SurgSora: Object-Aware Diffusion Model for Controllable Surgical Video Generation

Tong Chen[1*], Shuya Yang[2*], Junyi Wang[3*], Long Bai[3†], Hongliang Ren[3], and Luping Zhou[1†]

[1] The University of Sydney, Sydney, Australia
[2] The University of Hong Kong, Hong Kong SAR, China
[3] The Chinese University of Hong Kong, Hong Kong SAR, China
tche2095@uni.sydney.edu.au, b.long@ieee.org, luping.zhou@sydney.edu.au

Abstract. Surgical video generation can enhance medical education and research, but existing methods lack fine-grained motion control and realism. We introduce SurgSora, a framework that generates high-fidelity, motion-controllable surgical videos from a single input frame and user-specified motion cues. Unlike prior approaches that treat objects indiscriminately or rely on ground-truth segmentation masks, SurgSora leverages self-predicted object features and depth information to refine RGB appearance and optical flow for precise video synthesis. It consists of three key modules: (1) the Dual Semantic Injector, which extracts object-specific RGB-D features and segmentation cues to enhance spatial representations; (2) the Decoupled Flow Mapper, which fuses multi-scale optical flow with semantic features for realistic motion dynamics; and (3) the Trajectory Controller, which estimates sparse optical flow and enables user-guided object movement. By conditioning these enriched features within the Stable Video Diffusion, SurgSora achieves state-of-the-art visual authenticity and controllability in advancing surgical video synthesis, as demonstrated by extensive quantitative and qualitative comparisons. Our human evaluation in collaboration with expert surgeons further demonstrates the high realism of SurgSora-generated videos, highlighting the potential of our method for surgical training and education. Our project is available at surgsora.github.io.

Keywords: Surgical Video · Diffusion Model · Video Generation.

1 Introduction

Surgical video generation has the potential to enhance medical education, clinician training, and AI-driven surgical analysis by providing realistic and controllable visual representations of complex procedures [2, 15, 19]. However, existing methods face two key challenges: visual authenticity and generation controllability. Different strategies have been proposed to generate high-quality endoscopic

* Equal Contribution; † Corresponding Author.

videos. Endora [9] is an endoscopy simulator capable of replicating diverse endoscopic scenarios while lacking precise controlling abilities. MedSora [24] focuses on past temporal coherence to achieve high-fidelity forward video generation, while predicting based on the optical flow of past frames is challenging due to the complexity of surgical scenaris. Moreover, most current models indiscriminately process entire scenes, failing to differentiate individual objects, leading to blurred boundaries and unrealistic motion dynamics. Iliash et al. [7] attempt to improve object awareness using segmentation masks, while this approach requires additional object-level annotations and hard object boundaries, limiting the ability to model smooth transitions and fine anatomical details.

Beyond realism, precise motion control is critical for replicating surgical actions such as tool manipulation and tissue interaction [22]. However, most existing methods lack this capability, making it difficult to generate videos that follow user-specified actions. In the context of general video generation, text prompts are a commonly used control method [2,7,10,19,23]. However, surgical scenarios often involve precise and detailed actions, which are difficult to fully capture through textual descriptions. This leads to a domain gap between language and motion, making it challenging for text-based guidance to express fine-grained surgical actions. In addition, surgical triplets (tissue-instrument-action) have also been used for video generation tasks [13,14,27], while triplets do not include specific motion and directional information, making it difficult to comprehensively represent the rich information contained in surgical videos.

To this end, we propose SurgSora, an object-aware trajectory-controlled RGB-Depth Flow Diffusion model that introduces self-predicted object-relevant RGB and depth features for precise object localization — eliminating the need for segmentation masks. Instead of relying on text-based control, SurgSora utilizes motion trajectories, offering fine-grained, user-directed control over object movements, enabling the generation of realistic and clinically relevant surgical videos. We can freely generate realistic videos by providing motion guidance using our SurgSora, offering abundant surgical training and education materials.

Our contribution can be summarized as: **(i)** We present the first work on motion-controllable surgical video generation using a diffusion model, which allows fine-grained control (both direction and magnitude) over the motion of surgical instruments and tissues, guided by intuitive motion cues provided by simple clicks. **(ii)** We propose the Dual Semantic Injector (DSI), which integrates object-aware RGB-D semantic understanding. The DSI combines appearance (RGB) and depth information to better discriminate objects and capture complex anatomical structures, providing an accurate representation of the surgical scene. **(iii)** We introduce the Decoupled Flow Mapper (DFM), which effectively fuses optical flow with semantic-RGB-D features at multiple scales. This fusion serves as the guidance conditions for a frozen Stable Video Diffusion model to generate realistic surgical video sequences. **(iv)** Extensive experiments on the CoPESD dataset demonstrate the effectiveness of SurgSora in generating high-quality, motion-controllable surgical videos. We further involve surgeons verifying the authenticity of our videos to justify their usefulness in medical use.

2 Methodology

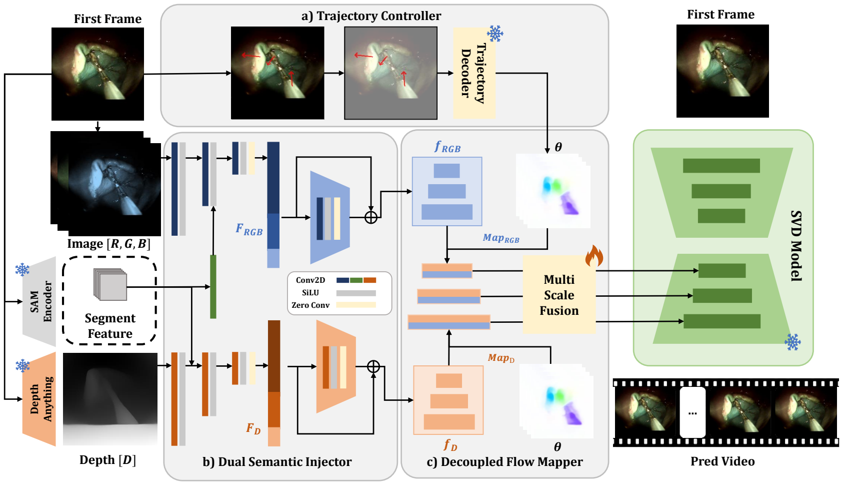

Fig. 1. SurgSora Pipeline: a) Trajectory Controller module encodes trajectories into sparse optical flow. **b)** Dual Semantic Injector merges RGB and depth features with segment features separately. **c)** Decoupled Flow Mapper maps RGB and depth features into optical flow, then sends to Multi-Scale Fusion Block as condition.

Our SurgSora framework, as illustrated in Fig. 1, comprises three key modules: the Dual Semantic Injector (DSI) introduced in Sec. 2.1, the Decoupled Flow Mapper (DFM) described in Sec. 2.2, and the Trajectory Controller (TC) module detailed in Sec. 2.3. Our model takes the first image frame $I_{RGB} \in \mathbb{R}^{3\times H\times W}$ as input. Based on I_{RGB}, the corresponding segmentation features f_{seg} and depth image $I_D \in \mathbb{D}^{1\times H\times W}$ are generated from the pretrained Segment Anything Model [8] and the Depth Anything V2 [26]. The segment feature is injected into the RGB and Depth features in the DSI module to extract object-aware image features f^r_{RGB} and depth features f^r_D at multi-scales r. These features are then processed in the DFM module, where the optical flow $\theta \in \mathbb{O}^{(T-1)\times 2\times H\times W}$ (with T as the total number of frames of the generated video), is resized and used to transform f^r_{RGB} and f^r_D independently. The transformed features are fused using the Multi-Scale Fusion (MSF) Block at different scales. These multi-scale fused features are then used as conditions for a frozen Stable Video Diffusion (SVD) model to generate the video.

2.1 Dual Semantic Injector

Traditional methodologies primarily rely on RGB images as input to create dynamic visual content. While effective in certain applications, this approach suffers from significant limitations in depth perception and scene understanding.

Specifically, relying solely on RGB data complicates accurately capturing spatial relationships between objects, leading to deficiencies in visual coherence and object segmentation in generated videos. To address these challenges, we introduce the Dual Semantic Injector (DSI) module, a dual-branch architecture that enhances object awareness by integrating segmentation features into both the RGB and depth feature branches. Unlike traditional methods that depend solely on RGB images, we estimate and incorporate a depth map to provide crucial geometric cues. These cues improve the understanding of spatial relationships between objects and overall scene structure, making it especially beneficial for complex tasks like surgical video synthesis. Furthermore, to better discriminate between objects, object segmentation is leveraged to refine both RGB and depth features. The segment features f_{seg} are combined with RGB images I_{RGB} and depth images I_D by passing through two separate processors ϕ_{RGB} and ϕ_D for feature extraction and fusion, followed by two separate encoders for further encoding. The Dual Semantic Injector can be formulated as:

$$f^r = \begin{cases} \boldsymbol{Encoder^r_{RGB}}(\boldsymbol{\phi_{RGB}}(I_{RGB}, f_{seg})), or \\ \boldsymbol{Encoder^r_{D}}(\boldsymbol{\phi_{D}}(I_D, f_{seg})). \end{cases} \tag{1}$$

Recall that the superscript r indicates different scales of feature maps extracted by the encoders. This dual encoding method synchronizes and harmonizes the enhanced features from RGB and depth channels to optimize the overall representation. The segmentation features enhance the semantic understanding compared with using the original RGB and depth features, significantly improving the discrimination of foreground and background, enhancing depth estimation, and ultimately contributing to more realistic and referenceable video predictions.

2.2 RGB-Depth Frame Mapper

Recent studies [12, 17, 30] have highlighted the advantages of integrating additional latent space information into diffusion models to enhance their output quality. Building on this insight, our Decoupled Flow Mapper (DFM) module integrates spatial and temporal data from image and optical features to generate sequential videos effectively. The DFM uses object-aware RGB and depth features from the Dual Semantic Injector (DSI) module, transformed spatially by resized optical flow to produce enriched video sequences.

Specifically, feature maps $f^r \in \mathbb{R}^{C_r \times H_r \times W_r}$ from the DSI module are modified according to the resized optical flow $\theta^r \in {\mathbb{O}'}^{(T-1)\times 2 \times H_r \times W_r}$, aligning them spatially for each frame. The features are shifted by $\mathrm{d}x$ and $\mathrm{d}y$, the displacements provided by θ^r_t, with new positions computed as $x' = x + \mathrm{d}x$ and $y' = y + \mathrm{d}y$. Interpolation determines the new pixel values at these coordinates, effectively merging the depth information's structural insights with the RGB data's textural details. This decoupled approach ensures that each feature type is optimally utilized, enhancing the video generation process. The fused features from separate streams are then passed the *Multi-Scale Fusion Block* (MSF) for final video output. The MSF block shall fuse the optical-flow-transformed RGB and depth features by concatenating them at different scales and then fusing them with 3D convolution blocks and an activation block, which can be formulated as:

$$\check{f}^{r}_{fuse} = \mathrm{SiLU}(\mathrm{Conv3d}(\mathrm{Conv3d}(\mathrm{ConCAT}(\hat{f}^{r}_{RGB}, \hat{f}^{r}_{D})))). \tag{2}$$

The fused feature $\check{f}^{r}_{fuse}$ is then used to assist a pretrained Stable Video Diffusion (SVD) Model for conditional video generation. The integration of optical flow information substantially enhances the temporal continuity and smoothness for generation, enabling accurate capture of scenes and complex objective motions. By utilizing fused data, the model gains a richer understanding of scene depth and structure, ensuring visual authenticity and adaptability to nuanced changes. This approach improves video quality through diversity-aware fusion, providing a dynamic and precise scene representation.

2.3 Trajectory Controller

Surgical videos demand higher precision compared to natural scene videos, particularly because generating them from a single image introduces significant challenges and reduces referential accuracy. Due to this we employ trajectory for precision control, which employs a pre-trained trajectory decoder from [29]. The surgeon inputs the first frame image and then clicks to set trajectories. The trajectories and image will be encoded separately, concatenated together, and then decoded by the trajectory decoder into optical flows as a condition to guide the following generation. By involving the TC module, generated videos will be more referencable and convenient for generating customized surgical videos.

3 Experiment

Dataset and Implementation Details. We utilize the publicly available CoPESD dataset [21], which was collected from 20 videos using both conventional endoscopic submucosal dissection (ESD) and the DREAMS system [3], performed on in-vivo porcine models. The videos were recorded at a frame rate of 30 Hz with an original resolution of 1920 × 1080, which was cropped to 1300 × 1024. After an expert surgeon provided temporal annotations of ESD activities, video segments corresponding to submucosal dissection were separately extracted. We extract 21-frame video clips from the datasets, then resize them into 256 × 197 resolution and pad them to 256 × 256 resolution as the training and testing set. for training. For the training process, we use the AdamW [11] with a learning rate of 2×10^{-5}, a batch size of 4, and train on two A6000 GPUs.

Comparison Methods and Evaluation Metrics. We assess the performance of our model against leading video generation models: the unconditional models Endora [9] and StyleGAN-V [18]; and the conditional model MOFA-Video [12]. For video quality evaluation, we employ metrics such as Fréchet Video Distance (FVD) [20] and Content-Debiased Fréchet Video Distance (CD-FVD) [4] to assess temporal consistency and video realism. Frame realism and diversity are evaluated using Fréchet Inception Distance (FID) [5] and Inception Score (IS) [1]. Frame consistency is measured by CLIP cosine similarity [16] between consecutive frames. For conditional generation, we employ PSNR [6] and SSIM [25] for content and structural accuracy, alongside optical flow metrics F1-epe and F1-all, to verify flow consistency with the input trajectories.

3.1 Experimental Results

Table 1. Quantitative Comparisons on CoPESD Dataset [21].

Models	Discriminate Avg Acc	Frame Consistency↑	Video						Optical-Flow	
			FVD↓	CD-FVD↓	FID↓	IS ↑	PSNR ↑	SSIM ↑	F1-epe↓	F1-all↓
Endora [9]	–	97.51%	1146.54	1289.59	205.93	2.239	–	–	–	–
StyleGAN-V [28]	–	98.02%	857.16	980.11	166.03	2.267	–	–	–	–
MOFA-Video [12]	–	95.59%	671.66	692.44	96.31	2.685	19.06	48.28%	0.2620	265.54
SurgSora (Ours)	53.33%	**98.70%**	**395.65**	**535.95**	**87.94**	**3.278**	**20.71**	**55.94%**	**0.1477**	**149.89**

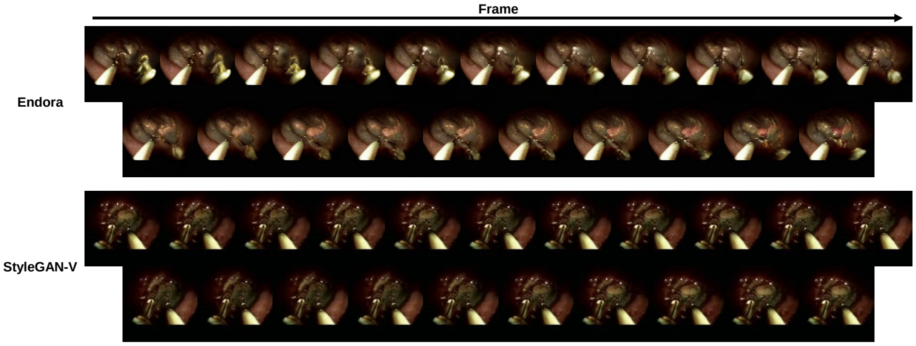

Fig. 2. Comparison of unconditional results on CoPESD dataset.

Table 1 quantitatively assesses SurgSora alongside existing generative models, highlighting our approach's superior performance in video generation. SurgSora achieves the highest frame consistency at 98.70%, surpassing Endora (97.51%) and StyleGAN-V (98.02%), indicating enhanced temporal stability. Besides, it also achieves the best FVD score of 395.65, outperforming StyleGAN-V and Endora, which showcase FVD scores of 857.16 and 1146.54, respectively. Additionally, SurgSora excels in content and dynamics with a CD-FVD score of 535.95, markedly better than MOFA-Video (671.66). Furthermore, SurgSora leads in visual fidelity with the best FID of 87.94, demonstrating its capacity to generate videos that closely mimic real surgical scenes. The model also tops IS at 3.278, indicating superior object and diversity representation. For conditional generation, SurgSora outperforms MOFA-Video with scores of 20.71 dB in PSNR and 55.94% in SSIM compared to MOFA-Video (19.06 dB and 48.28%), which confirms promising capacity in preserving generation quality and structural similarity. Visual evidence from Fig. 2 and Fig. 3 shows SurgSora's superior

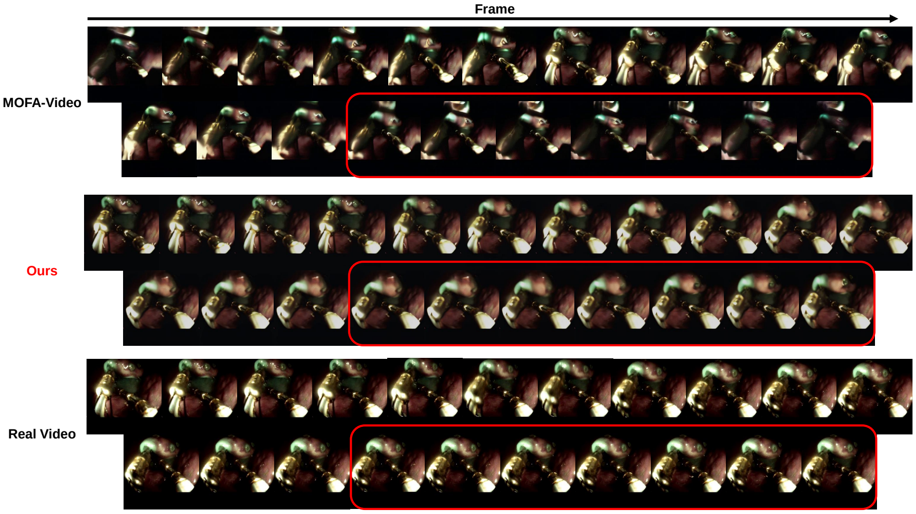

Fig. 3. Comparison of conditional generation results trained on CoPESD dataset.

performance, with videos that maintain authenticity and suffer less distortion compared to other models, showing its practical efficacy in medical applications.

Customize Trajectory Video Generation. To address the effectiveness of our Trajectory Controller block, we generate a few demos by using the TC module. Fig. 4 (a)(b) shows videos generated from varying surgical image trajectories, clearly depicting the dynamic movement and transformation of the objects within the images in accordance with the specified trajectories. Further demonstrating the module's capacity for precise control, we generated distinct trajectories within the same image, as presented in Fig. 4(c)(d)(e). We manipulated tissues and instruments to move in designated directions, and the visual results showed that objects along the set paths moved without noticeable distortion. The heatmap indicates that devices have undergone obvious changes according to the trajectory requirements while maintaining the background unchanged. These outcomes validate the high performance and accuracy of our module in controlling and generating detailed movement in medical imagery.

Human Evaluation. A blind test was conducted to assess the perceptual quality of video clips generated by SurgSora. We randomly selected seven clips from SurgSora-generated videos and eight from real surgical recordings (15 clips in total). These clips were shuffled and anonymized before being presented to six surgeons, who were individually asked to distinguish between generated and real videos. The results, shown in Fig. 1, indicate an average accuracy of 53.33%, which is close to random guessing (50%), suggesting that SurgSora produces highly realistic surgical videos that are difficult to differentiate from real footage.

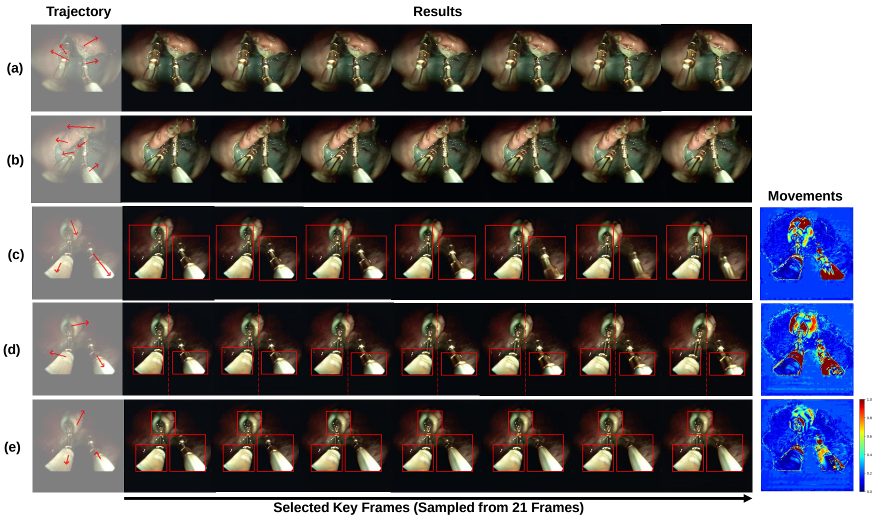

Fig. 4. Quantitative results of SurgSora within different trajectories and samples.

Table 2. Ablation experiments of our SurgSora on the CoPESD Dataset [21].

Segment Feature	Depth Branch	Multi-Scale Fusion	Frame Consistancy↑	FVD↓	CD-FVD↓	FID↓	IS ↑	PSNR↑	SSIM↑
✓	✗	–	98.08%	442.66	584.46	90.97	3.199	20.47	53.59%
✗	✓	✗	96.99%	510.11	782.51	115.72	2.586	19.49	54.59%
✗	✓	✓	98.35%	479.13	624.63	88.85	3.076	17.59	53.18%
✓	✓	✗	97.53%	422.06	603.34	88.54	3.270	20.64	51.33%
✓	✓	✓	**98.70%**	**395.65**	**535.95**	**87.94**	**3.278**	**20.71**	**55.94%**

Ablation Study. We carried out ablation studies on the SurgSora model using the CoPESD Dataset [21] to assess the impact of different components, summarized in Table 2. Results highlight that removing the segment feature increases FVD/CD-FVD from 395.65/535.95 to 479.13/624.63, underscoring its role in improving visual and temporal coherence. Eliminating the MSF block further degrades performance, with Frame Consistency dropping to 96.99% and a larger increase in FVD/CD-FVD to 510.11/782.51, emphasizing the depth branch's role in spatial integration). Disabling only the MSF block results in a milder performance drop. The worst structural preservation indicated by the lowest SSIM occurs when decoupled-flow features are used alone, but performance improves with the addition of the depth branch. Integrating all three components yields the best results, confirming their collective utility in video generation quality.

4 Conclusion

In this study, we propose SurgSora, a customized RGBD-flow-guided conditional diffusion video model. SurgSora incorporates a separate depth branch, the Dual

Semantic Injector (DSI), which increases object semantics information for dual features, and the Decoupled Flow Mapper (DFM) to provide a more suitable and richer feature representation for the Stable Video Diffusion model. Quantitative and qualitative experiments demonstrate superior performance in medical video generation and the ability to generate reasonable videos with simple trajectories. SurgSora provides a brand new view on the medical video generation field. Future works will focus on high-quality long medical clip generation and multimodal conditional medical video generation.

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

References

1. Barratt, S., Sharma, R.: A note on the inception score. arXiv preprint arXiv:1801.01973 (2018)
2. Cho, J., Schmidgall, S., Zakka, C., Mathur, M., Kaur, D., Shad, R., Hiesinger, W.: Surgen: Text-guided diffusion model for surgical video generation. arXiv preprint arXiv:2408.14028 (2024)
3. Gao, H., Yang, X., Xiao, X., Zhu, X., Zhang, T., Hou, C., Liu, H., Meng, M.Q.H., Sun, L., Zuo, X., et al.: Transendoscopic flexible parallel continuum robotic mechanism for bimanual endoscopic submucosal dissection. The International Journal of Robotics Research **43**(3), 281–304 (2024)
4. 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 (CVPR) (2024)
5. 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)
6. Huynh-Thu, Q., Ghanbari, M.: Scope of validity of psnr in image/video quality assessment. Electronics letters **44**(13), 800–801 (2008)
7. Iliash, I., Allmendinger, S., Meissen, F., Kühl, N., Rückert, D.: Interactive generation of laparoscopic videos with diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 109–118. Springer (2024)
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
9. 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)
10. Li, Y., Min, M., Shen, D., Carlson, D., Carin, L.: Video generation from text. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
11. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
12. Niu, M., Cun, X., Wang, X., Zhang, Y., Shan, Y., Zheng, Y.: Mofa-video: Controllable image animation via generative motion field adaptions in frozen image-to-video diffusion model. arXiv preprint arXiv:2405.20222 (2024)

13. Nwoye, C.I., Bose, R., Elgohary, K., Arboit, L., Carlino, G., Lavanchy, J.L., Mascagni, P., Padoy, N.: Surgical text-to-image generation. Pattern Recognition Letters (2025)
14. 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. Medical Image Analysis **78**, 102433 (2022)
15. Ozawa, T., Hayashi, Y., Oda, H., Oda, M., Kitasaka, T., Takeshita, N., Ito, M., Mori, K.: Synthetic laparoscopic video generation for machine learning-based surgical instrument segmentation from real laparoscopic video and virtual surgical instruments. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization **9**(3), 225–232 (2021)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
17. 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)
18. 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)
19. Sun, W., You, X., Zheng, R., Yuan, Z., Li, X., He, L., Li, Q., Sun, L.: Bora: Biomedical generalist video generation model. arXiv preprint arXiv:2407.08944 (2024)
20. 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)
21. Wang, G., Xiao, H., Gao, H., Zhang, R., Bai, L., Yang, X., Li, Z., Li, H., Ren, H.: Copesd: A multi-level surgical motion dataset for training large vision-language models to co-pilot endoscopic submucosal dissection. arXiv preprint arXiv:2410.07540 (2024)
22. Wang, S., Du, Y., Guo, X., Pan, B., Qin, Z., Zhao, L.: Controllable data generation by deep learning: A review. ACM Computing Surveys **56**(9), 1–38 (2024)
23. Wang, X., Zhang, S., Yuan, H., Qing, Z., Gong, B., Zhang, Y., Shen, Y., Gao, C., Sang, N.: A recipe for scaling up text-to-video generation with text-free videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6572–6582 (2024)
24. Wang, Z., Zhang, L., Wang, L., Zhu, M., Zhang, Z.: Optical flow representation alignment mamba diffusion model for medical video generation. arXiv preprint arXiv:2411.01647 (2024)
25. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
26. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)
27. Yeganeh, Y., Lazuardi, R., Shamseddin, A., Dari, E., Thirani, Y., Navab, N., Farshad, A.: Visage: Video synthesis using action graphs for surgery. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 146–156. Springer (2024)

28. Yu, S., Tack, J., Mo, S., Kim, H., Kim, J., Ha, J.W., Shin, J.: Generating videos with dynamics-aware implicit generative adversarial networks. In: International Conference on Learning Representations (2022)
29. Zhan, X., Pan, X., Liu, Z., Lin, D., Loy, C.C.: Self-supervised learning via conditional motion propagation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1881–1889 (2019)
30. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)