different perspectives. Our extensive testing demonstrates that *Endora* can produce highly realistic endoscopic videos, showcasing its effectiveness and potential for medical video generation with rich dynamics. In

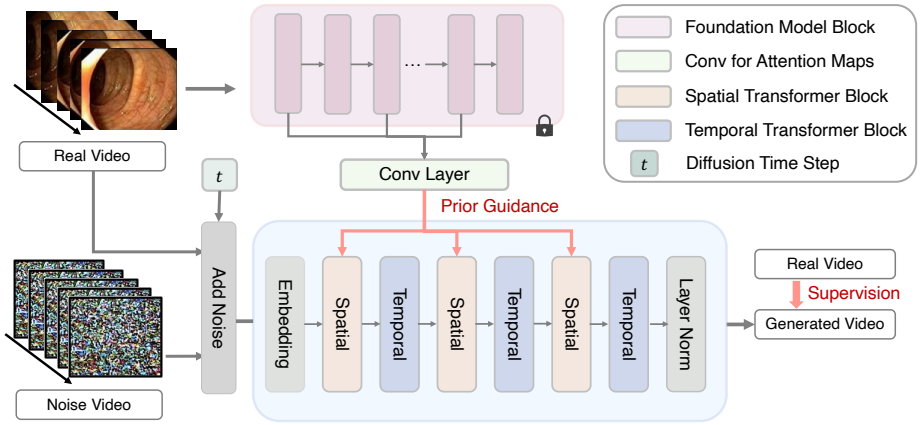

Fig. 1. ***Endora* Training Overview.** Starting from the noised input video sequences, the diffusion model iteratively removes the noise and recover the clean sequence. The long-range spatial-temporal dynamics is modeled by an interlaced cascading of several spatial-temporal Transformer blocks. We further instill the prior from 2D vision foundation model (DINO [7]) to guide feature extraction.

a nutshell, *Endora* leads the effort in creating complex, high-dimensional surgical video content, setting a benchmark for future medical generative AI research. Our key contributions include: (**i**) Introducing a high-fidelity medical video generation framework, tested on endoscopy scenes, laying the groundwork for further advancements in the field. (**ii**) Creating the first public benchmark for endoscopy video generation, featuring a comprehensive collection of clinical videos and adapting existing general-purpose generative video models for this purpose. (**iii**) Developing a novel technique to infuse generative models with features distilled from a 2D visual foundation model, ensuring consistency and quality across different scales. (**iv**) Demonstrating *Endora*'s versatility through successful applications in video-based disease diagnosis and 3D surgical scene reconstruction, highlighting its potential for downstream medical tasks.

2 Method

We train *Endora* to generate plausible endoscopy videos given the collection of actual clinical observations. Firstly, the diffusion backbone is introduced and lifted to handling video formats (Sec. 2.1). We further introduce cascaded transformer blocks interlaced for spatial and temporal modeling (Sec. 2.2). The prior from image foundation models is distilled to effectively guide the generation of the high-dimensional intricate video representation.

2.1 Diffusion Model for Video Generation

Generative diffusion models based on Denoising Diffusion Probability Models (DDPM) specializes in transforming disordered noise into desirable samples. By progressively removing noise from Gaussian noise $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{0}, \mathbf{I})$, these models generate samples aligned with the target data distribution. The forward diffusion step, denoted as $q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$, adds Gaussian noise into the image $\boldsymbol{x}_t$. The corresponding marginal distribution can be expressed as: $q(\boldsymbol{x}_t \mid \boldsymbol{x_0}) = \mathcal{N}\left(\alpha_t \boldsymbol{x_0}, \sigma_t^2 \boldsymbol{I}\right)$, where α_t and σ_t are designed to converge to $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ when t reaches the end of the forward process [13]. The reverse diffusion process $p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ is designed as a noise estimator $\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)$ that estimates noise from noisy images. The training process involves optimizing the weighted evidence lower bound (ELBO) [13],

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

where $\boldsymbol{\epsilon}$ is drawn from $\mathcal{N}(\mathbf{0}, \mathbf{I})$, the timestep t follows a uniform sampling, and $w(t)$ serves as a weighting function with $w(t) = 1$.

Lifting diffusion models for videos escalates computational overhead and representation complexity. Latent Diffusion Models[1] performs the diffusion processes in the encoded latent space rather than the pixel space, improving model efficiency [16]. Another strategy [11] trains video and image generation simultaneously to improve video generation quality. We adopt similar strategies in our framework but further introduce new innovations detailed below.

2.2 Spatial-temporal Transformer

Drawing insights from ViT [9] on spatial correlation capture [35], a transformer that exclusively extracts spatial information from tokens that share the same temporal index is introduced as a Spatial Transformer. We employ the patch embedding strategy to indicate the position embedding for this Spatial Transformer. A Temporal Transformer is further introduced to capture temporal information across video frames. We integrate temporal position embeddings accomplished by using an absolute position encoding strategy, which combines sinusoidal functions of varying frequencies. This strategy enables the model to accurately determine the exact position of each frame within the video sequence.

Specially, given a video clip in the latent space, denoted as $V \in \mathbb{R}^{F \times H \times W \times C}$, where F, H, W, and C denote the number of video frames, height, width, and channel of latent feature maps. We convert V into a sequence of tokens, represented as $\hat{Z} \in \mathbb{R}^{N_F \times N_H \times N_W \times D}$. The total number of tokens within a video clip in the latent space is $N_F \times N_H \times N_W$ and D represents the dimension of each token, respectively. A spatial-temporal positional embedding PE is integrated into $\hat{Z}$. Consequently, $Z = \hat{Z} + PE$ serves as the input for the Transformer backbone. We reshape Z into $Z^{\mathcal{S}} \in \mathbb{R}^{N_F \times L \times D}$ to serve as the input for the Spatial Transformer block, which captures spatial information. Here, $L = N_H \times N_W$

[1] https://github.com/CompVis/latent-diffusion

represents the token count of each temporal index. Subsequently, $Z^{\mathcal{S}}$, containing spatial information, is reshaped into $Z^{\mathcal{T}} \in \mathbb{R}^{L \times N_F \times D}$ as the input for Temporal Transformer block, which is used to capture temporal information. By interlacing a series of Spatial and Temporal Transformers, our model enables modeling long-range spatial correlations and temporal dynamics comprehensively.

2.3 Prior-guided Feature Facilitation

Compared to 2D contents, recovering video frames from noise is difficult, as we cannot adequately approximate the continuity of the temporal dimension, confined to sampling at specific quantized timestamps. Hence, it is better to optimize diffusion models to enable the inverse back projection of noise sequences into a latent time-continuous video space aligned with human perception [34,25,32]. However, they primarily consider semantic correlations between frames and do not adequately address dense correlations (e.g., patches, key points) across frames, which are crucial for frame continuity at a finer granularity.

Inspired by recent efforts of facilitating dense task [3] via vision foundation models like DINO [7], we consider leveraging the DINO features as they exhibit not only strong semantic correlation but also a potent ability to extract dense correspondence [3]. We propose to integrate the multi-scale representation [21] produced by DINO encoder, ranging from the outputs of its shallow layers to the deeper layers, as a prior obtained by large-scale 2D pre-training to guide the video diffusion training. We use attention maps from DINO [7], and apply a convolution operation with a stride of 2 and a 3x3 kernel to accommodate them with the dimension of diffusion attention maps. Considering the disparate regimes between DINO and *Endora*, we employ the relative distribution similarity to match features, by *Pearson correlation*,

$$Corr(A_{\textit{Endora}}, \mathrm{Conv}(A_{\texttt{DINO}})) = \frac{Cov\left(A_{\textit{Endora}}, \mathrm{Conv}\left(A_{\texttt{DINO}}\right)\right)}{\sqrt{Var(A_{\textit{Endora}})}\sqrt{Var\left(\mathrm{Conv}\left(A_{\texttt{DINO}}\right)\right)}} \tag{2}$$

where A is attention maps, $\mathrm{Conv}(\cdot)$ is the aforementioned convolution layer, $Cov(\cdot)$ is covariance, and $Var(\cdot)$ is variance. We simultaneously maximize this correlation of attention maps at multiple levels (four layers) between DINO encoder and *Endora* Spatial Transformer blocks, as shown in Fig. 1. With the guidance of the discriminative DINO prior, the semantic and spatial dependence can be thoroughly enhanced, boosting the photorealism of generated videos.

Finally, the overall optimization objective for training *Endora* is to minimize the following combination of the quantity in Eq. 1 and Eq. 2, $\mathcal{L}_{DDPM} + \alpha\mathcal{L}_{Prior} = ELBO + \alpha(1 - Corr)$. We set $\alpha = 0.5$ by grid searching.

3 Experiments

3.1 Experiment Settings

Datasets and Evaluation. We conduct comprehensive experiments on three public endoscopy video datasets: Colonoscopic [29], Kvasir-Capsule [6], CholecTriplet [31]. Following common practice [28], we extract 16-frame video clips

Table 1. Quantitative Comparisons on Endoscopic Video Datasets.

Method	Colonoscopic [29]			Kvasir-Capsule [6]			CholecTriplet [31]		
	FVD↓	FID↓	IS↑	FVD↓	FID↓	IS↑	FVD↓	FID↓	IS↑
StyleGAN-V [34](CVPR'22)	2110.7	226.14	2.12	183.5	31.61	2.77	594.1	87.46	3.36
LVDM [10](Arxiv'23)	1036.7	96.85	1.93	1027.8	200.9	1.46	1361.5	91.25	2.65
MoStGAN-V [33](CVPR'23)	468.5	53.17	3.37	82.77	17.34	2.53	416.2	72.87	3.56
Endora (Ours)	460.7	13.41	3.90	72.25	10.61	2.54	236.2	11.20	4.09

Table 2. Semi-supervised Classification Result (F1 Score) on PolyDiag [36].

Method	Colonoscopic [29]	CholeTriplet [31]
Supervised-only	74.5	74.5
LVDM [10]	76.2 (+1.7)	78.0 (+3.5)
Endora (Ours)	**87.0** (+**12.5**)	**82.0** (+**7.5**)

from these datasets using a specific sampling interval, with each frame resized to 128×128 resolution for training. In the assessment of quantitative comparisons, we employ three evaluation metrics: Fréchet Video Distance (FVD), Fréchet Inception Distance (FID), and Inception Score (IS). Adhering to the evaluation rules in StyleGAN-V [34], we compute FVD scores by analyzing 2048 video clips, each comprising 16 frames.

Implementation Details. We employ the AdamW optimizer, with a constant learning rate of 1×10^{-4} for training all models. We simply apply the basic data augmentation as horizontal flipping. Following standard practices in generative models, we use the exponential moving average (EMA) strategy [23] and report the performance on EMA model for final result sampling. We directly use the pre-trained variational autoencoder [8,39] from Stable Diffusion. Our model is constructed by $n = 28$ Transformer blocks, with a hidden dimension of $d = 1152$ with $n = 16$ multi-head attention in each block, following ViT [9].

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 endoscopic video datasets, including StyleGAN-V [34], MoStGAN-V [33] and LVDM [10]. As shown in Tab. 1, *Endora* excels over the state-of-the-art methods based on GAN [34,33] for endoscopic video generation in terms of high visual fidelity by all three metrics. Furthermore, *Endora* surpasses the advanced diffusion-based method, LVDM [10] in all aspects, indicating that *Endora* effectively generates a accurate video representation of endoscopic scenes. Fig. 2 further showcases the qualitative results of *Endora* and prior state-of-the-arts. We can observe that other techniques result in visually discordant distortions (row 1), restricted content variations (rows 2 and 4), and discontinuous inter-frame transitions (row 5,

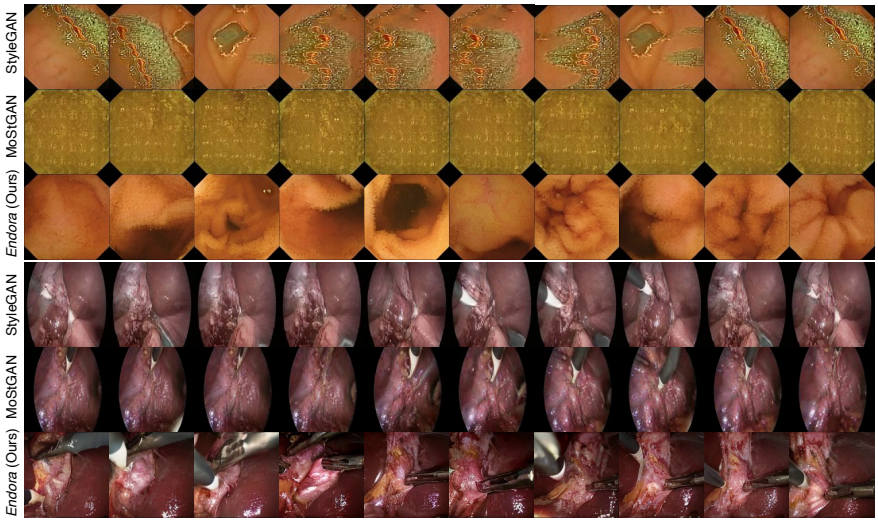

Fig. 2. Qualitative Comparison on Kvasir-Capsule [6] and Cholec [31] Datasets.

Table 3. Ablation Studies of Proposed Components on Colonoscopic [29] Dataset.

Modified Diffusion	Spatiotemporal Encoding	Prior Guidance	FVD↓	FID↓	IS↑
✗	✗	✗	611.9	22.44	3.61
✓	✗	✗	593.7	17.75	3.65
✓	✓	✗	493.5	13.88	3.89
✓	✓	✓	**460.7**	**13.41**	**3.90**

abrupt intrusion of surgical instruments). In contrast, the video frames generated by *Endora* (rows 3 and 6) avoid discordant visual distortions (row 1), retain more visual details, and offer superior visual representation of tissues.

3.3 Further Empirical Studies

In this section, we illustrate several potential applications of leveraging the generated videos of our *Endora* and conduct rigorous ablations on our key strategies.
Case I: *Endora* as a Temporal Data Augmenter. We explore the case of using generated videos as the unlabeled instances for semi-supervised training (by FixMatch [1]) on the video-based disease diagnosis benchmark (PolyDiag [36]). Specially, we use the randomly selected $n_l = 40$ videos in training set of PolyDiag as labeled data, and $n_u = 200$ generated videos as the unlabeled data in Colonoscopic [29]and CholecTriplet [31], respectively. Tab. 2 depict the F1 score of disease diagnosis, with the **gain** over the baseline using merely labeled training instances labeled (Supervised-Only). The outcomes indicate a notable

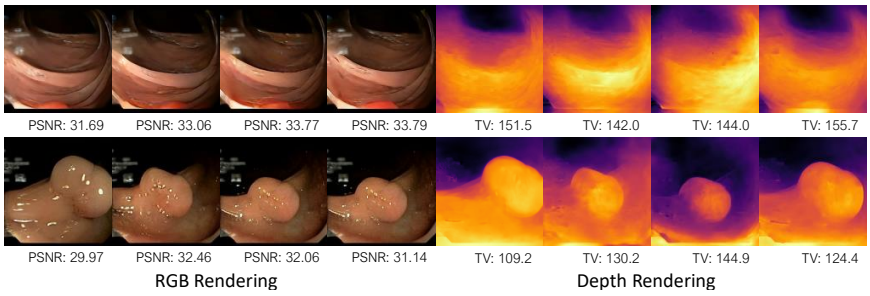

Fig. 3. RGB and 3D Depth Reconstructed from Generated Videos.

enhancement in downstream performance attributable to the data generated by *Endora* compared to not only the Supervised-only baseline but also other video generation methodologies, confirming the efficacy of *Endora* as a reliable video data augmenter for downstream video analysis.

Case II: *Endora* as a Surgical World Simulator. Emerging multi-view consistent properties in generated contents [24,14,18] inspire our exploration into whether a similar geometric consistency exists in our generated surgical videos. Specially, from a generated video, we take some frames to as training data for 3D reconstruction (training views), and keep the other frames as testing data (novel views). We then preprocess the training views with COLMAP [2] and then run an off-the-shelf 3D reconstruction pipeline (EndoGaussian [27]) to obtain a reconstructed 3D endoscopy scene. Fig. 3 provides the visualization of the rendered RGB images and depth maps at novel views, with the image PSNR and depth-wise total variation (TV) labeled. We can observe that the 3D scenes reconstructed from our generated videos exhibit realistic and continuous geometric structure, showing the potential of *Endora* to effectively perform surgical world simulation in a multi-view-consistent manner.

Ablation Studies. Table 3 presents an ablation study of the key components of the proposed *Endora*. Initially, we employ a plain video diffusion model without any of the proposed strategies as a baseline. Subsequently, we add the proposed three design strategies one at a time: modified diffusion (Sec. 2.1), spatial-temporal encoding (Sec. 2.2), and prior guidance (Sec. 2.3). We can observe they lead to a steady progression in the performance of model, confirming the crucial role of our designed strategies in enhancing the overall efficiency and effectiveness of the endoscopy video generation model.

4 Conclusion

*Endora*is a pioneering framework for medical video generation, producing high-quality, realistic endoscopy simulations. It combines a video transformer for long-range spatial-temporal modeling with priors from advanced 2D vision foundation models for enhanced feature extraction. Rigorous benchmarking shows *Endora*'s superior visual quality and potential for data augmentation in downstream video

analysis. It enables 3D surgical scene simulation when paired with existing endoscopy reconstruction methods. *Endora*advances medical generative AI, providing insights and foundations for future research in medical content generation.

Acknowledgments. This work was supported by Hong Kong Innovation and Technology Commission Innovation and Technology Fund ITS/229/22 and Research Grants Council (RGC) General Research Fund 14204321, 11211221.

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

Author contributions. Conceptualization: C. Li. Methodology: C. Li, Y. Liu, B. Feng. Implementation: C. Li, H. Liu, Y. Liu. Writing: C. Li, B. Feng. Experiment Design: C. Li, B. Feng, W. Li. Visualization: C. Li, H. Liu, B. Feng, W. Li. Supervision: X. Liu, Z. Chen, J. Shao, Y. Yuan.

References

1. https://github.com/google-research/fixmatch 7
2. https://github.com/colmap/colmap 8
3. Amir, S., Gandelsman, Y., Bagon, S., Dekel, T.: Deep vit features as dense visual descriptors. arXiv preprint arXiv:2112.05814 **2**(3), 4 (2021) 5
4. Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., Yang, B.: Medgan: Medical image translation using gans. Computerized medical imaging and graphics **79**, 101684 (2020) 1
5. Ben Abacha, A., Hasan, S.A., Datla, V.V., Demner-Fushman, D., Müller, H.: Vqa-med: Overview of the medical visual question answering task. In: Proceedings of CLEF 2019 Working Notes. 9-12 September 2019 (2019) 1
6. 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), 1–14 (2020) 5, 6, 7
7. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV. pp. 9650–9660 (2021) 2, 3, 5
8. Ding, Z., Dong, Q., Xu, H., Li, C., Ding, X., Huang, Y.: Unsupervised anomaly segmentation for brain lesions using dual semantic-manifold reconstruction. In: ICONIP. pp. 133–144. Springer (2022) 6
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 4, 6
10. 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 (2023) 6
11. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. In: NeurIPS (2022) 4
12. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models for medical image analysis: A comprehensive survey. arXiv preprint arXiv:2211.07804 (2022) 2

13. Kingma, D., Salimans, T., Poole, B., Ho, J.: Variational diffusion models. NeurIPS **34**, 21696–21707 (2021) 4
14. Li, C., Feng, B.Y., Fan, Z., Pan, P., Wang, Z.: Steganerf: Embedding invisible information within neural radiance fields. In: CVPR. pp. 441–453 (2023) 8
15. Li, C., Feng, B.Y., Liu, Y., Liu, H., Wang, C., Yu, W., Yuan, Y.: Endosparse: Real-time sparse view synthesis of endoscopic scenes using gaussian splatting. arXiv preprint arXiv:2407.01029 (2024) 1
16. Li, C., Lin, M., Ding, Z., Lin, N., Zhuang, Y., Huang, Y., Ding, X., Cao, L.: Knowledge condensation distillation. In: ECCV, pages=19–35, year=2022, organization=Springer 4
17. Li, C., Lin, X., Mao, Y., Lin, W., Qi, Q., Ding, X., Huang, Y., Liang, D., Yu, Y.: Domain generalization on medical imaging classification using episodic training with task augmentation. CBM **141**, 105144 (2022) 1
18. Li, C., Liu, H., Fan, Z., Li, W., Liu, Y., Pan, P., Yuan, Y.: Gaussianstego: A generalizable stenography pipeline for generative 3d gaussians splatting. arXiv preprint arXiv:2407.01301 (2024) 8
19. 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) 1
20. Li, C., Liu, X., Li, W., Wang, C., Liu, H., Yuan, Y.: U-kan makes strong backbone for medical image segmentation and generation. arXiv:2406.02918 (2024) 1
21. Li, C., Ma, W., Sun, L., Ding, X., Huang, Y., Wang, G., Yu, Y.: Hierarchical deep network with uncertainty-aware semi-supervised learning for vessel segmentation. Neural Computing and Applications pp. 1–14 (2022) 5
22. Li, C., Zhang, Y., Li, J., Huang, Y., Ding, X.: Unsupervised anomaly segmentation using image-semantic cycle translation. arXiv preprint arXiv:2103.09094 (2021) 2
23. Li, C., Zhang, Y., Liang, Z., Ma, W., Huang, Y., Ding, X.: Consistent posterior distributions under vessel-mixing: a regularization for cross-domain retinal artery/vein classification. In: ICIP. pp. 61–65. IEEE (2021) 6
24. Li, X., Zhou, D., Zhang, C., Wei, S., Hou, Q., Cheng, M.M.: Sora generates videos with stunning geometrical consistency. arXiv preprint arXiv:2402.17403 (2024) 8
25. Liang, Z., Rong, Y., Li, C., Zhang, Y., Huang, Y., Xu, T., Ding, X., Huang, J.: Unsupervised large-scale social network alignment via cross network embedding. In: CIKM. pp. 1008–1017 (2021) 5
26. Liu, H., Liu, Y., Li, C., Li, W., Yuan, Y.: Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. arXiv:2406.16073 (2024) 1
27. Liu, Y., Li, C., Yang, C., Yuan, Y.: Endogaussian: Gaussian splatting for deformable surgical scene reconstruction. arXiv:2401.12561 (2024) 1, 8
28. 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:2401.03048 (2024) 5
29. Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE TMI **35**(9), 2051–2063 (2016) 5, 6, 7
30. Mishra, R., Bian, J., Fiszman, M., Weir, C.R., Jonnalagadda, S., Mostafa, J., Del Fiol, G.: Text summarization in the biomedical domain: a systematic review of recent research. Journal of biomedical informatics **52**, 457–467 (2014) 1, 2
31. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. MedIA **78**, 102433 (2022) 5, 6, 7

32. Pan, P., Fan, Z., Feng, B.Y., Wang, P., Li, C., Wang, Z.: Learning to estimate 6dof pose from limited data: A few-shot, generalizable approach using rgb images. arXiv preprint arXiv:2306.07598 (2023) 5
33. Shen, X., Li, X., Elhoseiny, M.: Mostgan-v: Video generation with temporal motion styles. In: Computer Vision and Pattern Recognition. pp. 5652–5661 (2023) 6
34. Skorokhodov, I., Tulyakov, S., Elhoseiny, M.: Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: Computer Vision and Pattern Recognition. pp. 3626–3636 (2022) 5, 6
35. Sun, L., Li, C., Ding, X., Huang, Y., Chen, Z., Wang, G., Yu, Y., Paisley, J.: Few-shot medical image segmentation using a global correlation network with discriminative embedding. CBM **140**, 105067 (2022) 4
36. 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: MICCAI. pp. 88–98. Springer (2022) 6, 7
37. Wang, Y., Yao, H., Zhao, S.: Auto-encoder based dimensionality reduction. Neurocomputing **184**, 232–242 (2016) 2
38. Xu, H., Li, C., Zhang, L., Ding, Z., Lu, T., Hu, H.: Immunotherapy efficacy prediction through a feature re-calibrated 2.5 d neural network. Computer Methods and Programs in Biomedicine **249**, 108135 (2024) 1
39. Xu, H., Zhang, Y., Sun, L., Li, C., Huang, Y., Ding, X.: Afsc: Adaptive fourier space compression for anomaly detection. arXiv:2204.07963 (2022) 6
40. Zhang, Y., Li, C., Lin, X., Sun, L., Zhuang, Y., Huang, Y., Ding, X., Liu, X., Yu, Y.: Generator versus segmentor: Pseudo-healthy synthesis. In: MICCAI. pp. 150–160. Springer (2021) 1
41. Zhu, L., Wang, Z., Jin, Z., Lin, G., Yu, L.: Deformable endoscopic tissues reconstruction with gaussian splatting. arXiv preprint arXiv:2401.11535 (2024) 1