Mission Balance: Generating Under-represented Class Samples using Video Diffusion Models

Danush Kumar Venkatesh[1], Isabel Funke[1,2], Micha Pfeiffer[1], Fiona Kolbinger[3,4], Hanna Maria Schmeiser[4], Marius Distler[4], Jürgen Weitz[4], and Stefanie Speidel[1,2]

[1] Department of Translational Surgical Oncology, NCT/UCC Dresden, a partnership between DKFZ, Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden,HZDR, Germany
[2] The Centre for Tactile Internet with Human-in-the-Loop (CeTI), TUD Dresden
[3] Weldon School of Biomedical Engineering, Purdue University, USA
[4] University Hospital & Faculty of Medicine Carl Gustav Carus, TUD Germany
danushkumar.venkatesh@nct-dresden.de

Abstract. Computer-assisted interventions can improve intraoperative guidance, particularly through deep learning methods that harness the spatiotemporal information in surgical videos. However, the severe data imbalance often found in surgical video datasets hinders the development of high-performing models. In this work, we aim to overcome the data imbalance by synthesizing surgical videos. We propose a unique two-stage, text-conditioned diffusion-based method to generate high-fidelity surgical videos for under-represented classes. Our approach conditions the generation process on text prompts and decouples spatial and temporal modeling by utilizing a 2D latent diffusion model to capture spatial content and then integrating temporal attention layers to ensure temporal consistency. Furthermore, we introduce a rejection sampling strategy to select the most suitable synthetic samples, effectively augmenting existing datasets to address class imbalance. We evaluate our method on two downstream tasks—surgical action recognition and intra-operative event prediction—demonstrating that incorporating synthetic videos from our approach substantially enhances model performance.

Keywords: Video Diffusion · Surgical Data Science · Data Imbalance

1 Introduction

Computer-assisted intervention (CAI) aims to enhance surgical procedures by integrating advanced computational techniques [18]. The rapid growth of deep learning (DL) has further propelled CAI research, enabling applications such as surgical action recognition, phase recognition, and the identification of critical anatomical structures to offer context-aware guidance and decision support during surgery [18]. Notably, these applications can benefit from temporal context by analyzing surgical video sequences rather than individual frames.

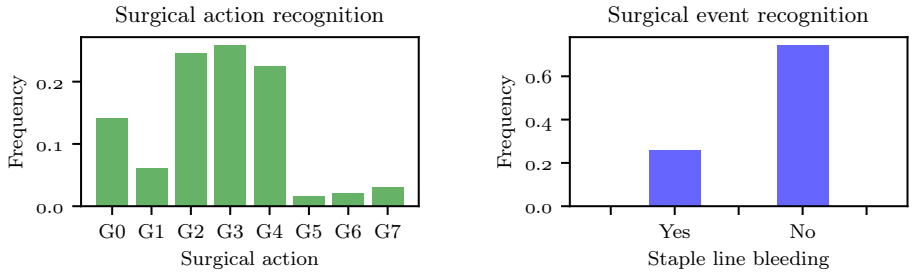

Fig. 1: Data imbalance among classes in surgical datasets. Left: Frequency of different actions (G0-G7) in the SAR-RARP50 [22] dataset. Right: Occurrence of staple line bleeding in distal pancreatectomy.

However, DL methods require large volumes of real, diverse, and annotated surgical video data [18]. Although the availability of public surgical datasets has increased, significant data imbalance remains a critical issue (see Fig. 1), which causes DL methods to become biased toward the majority classes. While oversampling and data augmentation can mitigate this problem, they increase the frequency of under-represented samples but do not address the lack of diversity.

Synthesizing data with generative models offers a promising solution to the lack and imbalance of data [26,7]. In particular, *diffusion models (DMs)* [5] have been applied to generate high-quality surgical images conditioned on tissue texture and shape [27,16]; however, these approaches lack temporal context—a critical factor for generating surgical videos. On the other hand, recent methods for surgical video synthesis either (i) do not condition the generation on class labels, which is needed to control the generation of new samples specifically for under-represented classes [17], (ii) require pre-existing instrument masks to control the synthesis of each video frame [14], (iii) rely on large volumes of data (100–200K frames), which is typically not available for under-represented cases [4], or (iv) use spatio-temporal models with high computational demand [17,4].

In contrast, we propose to leverage a pre-trained 2D latent diffusion model – *Stable Diffusion (SD)* [23] – and extend it into a video diffusion model by adding temporal layers to separately model video dynamics. By conditioning the model on text prompts, we generate videos for specific classes. By separating the spatial and temporal modeling, the training and inference efficiency is improved. In particular, we start with a pre-trained SD model, which we finetune into a common text-conditioned spatial diffusion model for all classes. Based on that, we create the video diffusion model by freezing the spatial layers and adding temporal layers, which are trained with additional class label conditioning to learn class-specific dynamics for each under-represented class.

In addition, we introduce a rejection sampling procedure to select the most suitable synthetic video samples for downstream surgical tasks.

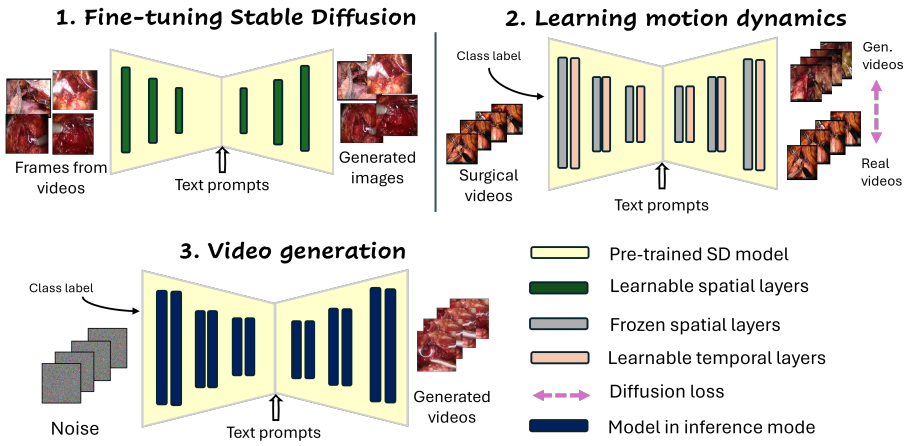

Fig. 2: Overview of the SurV-Gen method for surgical video generation.

We evaluate the abilities of *SurV-Gen*, our proposed surgical video diffusion model, to synthesize videos of under-represented classes on two downstream tasks: ① surgical action recognition on the SAR-RARP50 [22] dataset and ② video-level recognition of staple line bleeding (SLB) during pancreas transection on an in-house dataset. To the best of our knowledge, we are the first to tackle the data imbalance problem in surgical data science using diffusion models for video generation. Our codebase: https://gitlab.com/nct_tso_public/surgvgen.

2 Method

Our video diffusion method, SurV-Gen, is trained in two stages. In stage 1, we fine-tune a pre-trained Stable Diffusion model on video frames paired with text prompts to capture the spatial content. In stage 2, with spatial layers frozen, temporal transformer layers are inserted and trained to model dynamics, with text conditioning applied throughout. During inference, SurV-Gen generates class-specific videos from text prompts, which are refined via a sample selection pipeline. An overview of the method is shown in Fig. 2.

2.1 Stage 1: Fine-tuning Stable Diffusion

Stable Diffusion (SD) [23] was opted as the base text-to-image model in this work because it is open-sourced and produces high-quality surgical images [19,27,16]. In SD, the diffusion process takes place in latent space by encoding the image x_0 to z_0 with an encoder $\mathcal{E}(x_0)$. During the forward diffusion process, z_0 is perturbed to z_t via

$$z_t = \sqrt{1-\beta_t}\, z_{t-1} + \sqrt{\beta_t}\, \boldsymbol{\epsilon}_{t-1} = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \boldsymbol{\epsilon}_0, \quad 1 \leq t \leq T, \qquad (1)$$

where $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, and the predefined β_t controls the noise strength at each step t. A denoising network $\boldsymbol{\epsilon}_\theta(\cdot)$ is trained to reverse this process by predicting the noise that was added by minimizing loss L:

$$L = \mathbb{E}_{\mathcal{E}(x_0),\, y,\, \boldsymbol{\epsilon} \sim \mathcal{N}(0,I),\, t} \left[\left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\big(z_t, t, y\big) \right\|_2^2 \right], \tag{2}$$

where y is the text prompt associated with image x_0. $\boldsymbol{\epsilon}_\theta(\cdot)$ is implemented as a U-Net [24]. We fine-tune the SD model (pre-trained on natural images) in this stage using images extracted from the surgical videos together with text prompts that describe the corresponding class label.

2.2 Stage 2: Learning motion dynamics

In this stage, we extend the fine-tuned SD model to train directly on surgical video sequences consisting of 16 frames. We adopt the method proposed in recent works [12,3,8] by incorporating temporal *transformer* blocks [25] into the SD model, placing them after each spatial layer. Since this stage focuses solely on modeling temporal dynamics, the spatial layers are kept frozen.

Given a 5D video tensor $v \in \mathbb{R}^{b \times c \times f \times h \times w}$, where b is the batch size, c and f are the number of channels and frames respectively, and h and w are the spatial dimensions, the frozen spatial layers process v frame-wise by reshaping to $(bf) \times c \times h \times w$. In the temporal layers, the sequences at each spatial location are processed independently by reshaping v to $(bhw) \times f \times c$. Here, sequence $v_{\text{in}} \in \mathbb{R}^{(bhw) \times f \times c}$ is projected and – after adding sinusoidal positional encoding – processed using *self-attention* via:

$$v_{\text{out}} = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{c}}\right) V, \tag{3}$$

with $Q = v_{\text{in}} W_Q$, $K = v_{\text{in}} W_K$, and $V = v_{\text{in}} W_V$ being query, key and value vectors. In this way, each generated frame incorporates information from other frames in the video clip, capturing the motion dynamics over time. The self-attention is followed by a multi-layer perceptron. Following [28], we initialize the output projection layers of the temporal blocks to zero and include a residual connection. In addition to text prompts, we add an embedding of the class label during training to serve as an additional conditioning signal and train with the same loss as in Eq. 2.

2.3 Selection of generated data

A common approach for leveraging synthetically generated data in downstream tasks is to combine it with real datasets to improve performance [7,17]. However, prior works have shown that adding synthetic data can sometimes have adverse effects [2,1]. To address this, we introduce a *rejection sampling* (RS) procedure to select the most suitable synthetic videos from the pool of generated candidates.

Table 1: Image-level quality comparison of generated video frames for SLB recognition task after RS. Inf. time denotes the time for generating a video.

Method	Trainable params	Inf. time	CFID (↓)	Density(↑)	Coverage(↑)	CMMD(↓)
LVDM [11]	548M	43.3s	$185.45_{\pm 4.30}$	$0.20_{\pm 0.007}$	$0.40_{\pm 0.006}$	$3.98_{\pm 0.10}$
Endora [17]	675M	15.3s	$\underline{117.72_{\pm 2.93}}$	$\underline{0.81_{\pm 0.004}}$	$\mathbf{0.71_{\pm 0.001}}$	$\underline{2.78_{\pm 0.001}}$
SurV-Gen	435 M	6.55s	$\mathbf{108.30_{\pm 1.22}}$	$\mathbf{0.90_{\pm 0.002}}$	$\underline{0.65_{\pm 0.003}}$	$\mathbf{2.25_{\pm 0.01}}$

Specifically, we train a discriminative model to predict class labels on the available real datasets and then use this model to evaluate the synthetic videos. A synthetic video that was conditionally generated for class label l is only retained if the model's top-k predictions contain l.

3 Evaluation

3.1 Experiments

Downstream tasks For downstream evaluation, we use two challenging tasks on small-scale intraoperative video datasets with imbalanced class distribution.

Action recognition (Task ①): The task is to recognize which surgical action is performed at any time t in a video. Here, we use the SAR-RARP50[5] dataset [22], which consists of 50 videos of robot-assisted suturing with an average duration of about 5 min. We split them into 35, 5, and 10 videos for training, validation, and testing. Seven different actions plus a background class are defined, including the under-represented actions "Picking-up the needle" (G1), "Cutting the suture" (G6), and "Returning/dropping the needle" (G7) (see Fig. 1). Action G5 ("Tying a knot") is not considered further because it occurs only in one test video. As DL model, we train X3D [6], a 3D CNN for video recognition, to recognize the action at time t by classifying the 16-frame video clip that is centered around time t. To measure model performance, we compute the average video-wise Jaccard index for each action.

SLB recognition (Task ②): The task is to recognize whether or not staple line bleeding (SLB) occurs in a video of the pancreas transection phase during distal pancreatectomy. During pancreatic stapler transection, SLB represents a visually recognizable event that has been associated with postoperative pancreatic fistula [29]. We use an internal dataset of 39 videos, each 3 min long on average, where SLB occurred in 10 cases. The analysis of these deidentified data was approved by the local institutional review board, and informed patient consent was waived. We split the data into 25, 4, and 10 videos for training,

[5] Train set: https://doi.org/10.5522/04/24932529.v1, test set: https://doi.org/10.5522/04/24932499.v1, license: CC BY-NC-SA 4.0

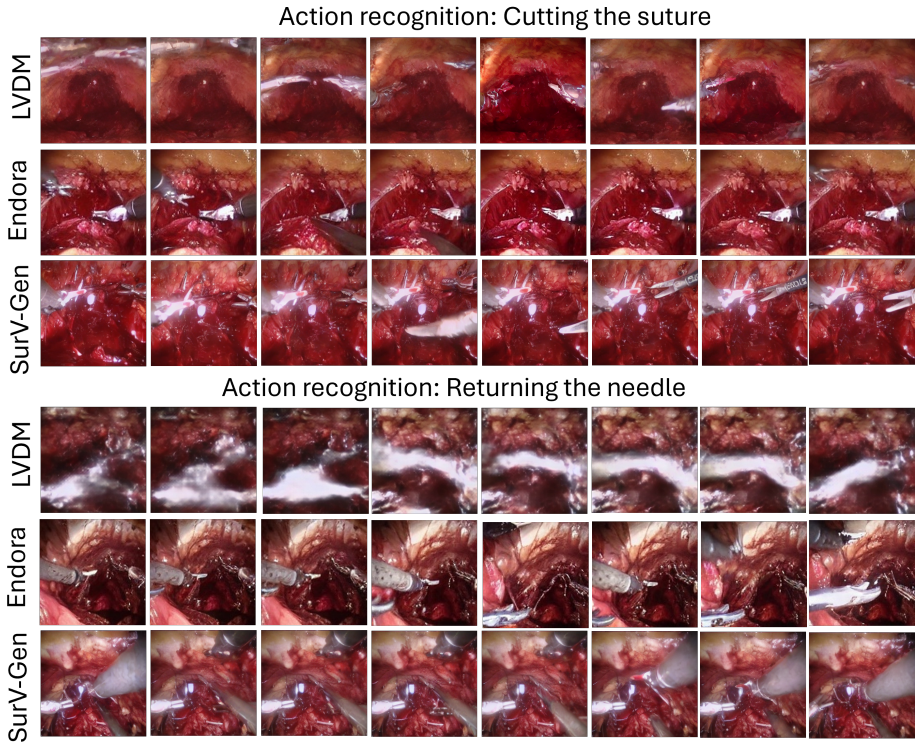

Fig. 3: Qualitative comparison of generated video frames. In our approach (SurV-Gen), the scissors are clearly visible during suture cutting (row 3), whereas in other methods, the scissors appear only partially. Similarly, during the needle return, the tools are consistently generated (row 6).

validation, and testing. As DL model, we train a ResNet-LSTM on the videos, where the ResNet [10] extracts visual features from individual video frames and the LSTM [13] aggregates these features over time. A linear classifier recognizes the occurrence of SLB based on the final LSTM state. To measure model performance, we compute balanced accuracy and F_1 score.

Notably, we want to investigate the impact of adding synthetic training data only for the under-represented classes and therefore chose solid baselines for the downstream tasks instead of more sophisticated models. We repeat all experiments on the downstream tasks three times and average the evaluation metrics over these runs.

Baselines We selected Endora [17] and LVDM [11] as the video diffusion baselines. Endora is a fully transformer-based diffusion model that jointly models spatio-temporal components whereas LVDM uses a 3D U-Net architecture for a 3D latent video diffusion approach.

Video generation On task ①, we trained a single SD model for all the selected classes (G1, G6, G7), after which separate spatio-temporal models were trained for the three classes using label conditioning. Text prompts were constructed based on the class label as "action recognition task of <gesture>" (task ①) and "a complication of staple line bleeding" (task ②) for both training and inference. We train the diffusion models on the respective (patient-specific) training sets of the downstream tasks and generate videos of 16 frames with resolutions of 256×256 pixels for SurV-Gen and LVDM, and 128×128 pixels for the Endora model.

Downstream evaluation For evaluating any of the video generation methods, we generated videos for each under-represented class and performed rejection sampling (RS) using a lightweight ResNet3D classifier [9] with thresholds of $k = 3$ (task ①) and $k = 1$ (task ②). With SurV-Gen, we further conducted an ablation study without rejection sampling.

3.2 Results and discussion

We evaluate the generated videos for both image quality and – by measuring their benefit for downstream tasks – realism, diversity, and temporal consistency.

Image quality We evaluate the fidelity and diversity of the generated frames using (i) the CFID metric [21] to quantify realism, (ii) the CMMD score [15] to quantify unbiased image quality, which is particularly effective for smaller datasets, and (iii) density and coverage metrics [20] to assess image quality and diversity.

The results in Tab. 1 indicate that our SurV-Gen model synthesizes high-quality surgical video frames compared to baseline methods. This finding is further supported by the qualitative results in Fig. 3, where SurV-Gen effectively generates surgical instruments, such as scissors and graspers, corresponding to specific gestures. In contrast, the baseline models either omitted these tools or produced poorly defined scenes, which we attribute to their dependence on larger datasets for learning temporal dynamics. Moreover, the SurV-Gen model features fewer trainable parameters and requires less inference time than the baselines. Additional generated videos are provided as supplementary material.

Downstream evaluation Tab. 2 presents the results for task ①. Extending the training set with synthetically generated videos (of under-represneted classes) enhanced performance across all baseline models for most classes. Notably, our SurV-Gen model with rejection sampling achieved improvements of 12 and 14 points for classes G6 and G7, respectively, relative to using only real videos. These results underscore the high quality of the synthesized samples and the effectiveness of the rejection sampling strategy. A modest improvement was observed for gesture G1 only with the Endora model. Further exploration of model architectures with the addition of more synthetic samples can improve performance for all the classes, which we leave for future work.

The results for SLB recognition is shown in Tab. 3. The addition of synthetic videos to the real training data improved performance for both the SurV-Gen

Table 2: Results on task ①. We report the Jaccard index for each class, including under-represented classes. (RS: rejection sampling)

Training data	RS	G0	G1	G2	G3	G4	G5*	G6	G7	Avg.
Only Real	-	$0.54_{\pm 0.002}$	$0.31_{\pm 0.004}$	$0.62_{\pm 0.004}$	$0.75_{\pm 0.03}$	$0.58_{\pm 0.02}$	$0.19_{\pm 0.10}$	$0.11_{\pm 0.03}$	$0.23_{\pm 0.06}$	$0.44_{\pm 0.11}$
Real + LVDM	✓	$0.47_{\pm 0.01}$	$0.30_{\pm 0.03}$	$0.60_{\pm 0.01}$	$\mathbf{0.77_{\pm 0.009}}$	$0.59_{\pm 0.01}$	$\underline{0.25_{\pm 0.13}}$	$0.20_{\pm 0.01}$	$0.23_{\pm 0.04}$	$0.44_{\pm 0.07}$
Real + Endora	✓	$\underline{0.54_{\pm 0.03}}$	$\mathbf{0.34_{\pm 0.06}}$	$\underline{0.61_{\pm 0.01}}$	$\mathbf{0.77_{\pm 0.01}}$	$\underline{0.60_{\pm 0.01}}$	$\mathbf{0.26_{\pm 0.13}}$	$\underline{0.20_{\pm 0.01}}$	$\underline{0.31_{\pm 0.02}}$	$\underline{0.48_{\pm 0.09}}$
Real + SurV-Gen	✓	$\mathbf{0.55_{\pm 0.01}}$	$\underline{0.31_{\pm 0.01}}$	$\mathbf{0.62_{\pm 0.005}}$	$\underline{0.76_{\pm 0.001}}$	$\mathbf{0.63_{\pm 0.02}}$	$0.18_{\pm 0.14}$	$\mathbf{0.23_{\pm 0.03}}$	$\mathbf{0.37_{\pm 0.01}}$	$\mathbf{0.50_{\pm 0.10}}$
Real + SurV-Gen	✗	$0.52_{\pm 0.10}$	$0.25_{\pm 0.05}$	$0.60_{\pm 0.01}$	$0.75_{\pm 0.002}$	$0.60_{\pm 0.01}$	$0.10_{\pm 0.09}$	$0.18_{\pm 0.009}$	$0.24_{\pm 0.003}$	$0.44_{\pm 0.14}$

*G5 occurs only in one test video.

Table 3: Results on SLB recognition (task ②). (RS: rejection sampling)

Training data	RS	Bal. Acc (↑)	F_1 (↑)
Only Real	-	$0.74_{\pm 0.08}$	$0.71_{\pm 0.01}$
Real + LVDM	✓	$0.72_{\pm 0.10}$	$0.69_{\pm 0.03}$
Real + Endora	✓	$\underline{0.76_{\pm 0.01}}$	$\underline{0.75_{\pm 0.10}}$
Real + SurV-Gen	✓	$\mathbf{0.81_{\pm 0.05}}$	$\mathbf{0.78_{\pm 0.04}}$
Real + SurV-Gen	✗	$0.74_{\pm 0.08}$	$0.70_{\pm 0.10}$

and Endora models compared to relying solely on real videos. Conversely, the LVDM model did not show any performance improvements.

The ablation on rejection sampling underscores the critical role of synthetic sample selection on both downstream tasks. For example on task ②, SurV-Gen with RS achieves a 7-point boost whereas the gains without rejection sampling are only modest.

4 Conclusion

In this work, we introduce SurV-Gen, a light-weight two-stage diffusion framework synthesizing high-fidelity surgical videos for underrepresented classes, thereby addressing the data imbalance in surgical datasets. By separating spatial and

temporal modeling, our approach can directly reuse pre-trained weights of large-scale image diffusion models, which efficiently helps to learn video generation with limited examples. For the first time, we show that synthesizing additional training videos for under-represented classes helps to improve the performance of video recognition models on two challenging surgical downstream tasks with data imbalance. Here, selecting the most suitable synthetic videos using rejection sampling proved to be a crucial step. Future work may extend this framework to model additional under-represented classes by incorporating conditional signals, e.g., segmentation maps, from surgical simulations. Furthermore, a more detailed analysis of the contribution of individual synthetic samples to downstream performance could provide valuable insights for further improving video generation.

Acknowledgments. This work is partly supported by BMBF (Federal Ministry of Education and Research) in DAAD project 57616814 (SECAI, School of Embedded Composite AI). Also partially funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 –Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

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

References

1. Alaa, A., Van Breugel, B., Saveliev, E.S., van der Schaar, M.: How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models. In: International Conference on Machine Learning. pp. 290–306. PMLR (2022)
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves ImageNet classification. TMLR (2023)
3. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22563–22575 (2023)
4. 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)
5. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. Advances in Neural Information Processing Systems **34**, 8780–8794 (2021)
6. Feichtenhofer, C.: X3D: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020)
7. Frisch, Y., Fuchs, M., Sanner, A., Ucar, F.A., Frenzel, M., Wasielica-Poslednik, J., Gericke, A., Wagner, F.M., Dratsch, T., Mukhopadhyay, A.: Synthesising rare cataract surgery samples with guided diffusion models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 354–364. Springer (2023)

8. Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., Dai, B.: AnimateDiff: Animate your personalized text-to-image diffusion models without specific tuning. International Conference on Learning Representations (2024)
9. Hara, K., Kataoka, H., Satoh, Y.: Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6546–6555 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
11. 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)
12. Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. NeruIPS **35**, 8633–8646 (2022)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. 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)
15. Jayasumana, S., Ramalingam, S., Veit, A., Glasner, D., Chakrabarti, A., Kumar, S.: Rethinking FID: Towards a better evaluation metric for image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9307–9315 (2024)
16. Kaleta, J., Dall'Alba, D., Płotka, S., Korzeniowski, P.: Minimal data requirement for realistic endoscopic image generation with stable diffusion. International Journal of Computer Assisted Radiology and Surgery **19**(3), 531–539 (2024)
17. 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)
18. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., Collins, T., Malpani, A., Fallert, J., Feussner, H., Giannarou, S., Mascagni, P., et al.: Surgical data science–from concepts toward clinical translation. Medical image analysis **76**, 102306 (2022)
19. Martyniak, S., Kaleta, J., Dall'Alba, D., Naskręt, M., Płotka, S., Korzeniowski, P.: Simuscope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 4268–4278 (February 2025)
20. Naeem, M.F., Oh, S.J., Uh, Y., Choi, Y., Yoo, J.: Reliable fidelity and diversity metrics for generative models. In: ICML. pp. 7176–7185. PMLR (2020)
21. Parmar, G., Zhang, R., Zhu, J.Y.: On aliased resizing and surprising subtleties in GAN evaluation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11410–11420 (2022)
22. Psychogyios, D., Colleoni, E., Van Amsterdam, B., Li, C.Y., Huang, S.Y., Li, Y., Jia, F., Zou, B., Wang, G., Liu, Y., et al.: SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. arXiv preprint arXiv:2401.00496 (2023)
23. 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)

24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
26. Venkatesh, D.K., Rivoir, D., Pfeiffer, M., Kolbinger, F., Distler, M., Weitz, J., Speidel, S.: Exploring semantic consistency in unpaired image translation to generate data for surgical applications. International Journal of Computer Assisted Radiology and Surgery pp. 1–9 (2024)
27. Venkatesh, D.K., Rivoir, D., Pfeiffer, M., Kolbinger, F., Speidel, S.: Data augmentation for surgical scene segmentation with anatomy-aware diffusion models. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 2280–2290 (February 2025)
28. 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)
29. Zimmitti, G., La Mendola, R., Manzoni, A., Sega, V., Malerba, V., Treppiedi, E., Codignola, C., Monfardini, L., Garatti, M., Rosso, E.: Investigation of intraoperative factors associated with postoperative pancreatic fistula following laparoscopic left pancreatectomy with stapled closure: a video review-based analysis. Surgical Endoscopy **35**, 941–954 (2021)