SimuScope: Realistic Endoscopic Synthetic Dataset Generation through Surgical Simulation and Diffusion Models

Sabina Martyniak[1] Joanna Kaleta[1,2] Diego Dall'Alba[3] Michał Naskręt[1]
Szymon Płotka[1,4] Przemysław Korzeniowski[1]

Sano Centre for Computational Medicine, Poland[1] Warsaw University of Technology, Poland[2]
University of Verona, Italy[3] University of Warsaw, Poland[4]

Abstract

Computer-assisted surgical (CAS) systems enhance surgical execution and outcomes by providing advanced support to surgeons. These systems often rely on deep learning models trained on complex, challenging-to-annotate data. While synthetic data generation can address these challenges, enhancing the realism of such data is crucial. This work introduces a multi-stage pipeline for generating realistic synthetic data, featuring a fully-fledged surgical simulator that automatically produces all necessary annotations for modern CAS systems. This simulator generates a wide set of annotations that surpass those available in public synthetic datasets. Additionally, it offers a more complex and realistic simulation of surgical interactions, including the dynamics between surgical instruments and deformable anatomical environments, outperforming existing approaches. To further bridge the visual gap between synthetic and real data, we propose a lightweight and flexible image-to-image translation method based on Stable Diffusion (SD) and Low-Rank Adaptation (LoRA). This method leverages a limited amount of annotated data, enables efficient training, and maintains the integrity of annotations generated by our simulator. The proposed pipeline is experimentally validated and can translate synthetic images into images with real-world characteristics, which can generalize to real-world context, thereby improving both training and CAS guidance. The code and the dataset are available at https://github.com/SanoScience/SimuScope.

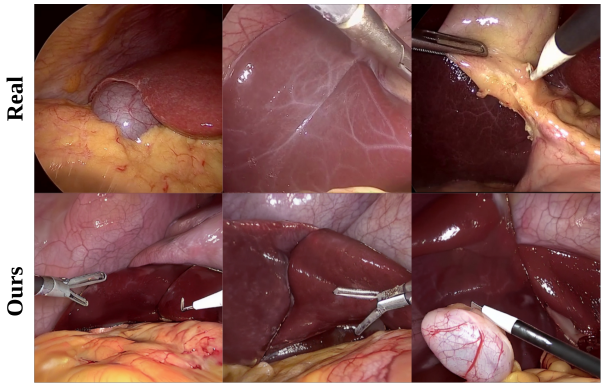

Figure 1. Comparison of real images and images generated by **SimuScope**. The real images used in this comparison are sourced from the CholecT45 dataset [33], which comprises authentic surgical footage. Both the real and generated images exhibit comparable coloration and textural details, posing a challenge in distinguishing between them at a glance. This similarity underscores the fidelity and realism achieved by SimuScope in simulating surgical scenarios.

1. Introduction

Computer Assisted Surgery (CAS) is a rapidly evolving field that aims to enhance surgical procedures by providing advanced technological support to surgeons [51]. By integrating sophisticated computational tools, CAS systems can improve the precision, safety, and outcomes of surgeries [36, 37, 43, 62]. One of the most extensively studied procedures within this domain is cholecystectomy, a surgical procedure for the removal of the gallbladder [48]. Despite its prevalence in surgery, cholecystectomy can lead to serious complications such as bile duct injury, which underscores the need for improved surgical support systems [48]. CAS systems can leverage deep learning (DL) methods to perform various supporting tasks. These include the segmentation of anatomical structures and surgical instruments [1, 27], as well as the temporal modeling [13] of the procedure at different levels of detail. These DL systems' effectiveness heavily depends on large volumes of annotated data. However, obtaining such data is challenging due to the need for detailed and time-consuming annotations, which can only be performed by expert personnel. This requirement poses a significant limitation on the development and

scalability of CAS systems.

To address this limitation, synthetic data generated through virtual simulators is proposed. These simulators automatically generate the necessary annotations, reducing the dependency on manual annotation. However, existing simulators often lack realism and provide limited interactions between surgical tools and tissues. Additionally, the annotations generated by these simulators are typically restricted to specific tasks, such as semantic segmentation of the scene.

In this work, we propose an advanced virtual simulator that overcomes these limitations. Our simulator offers a highly realistic biomechanical simulation of tissues and rich interactions, such as tissue grasping, tearing, cutting, thermocoagulation, and vessel clipping with various types of surgical instruments, including both laparoscopic and robotic ones. This enhanced realism is crucial for training DL models to support the execution of surgical procedures effectively. Furthermore, our simulator generates an extensive set of annotations, including pixel-level semantic segmentation, depth and normal maps, optical flow, as well as temporal and high-level text information, such as surgical action-triplets [33] and instrument poses in both 2D and 3D. These detailed annotations provide a more comprehensive dataset for training DL models, improving their performance and reliability.

To further enhance the visual realism of the main endoscopic RGB image rendered by our simulator, we propose an image-to-image (img2img) translation approach. It is based on SD combined with ControlNet, which allows us to preserve the automatically generated annotations while significantly improving the visual quality of the rendered images. Unlike traditional methods based on Generative Adversarial Networks (GANs) [5], our approach is lightweight and efficient, utilizing multiple LoRA modules [9]. This enables us to achieve high-quality results with limited annotated data and efficient training processes. The main contributions of our work are as follows:

1. We introduce a high-fidelity virtual simulator that provides a realistic biomechanical simulation of tissues and a rich set of interactions between surgical instruments and tissues, such as grasping, tearing, cutting, clipping, and thermocoagulation. To the best of our knowledge, this is the first work that supports such a wide range of surgical interactions, which is indispensable for training effective DL models.

2. We propose a novel image-to-image translation approach that improves the visual realism of synthetic data generated by the simulator while preserving the integrity of all annotations. This approach is both lightweight and efficient, making it a practical solution for enhancing synthetic data quality.

3. The simulator generates a comprehensive set of annotations, including detailed temporal information, which significantly enhances the training dataset for DL models.

2. Related work

Recently, several approaches have been proposed for generating synthetic data with realistic characteristics, either for specific surgical procedures or general anatomical structures (e.g., [21, 30]). The combination of synthetic images and real segmentation maps is used to train GANs for image analysis and surgical applications, as seen in [25, 34, 42, 44].

While GAN-based approaches show potential, they have limitations, such as early convergence of discriminators and instability of adversarial training, leading to mode collapse and reduced diversity in generated data [10]. Diffusion models (DMs) [8] emerge as a promising alternative, surpassing GANs in computer vision tasks.

Diffusion models are widely adopted in the medical domain with several applications. One is Image-to-Image Translation, such as CT-to-MRI translation [20, 67]. To address the lengthy training times required for diffusion models, some works focus on zero-shot approaches [24, 53, 54]. Diffusion models provide powerful representations useful for image understanding tasks, including segmentation [2, 6, 18, 38, 56], classification [60], and anomaly detection [55, 57]. Other applications include image reconstruction [26] and image registration [17].

Data generation is one of the primary objectives of diffusion models, which are widely applied in various styles. Generated data targets different aspects, such as temporal consistency [19], data debiasing [46], and multimodal generation [12, 63].

Although multiple generative works exist in the medical and surgical fields, a gap remains in surgical data generation, particularly for fully labeled simulator-based data with accurate and detailed instrument-tissue interactions [16, 17]. Physically accurate data generation with rich labeling, such as depth, normals, and triplets, is potentially useful for robotic tasks. The closest generative works focusing on laparoscopic cholecystectomy include two studies [35, 40]—one on photo collection and one for videos with temporal consistency, both based on GANs. These works require large datasets and use simple simulators with long training times. Although they release a large open-source simulation dataset, it has limitations, including a very simple simulator that lacks tool-tissue interactions and requires long training times.

In [50], semantic consistency in unpaired image translation to generate data for surgical applications is investigated, providing insightful analysis and proposing techniques to enhance GAN performance. Based on simulation

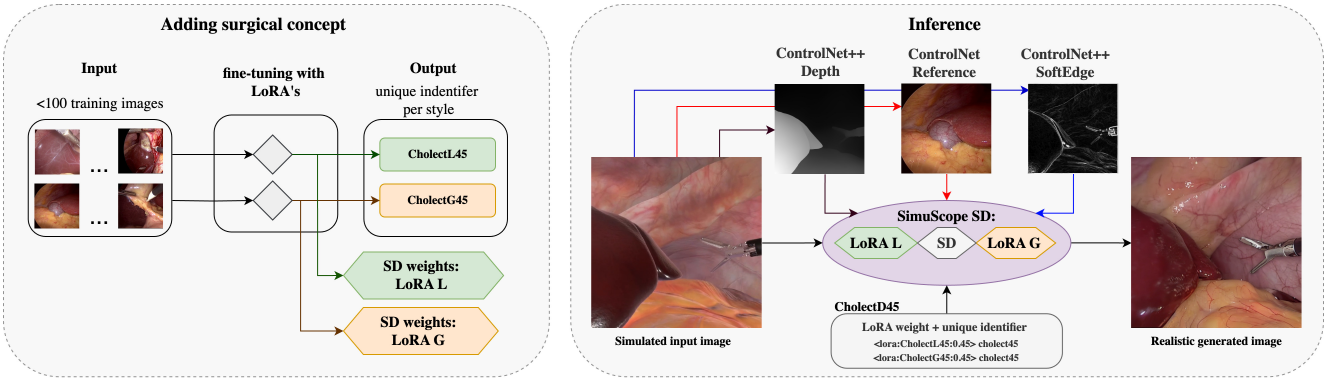

Figure 2. An overview of SimuScope fine-tuning and inference stage. SD model undergoes fine-tuning using LoRA, a framework that associates a unique LoRA identifier and weight with the newly integrated cholect45 style. During inference, the enhanced SimuScope leverages three ControlNet/ControlNet++ models for comprehensive conditioning. The raw input sample, along with the prompts 'lora:CholectL45:0.45 cholect45' and 'lora:CholectG45:0.45 cholect45,' is fed into SimuScope. The SoftEdge ControlNet++ processes edges predicted by HED, the Depth ControlNet++ handles depth detected by MiDaS, and the Reference ControlNet utilizes additional real input sample as a reference. This multi-model integration enhances SimuScope's capability to generate realistic surgical simulations enriched with detailed texture, edge, depth, and reference data.

data, [16] generates large fully labeled realistic endoscopic images, addressing the minimal data requirement. However, this work also relies on a very simple simulator lacking tool-tissue interactions.

3. Methods

Our work involves a three-step process as follows. We generate synthetic images of the cholecystectomy procedure from our simulator. We fine-tune the Stable Diffusion (SD) model on a small subset of real images using two LoRAs. Finally, in the sim-to-real phase, we perform inference using the fine-tuned SD model in image-to-image mode, enhanced with three versions of the ControlNet architecture, to ensure consistency between the labels and the generated images. An overview of our method is presented in Figure 2.

3.1. Simulation system

The simulation software runs on a desktop computer equipped with an Intel Core i9-12900K, 64GB RAM, and NVIDIA RTX 3080 GPU. We use two 3D Systems Touch devices for haptic input. The development environment includes the Unity3D game engine integrated with our custom-built surgical simulation framework.

A multi-threaded implementation of the Extended Position-Based Dynamics (XPBD) physics solver, programmed in C/C++, is employed for real-time soft tissue simulation. Recent research (e.g., [28, 29, 32]) establishes XPBD as a competitive alternative to more complex methods for simulating non-linear tissue behavior due to its accuracy, stability, speed, and ease of implementation. The local nature of the non-linear Gauss-Seidel solver within XPBD avoids the limitations associated with global, matrix-based solvers. This enables efficient and accurate implicit simulation of arbitrary elastic and dissipative energy potentials, leading to robust handling of equality and inequality constraints. Additionally, XPBD provides constraint force estimates crucial for accurate haptic feedback calculations in surgical simulation.

To create a virtual model of the liver, gallbladder with cystic duct and artery, and surrounding tissues, we use volumetric tetrahedral meshes consisting of about 50,000 elements in total. A Neo-Hookean constitutive model [28] is employed to simulate the near incompressibility of soft tissue. The model can conserve volume more than corotational finite element or Saint-Venant-Kirchhoff models and can recover from inverted element configurations (i.e., flipped tetrahedrons). The following equation presents the energy-based formulation of the Neo-Hookean model adopted in this simulator:

$$\begin{aligned}\Psi_{Neo} &= \Psi_H + \Psi_D = \\ &= \frac{\lambda}{2}\left(\det(F)-1\right)^2 + \frac{\mu}{2}\left(\mathrm{tr}(F^T F)-3\right),\end{aligned} \tag{1}$$

where F is a 3×3 deformation gradient matrix, λ and μ are the Lamé parameters, Ψ_H represents the hydrostatic energy component resisting volume changes, and Ψ_D signifies the deviatoric energy component resisting distortion.

The proposed simulator leverages XPBD's iterative constrained optimization approach, which only requires the computation of first-order gradients. This eliminates the need for complex calculations involving Hessian matrices, eigenvalue decomposition, and sophisticated linear

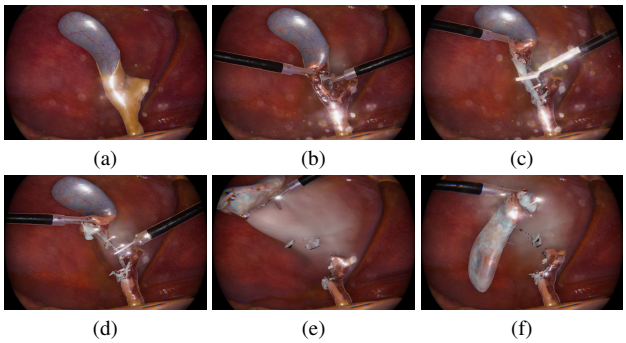

Figure 3. An overview of the stages of a virtual cholecystectomy: (a) the start of the procedure, showing the initial setup with surgical instruments inserted into the abdominal cavity; (b) the dissection of Calot's triangle using a grasper and diathermy hook; (c) the clipping of the cystic duct and artery with a clipping tool; (d) the cutting of the cystic duct and artery with scissors; (e) the dissection of the gallbladder from the liver bed using a hook; and (f) the gallbladder fully dissected and ready for removal from the abdominal cavity.

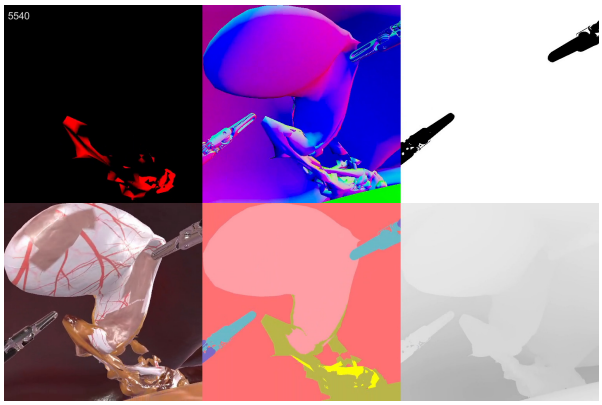

Figure 4. The output from the simulator shows partially dissected Calot's triangle. Top row: blood map - marking the bleeding tissue, normal map - providing detailed surface orientation information to enhance the realism of the simulation, and tools mask - highlighting the specific surgical instruments in use. Bottom row: color - depicting the visual appearance of the surgical scene, segmentation - categorizing different anatomical structures and tools for precise identification, and depth maps - offering information on the distance of objects within the scene to facilitate accurate spatial understanding.

solvers, which are characteristic of Newton-method-based approaches [28].

The compliance parameters governing both hydrostatic and deviatoric constraints are visually tuned to approximate the behavior of real anatomical structures. This tuning process is guided by the feedback of experienced surgeons interacting with the virtual environment. Despite the computational demands, refresh rates up to 1.5 kHz are achieved, allowing for small simulation time steps (0.75–1.0 ms), which are crucial for accurate haptic interaction.

3.2. Virtual Cholecystectomy Surgery

We chose cholecystectomy for its prevalence in general surgery. The simulation starts with surgical instruments inserted into the abdominal cavity inflated with carbon dioxide gas. Simulation visualizes the liver, gallbladder, cystic duct and artery, which are initially covered with fatty connective tissue (Figure 3(a)). The first stage of the procedure is to use grasper and diathermy hook to dissect the hepatocystic triangle (Figure 3(b)). Next, the operator needs to establish the critical view of safety, which allows to safely clip with the clipping tool (Figure 3(c)) and cut cystic duct and artery with scissors (Figure 3(d)). Finally, the gallbladder is separated from the liver bed using the hook (Figure 3(e) and Figure 3(f)). At this development stage, the simulator does not provide support for irrigation and removal of the gallbladder from the abdominal cavity using a specimen bag.

The simulator, in addition to the endoscopic RGB image, outputs a range of corresponding ground-truth image data, such as depth and normal maps, optical flow, tool masks, semantic segmentation, and procedure-specific masks, including tissue bleeding, damage, and coagulation (Figure 4). It also provides 2D/3D instrument poses and surgical action triplets (e.g., 'grasper retracts gallbladder,' 'hook coagulates cystic duct') [33].

3.3. Fine-tuning with LoRAs

LoRA [9] is an innovative method for efficiently fine-tuning large language models. Instead of retraining all the model's parameters, LoRA introduces trainable rank decomposition matrices into each layer of the model's architecture while keeping the pre-trained model weights frozen. This approach significantly reduces the number of trainable parameters for downstream tasks, leading to a substantial decrease in GPU memory requirements and overall costs, without compromising model performance. LoRA is primarily applied to the attention blocks of Transformers in Large Language Models [49], where researchers find that LoRA fine-tuning provides similar quality to full model fine-tuning while being faster and requiring less computational resources.

For instance, fine-tuning an SD model can be achieved using DreamBooth (DB) in LC-SD [41], a method that adjusts the entire model to align with a particular concept or style. Although DB yields impressive outcomes [15, 16], it has a major drawback: the size of the fine-tuned model. Since DB updates the entire model, the resulting checkpoint can be quite large (around 2 to 7 GB) and demands substantial GPU resources for training. In contrast, a LoRA adapter requires significantly less GPU power, yet its inferences are

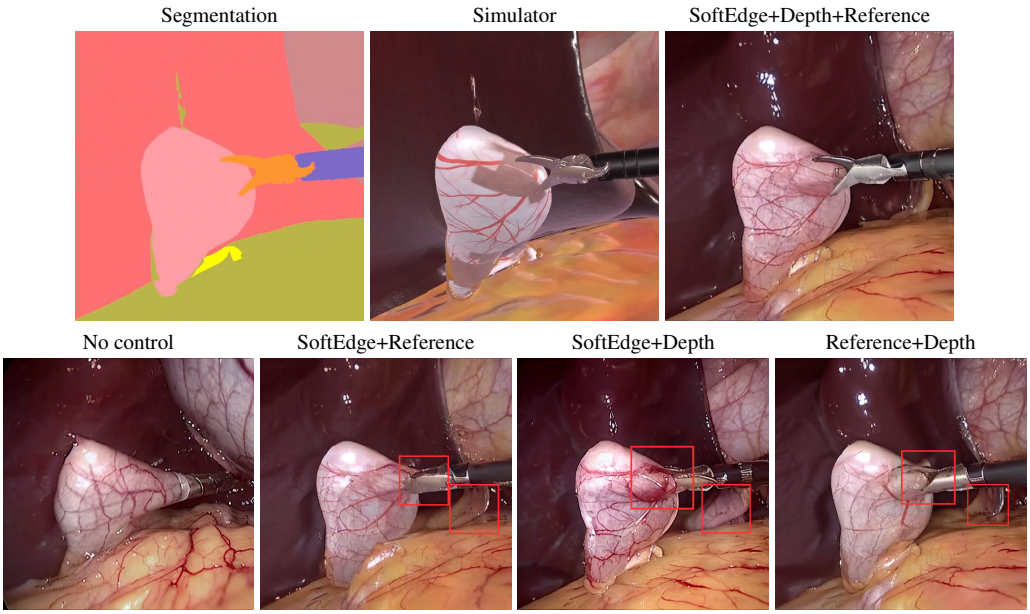

Figure 5. Visual comparison of sample images generated using different types of controls. Without any control, the overall image consistency is degraded, resulting in inconsistent textures and inaccurate color representation.

still comparable to those of a DB fine-tuned checkpoint. For comparison, a model fine-tuned using DB occupies about 2GB [16] while LoRA takes up 0.10 GB, which reduces the required space by over 20 times. This allows for a significant reduction in file sizes while maintaining comparable performance, resulting in lower GPU resource requirements in our models.

We fine-tune SD model using two LoRAs: *CholectG45*, which focuses on accurately representing the gallbladder, and *CholectL45*, which specializes in generating a detailed liver model and surrounding tissues. Together, they ensure anatomical consistency and highly detailed textures.

3.4. Inference with ControlNet

Realistic tissues are generated from simulation scenes using image-to-image inference with LoRA adapters. By combining *CholectG45* and *CholectL45*, introduced in the previous section, we create a new style called *CholectD45* (see Figure 2).

To improve the precision of text-to-image diffusion models, ControlNet [64] introduces image-based conditioning controls. ControlNet++ [22] extends this by optimizing alignment between input conditions and generated images using a pre-trained discriminative reward model for cycle consistency, unlike methods relying on latent diffusion denoising. This enhances controllability across various conditions.

Alternative control tools. Various methods impose control on diffusion models. Prompt engineering [23, 61, 65] is useful for general tasks but struggles with enforcing constraints in specialized domains like surgery. T2I-Adapter [31], conceptually most similar to ControlNet, offers composability and generalization but it is not as widely adopted as ControlNet. Cross-attention constraints [4, 14, 58] and instance-based controllable generation [52, 66] provide regulated generation but lack the precision needed for surgical applications. Compared to the mentioned works, ControlNet and ControlNet++ are widely adopted, offering precise control from various inputs applicable to unusual data domains.

Applied types of control. In our work, we utilize both ControlNet [64] and ControlNet++ [22]. SoftEdge and Depth ControlNet++ models play a crucial role in preserving details and improving shapes and boundaries in the output images. The SoftEdge ControlNet++ model focuses on generating natural tissue appearances by controlling detail through soft edges. We use the HED preprocessor [59] to detect edges from the input sample, which are then fed into the SoftEdge ControlNet++. Depth ControlNet++ leverages depth information obtained from the MiDaS model [39] to further enhance image realism and structural accuracy. Reference ControlNet [1] enhances control and customization in image generation by enabling users to guide outputs based on a reference image. The Reference model establishes a direct link between the reference image and the model's attention layers, allowing focused replication of specific elements from the reference. It serves as a valuable tool for generating images that closely resemble or are inspired

[1] https://github.com/Mikubill/sd-webui-controlnet/discussions/1236

by the reference, enabling greater precision and alignment with user intent. To achieve the desired balance, we utilize stronger Depth and Reference controls in combination with a weaker SoftEdge control. Removing only one control diminishes the quality of the generated images (see Figure 5). To generate data on a large scale while maintaining reasonable inference time and acceptable image quality, we limit the denoising steps to 20. The denoising strength and CFG scale were determined experimentally, as in [16], with values listed in Supplementary Material Table 1.

4. Experiments and Results

In this section, we describe the dataset used for training and evaluating SimuScope. Following this, we outline the implementation details and evaluation metrics. Finally, we present our quantitative results.

4.1. Dataset

To train SimuScope, we utilized two diverse real image sets with varying visual properties. The set for CholectL45 focuses on liver images from endoscopic views while also adapting to other tissues. The image set for CholectG45 emphasizes gallbladder images, ensuring diversity while incorporating other tissues and instruments. Both LoRAs were fine-tuned on fewer than 85 manually selected training images sourced from CholecT45 [33]. These training images were carefully chosen to represent different stages of the procedure and include a variety of tissues and instruments to ensure comprehensive coverage.

For inference, we created a dataset using video footage from a simulator, which included DaVinci instruments (see Figure 3 in Supplementary Material). The simulator output depicted partially dissected Calot's triangle and provided data on color, segmentation, depth maps, optical flow, normal maps, and tool masks (see Figure 4).

4.2. Implementation details

For training the diffusion model, we utilize the implementation from [64]. We perform all experiments using a single NVIDIA A100 80GB GPU. To maximize the benefits of the pre-trained model, we scale the training samples to a resolution of 512 $\times$ 512 pixels, matching the final pretraining resolution of Stable Diffusion 1.5. During the training phase, we only fine-tune the ControlNet branch of the model. This configuration allows us to train the model with a batch size of 1 through gradient accumulation. The model is trained with a learning rate of 1.2×10^{-3} for 20 epochs, with a maximum timestep of 1,000 and a text encoder learning rate of 1.2×10^{-3}. We minimize the Mean Squared Error (MSE) loss function with the Adafactor [45] optimizer.

4.3. Evaluation metrics

For evaluation, we employ key metrics to assess the efficacy and fidelity of generated data in computer vision applications. The mean Intersection over Union (mIoU) metric measures the semantic segmentation accuracy by evaluating the overlap between predicted segmentation masks and ground truth annotations.

The Fréchet Inception Distance (FID) [7] quantifies the distance between feature distributions of real and generated images:

$$\text{FID} = \|\mu_X - \mu_Y\|_2^2 + \text{Tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{1/2}), \quad (2)$$

where (μ_X, Σ_X) and (μ_Y, Σ_Y) represent the mean and covariance matrix of the feature distributions of real and generated images, respectively.

The Kernel Inception Distance (KID) [3] evaluates the distributional similarity between real and generated datasets using kernel embeddings:

$$\text{KID} = \|\mu_X - \mu_Y\|_2^2 + \text{Tr}(\Sigma_X + \Sigma_Y - 2(\Sigma_X\Sigma_Y)^{1/2}), \quad (3)$$

where μ_X, Σ_X and μ_Y, Σ_Y denote the mean and covariance matrix of kernel embeddings for real and generated data distributions. Additionally, to assess the quality of the image based on features extracted from the simulator, we use CMMD [11] to calculate the result. It is suitable for smaller datasets because it is unbiased, which we cannot claim in the case of FID. To assess sample fidelity, we use the Density and Coverage [47] metrics, both of which are derived from nearest neighbors in the representation space. Density quantifies how many neighborhood spheres of real samples encompass a given sample. For measuring sample diversity, we use Coverage, which also relies on the nearest neighbors in the representation space. Density and Coverage indicate the perceptual quality and diversity of the generated images, respectively:

$$\left(\{x_i^g\}_{i=1}^n, \{x_j^r\}_{j=1}^m\right) = \\ = \frac{1}{kn} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}\left(x_i^g \in B\left(x_j^r, \text{NND}_k\left(x_j^r\right)\right)\right), \quad (4)$$

$$\left(\{x_i^g\}_{i=1}^n, \{x_j^r\}_{j=1}^m\right) = \\ = \frac{1}{m} \sum_{j=1}^{m} \max_{i=1,\ldots,n} \mathbf{1}\left(x_i^g \in B\left(x_j^r, \text{NND}_k\left(x_j^r\right)\right)\right). \quad (5)$$

where $\mathbf{1}(\cdot)$ indicate the indicator function, $S(\{x_j^r\}_{j=1}^m) = \bigcup_{j=1}^m B(x_j^r, \text{NND}_k(x_j^r))$, where $B(x, r)$ defines a Euclidean ball centered at x with radius r, and $\text{NND}_k(x_j^r)$ is the distance between x_j^r and its k-th nearest neighbour in $\{x_j^r\}_{j=1}^m$, excluding itself [47].

Table 1. Quantitative results on **our** simulator.We compare the raw simulator output with data generated using a concurrent method. We demonstrate the effectiveness of our approach using key metrics including mIoU, FID, KID, CMMD, Density, and Coverage.

Method	Style	mIoU [%] ↑	FID ↓	KID ↓	CMMD ↓	Density ↑	Coverage ↑
N/A	Raw	60.79	202.54	.1888	3.442	0.011	0.004
LC-SD [16]	Mixed styles	67.42	**68.26**	**.0635**	1.017	0.077	0.104
SimuScope	CholectD45	**70.65**	79.80	.0690	**0.697**	**0.114**	**0.117**

4.4. Quantitative results

Table 2. Ablation study results showing mIoU (%) for different combinations of SoftEdge, Reference, and Depth controls.

SoftEdge	Reference	Depth	mIoU [%]
×	×	×	41.00
✓	✓	×	64.39
✓	×	✓	54.37
×	✓	✓	58.29
✓	✓	✓	**70.65**

We evaluate the efficacy of SimuScope in generating realistic surgical images by comparing them against the raw simulator output using key metrics: mIoU, FID, and KID. Our results, as shown in Table 1, demonstrate substantial improvements over the baseline simulator. Specifically, employing *SimuScope* with style CholectD45 yields a notable increase in mIoU from 60.79% to 70.65%, indicating enhanced semantic segmentation accuracy. Moreover, the FID decreases significantly from 202.54 to 79.80, the KID also showed improvement from 0.1888 to 0.0690, the CMMD decrease from 1.017 to 0.697, the Density increased from 0.077 to 0.114 and the Coverage from 0.104 to 0.117. These metrics underscore the capability of SimuScope to produce surgical images that closely resemble real-world scenarios, showcasing its potential for advancing computer-assisted surgery through enhanced training and simulation capabilities.

For comparison, we utilized LoRAs on our simulator using the style from LC-SD, which, according to the article, yielded the best results. The visual comparison revealed that the data generated by our method achieves a similar perceptual realism to the work of LC-SD [16] (see Figure 6). Additionally, it should be noted that our simulator focuses on details and has less blurred details compared to LC-SD. We used the Mixed style, which, according to the article by [16], achieved the best results for comparison on our simulator. As observed, we achieved an improvement in mIoU from 67.42% to 70.65%. Additionally, according to the proposed metrics such as CMMD, Density, and Coverage, we observe a significant difference. These metrics highlight an increase in both fidelity and diversity of the images compared to LC-SD. Conversely, FID and KID perform worse compared to LC-SD.

Figure 5 and Table 2 provide an overview of mIoU values for different types of control and their enhancements compared to no control inference. Table 2 shows that combined control models yielded the best overall results. The highest performance was achieved by using three ControlNets together. For the CholectD45 style without control, the result was 41%. With the application of SoftEdge and Depth, the mIoU increased to 54.37% (+13.37%). A further increase to 58.29% (+17.19%) was observed with the combination of ControlNet Reference and Depth. The most significant improvement of 33.35% was achieved with the combination of SoftEdge, Reference, and Depth control, resulting in an mIoU of 70.65%.

5. Discussion and Conclusions

In this work, we introduced SimuScope, a novel framework that combines high-fidelity surgical simulation with advanced diffusion models. This innovative system facilitates the generation of photo-realistic surgical footage, which is automatically and fully labeled with essential annotations such as semantic segmentation, depth and normal maps, optical flow, action triplets, and 2D/3D instrument poses. By leveraging image-to-image translation techniques, we effectively transformed rendered images from the simulator into their realistic equivalents using SD, conditioned by multiple ControlNets.

Compared to previous works, like [16], which only considered a simple "fly-through" a static 3D anatomy, SimuScope supports a wide range of surgical interactions such as tissue grasping, tearing, cutting, thermo-coagulation and clipping. Such interactions are ubiquitous in surgery and containing them in the generated data is indispensable for context-aware, comprehensive and fine-grained analysis of surgical activities and workflow using deep-learning.

Our comprehensive evaluation, employing metrics such as mIoU, FID, KID, CMMD, Density and Coverage has demonstrated that SimuScope is capable of producing photorealistic surgical images that accurately align with the simulator's output. This high degree of fidelity ensures that

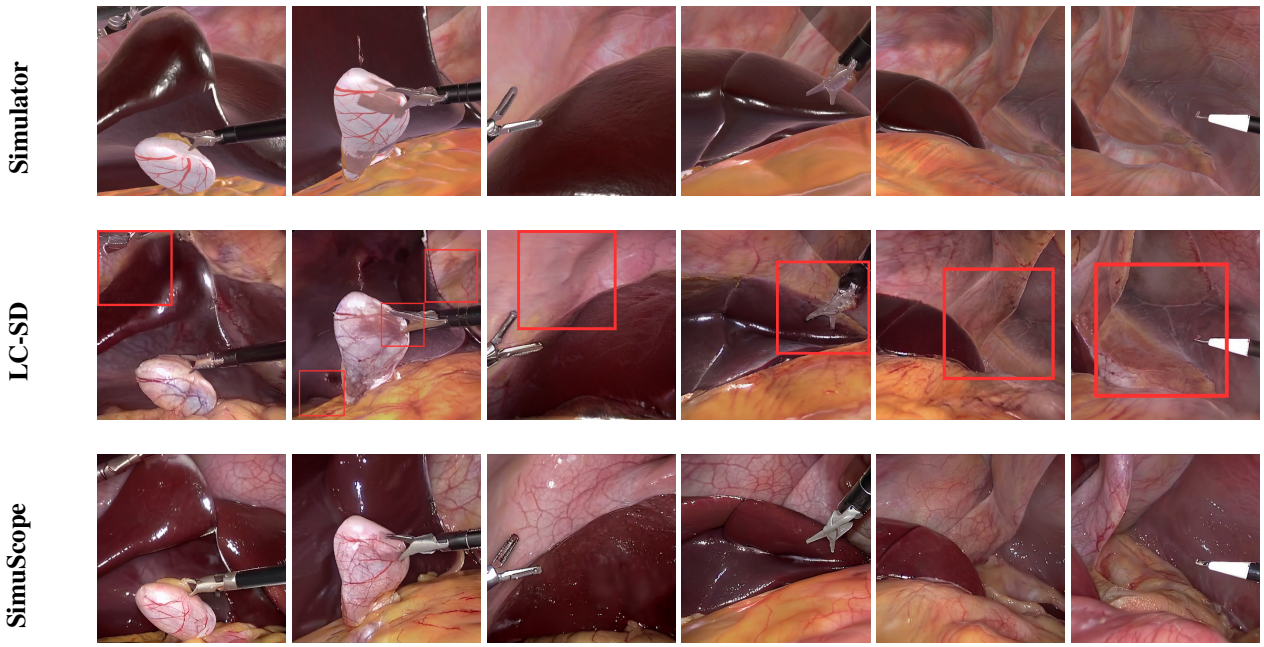

Figure 6. Comparison of applying LC-SD styles on our simulator to our result. Images generated using LC-SD styles appear less realistic, with blurred details. A lack of depth in the generated tissues can also be observed.

the synthetic data generated is both visually convincing and semantically consistent. SimuScope addresses the critical bottleneck of data scarcity by efficiently generating large datasets of high-quality imaging data. These datasets are indispensable for training and developing data-hungry deep learning models that underpin modern CAS systems.

The main limitation of this work is the lack of temporal coherency between the generated frames. As the generative AI community has recently shifted focus towards this problem (i.e., video-to-video generation), we aim to address this issue in future work.

By providing a robust and scalable solution for synthetic data generation, SimuScope holds the potential to enhance the capabilities of CAS systems significantly. This advancement can lead to safer and more precise surgical outcomes, with fewer errors and complications. Moreover, generating diverse and comprehensive datasets can accelerate research and innovation in the field, ultimately contributing to improved surgical training, better-informed clinical decisions, and enhanced patient care.

To conclude, SimuScope represents a significant step forward in integrating realistic surgical simulation with cutting-edge generative models. The success of this approach opens up new avenues for the development of sophisticated AI-driven tools in surgery, paving the way for a future where advanced CAS systems can deliver unprecedented levels of support to surgeons, thereby improving both the safety and efficacy of surgical procedures.

Acknowledgements

This paper received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857533. The research is supported by Sano project carried out within the International Research Agendas programme of the Foundation for Polish Science, co-financed by the European Union under the European Regional Development Fund. The research was created within the project of the Minister of Science and Higher Education "Support for the activity of Centers of Excellence established in Poland under Horizon 2020" on the basis of the contract number MEiN/2023/DIR/3796. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017108.

References

[1] Binod Bhattarai, Ronast Subedi, Rebati Raman Gaire, Eduard Vazquez, and Danail Stoyanov. Histogram of oriented gradients meet deep learning: A novel multi-task deep network for 2d surgical image semantic segmentation. *Medical Image Analysis*, 85:102747, 2023. 1

[2] Florentin Bieder, Julia Wolleb, Alicia Durrer, Robin Sandkuehler, and Philippe C Cattin. Memory-efficient 3d denoising diffusion models for medical image processing. In *Medical Imaging with Deep Learning*, 2023. 2

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018. 6

[4] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5343–5353, 2024. 5

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[6] Moein Heidari, Amirhossein Kazerouni, Milad Soltany, Reza Azad, Ehsan Khodapanah Aghdam, Julien Cohen-Adad, and Dorit Merhof. Hiformer: Hierarchical multi-scale representations using transformers for medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6202–6212, 2023. 2

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[9] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. 2, 4

[10] Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The GAN is dead; long live the GAN! a modern baseline GAN. In *ICML 2024 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2024. 2

[11] Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9307–9315, 2024. 6

[12] Lan Jiang, Ye Mao, Xiangfeng Wang, Xi Chen, and Chao Li. Cola-diff: Conditional latent diffusion model for multi-modal mri synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 398–408. Springer, 2023. 2

[13] Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging*, 40(7):1911–1923, 2021. 1

[14] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15988–15998, 2023. 5

[15] Joanna Kaleta, Diego Dall'alba, Szymon Plotka, and Przemyslaw Korzeniowski. Lc-sd: Realistic endoscopic image generation with limited training data. In *Deep Generative Models for Health Workshop NeurIPS 2023*. 4

[16] Joanna Kaleta, Diego Dall'Alba, Szymon Płotka, and Przemysław Korzeniowski. Minimal data requirement for realistic endoscopic image generation with stable diffusion. *International journal of computer assisted radiology and surgery*, 19(3):531–539, 2024. 2, 3, 4, 5, 6, 7

[17] Boah Kim, Inhwa Han, and Jong Chul Ye. Diffusemorph: Unsupervised deformable image registration using diffusion model. In *European Conference on Computer Vision*, pages 347–364. Springer, 2022. 2

[18] Boah Kim, Yujin Oh, and Jong Chul Ye. Diffusion adversarial representation learning for self-supervised vessel segmentation. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[19] Boah Kim and Jong Chul Ye. Diffusion deformable model for 4d temporal medical image generation. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 539–548, Cham, 2022. Springer Nature Switzerland. 2

[20] Jonghun Kim and Hyunjin Park. Adaptive latent diffusion model for 3d medical image to image translation: Multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7604–7613, January 2024. 2

[21] Przemysław Korzeniowski, Szymon Płotka, Robert Brawura-Biskupski-Samaha, and Arkadiusz Sitek. Virtual reality simulator for fetoscopic spina bifida repair surgery. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 401–406, 2022. 2

[22] Ming Li, Taojiannan Yang, Huafeng Kuang, Jie Wu, Zhaoning Wang, Xuefeng Xiao, and Chen Chen. Controlnet++: Improving conditional controls with efficient consistency feedback. In *European Conference on Computer Vision*, pages 129–147. Springer, 2025. 5

[23] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023. 5

[24] Yunxiang Li, Hua-Chieh Shao, Xiao Liang, Liyuan Chen, Ruiqi Li, Steve Jiang, Jing Wang, and You Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE Transactions on Medical Imaging*, PP:1–1, 10 2023. 2

[25] Shan Lin, Fangbo Qin, Yangming Li, Randall A. Bly, Kris S. Moe, and Blake Hannaford. Lc-gan: Image-to-image translation based on generative adversarial network for endoscopic images. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2914–2920, 2020. 2

[26] Jiaming Liu, Rushil Anirudh, Jayaraman J Thiagarajan, Stewart He, K Aditya Mohan, Ulugbek S Kamilov, and Hyojin Kim. Dolce: A model-based probabilistic diffusion framework for limited-angle ct reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10498–10508, 2023. 2

[27] Ange Lou, Kareem Tawfik, Xing Yao, Ziteng Liu, and Jack Noble. Min-max similarity: A contrastive semi-supervised deep learning network for surgical tools segmentation. *IEEE Transactions on Medical Imaging*, 42(10):2832–2841, 2023. 1

[28] Miles Macklin and Matthias Muller. A constraint-based formulation of stable neo-hookean materials. In *Motion, Interaction and Games*, pages 1–7. 2021. 3, 4

[29] Miles Macklin, Kier Storey, Michelle Lu, Pierre Terdiman, Nuttapong Chentanez, Stefan Jeschke, and Matthias Müller. Small steps in physics simulation. In *Proceedings of the 18th annual ACM siggraph/eurographics symposium on computer animation*, pages 1–7, 2019. 3

[30] Nina Montaña-Brown, Shaheer U Saeed, Ahmed Abdulaal, Thomas Dowrick, Yakup Kilic, Sophie Wilkinson, Jack Gao, Meghavi Mashar, Chloe He, Alkisti Stavropoulou, et al. Saramis: simulation assets for robotic assisted and minimally invasive surgery. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[31] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4296–4304, 2024. 5

[32] Matthias Müller, Miles Macklin, Nuttapong Chentanez, Stefan Jeschke, and Tae Kim. Detailed rigid body simulation with extended position based dynamics. *Computer Graphics Forum*, 39:101–112, 12 2020. 3

[33] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. 1, 2, 4, 6

[34] Masahiro Oda, Kiyohito Tanaka, Hirotsugu Takabatake, Masaki Mori, Hiroshi Natori, and Kensaku Mori. Realistic endoscopic image generation method using virtual-to-real image-domain translation. *Healthcare Technology Letters*, 6(6):214–219, Nov. 2019. 2

[35] Micha Pfeiffer, Isabel Funke, Maria R Robu, Sebastian Bodenstedt, Leon Strenger, Sandy Engelhardt, Tobias Roß, Matthew J Clarkson, Kurinchi Gurusamy, Brian R Davidson, et al. Generating large labeled data sets for laparoscopic image processing tasks using unpaired image-to-image translation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*, pages 119–127. Springer, 2019. 2

[36] Szymon Płotka, Tomasz Szczepański, Paula Szenejko, Przemysław Korzeniowski, Jesús Rodriguez Calvo, Asma Khalil, Alireza Shamshirsaz, Robert Brawura-Biskupski-Samaha, Ivana Išgum, Clara I Sánchez, et al. Real-time placental vessel segmentation in fetoscopic laser surgery for twin-to-twin transfusion syndrome. *Medical Image Analysis*, 99:103330, 2025. 1

[37] Gian Andrea Prevost, Benjamin Eigl, Iwan Paolucci, Tobias Rudolph, Matthias Peterhans, Stefan Weber, Guido Beldi, Daniel Candinas, and Anja Lachenmayer. Efficiency, accuracy and clinical applicability of a new image-guided surgery system in 3d laparoscopic liver surgery. *Journal of gastrointestinal surgery*, 24(10):2251–2258, 2020. 1

[38] Aimon Rahman, Jeya Maria Jose Valanarasu, Ilker Hacihaliloglu, and Vishal M Patel. Ambiguous medical image segmentation using diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11536–11546, 2023. 2

[39] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 5

[40] Dominik Rivoir, Micha Pfeiffer, Reuben Docea, Fiona Kolbinger, Carina Riediger, Jürgen Weitz, and Stefanie Speidel. Long-term temporally consistent unpaired video translation from simulated surgical 3d data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3323–3333, 2021. 2

[41] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 4

[42] Manish Sahu, Ronja Strömsdörfer, Anirban Mukhopadhyay, and Stefan Zachow. Endo-sim2real: Consistency learning-based domain adaptation for instrument segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 784–794, Cham, 2020. Springer International Publishing. 2

[43] Crispin Schneider, Moustafa Allam, Danail Stoyanov, DJ Hawkes, K Gurusamy, and BR Davidson. Performance of image guided navigation in laparoscopic liver surgery–a systematic review. *Surgical Oncology*, 38:101637, 2021. 1

[44] Lalith Sharan, Gabriele Romano, Sven Koehler, Halvar Kelm, Matthias Karck, Raffaele De Simone, and Sandy Engelhardt. Mutually improved endoscopic image synthesis and landmark detection in unpaired image-to-image translation. *IEEE Journal of Biomedical and Health Informatics*, 26(1):127–138, 2022. 2

[45] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pages 4596–4604. PMLR, 2018. 6

[46] Grzegorz Skorupko, Richard Osuala, Zuzanna Szafranowska, Kaisar Kushibar, Nay Aung, Steffen E Petersen, Karim Lekadir, and Polyxeni Gkontra. Debiasing cardiac imaging with controlled latent diffusion models. *arXiv preprint arXiv:2403.19508*, 2024. 2

[47] George Stein, Jesse Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L Caterini, Eric Taylor, and Gabriel Loaiza-Ganem. Exposing

flaws of generative model evaluation metrics and their unfair treatment of diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[48] Tatsushi Tokuyasu, Yukio Iwashita, Yusuke Matsunobu, Toshiya Kamiyama, Makoto Ishikake, Seiichiro Sakaguchi, Kohei Ebe, Kazuhiro Tada, Yuichi Endo, Tsuyoshi Etoh, et al. Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surgical endoscopy*, 35:1651–1658, 2021. 1

[49] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 4

[50] Danush Kumar Venkatesh, Dominik Rivoir, Micha Pfeiffer, Fiona Kolbinger, Marius Distler, Jürgen Weitz, and Stefanie Speidel. Exploring semantic consistency in unpaired image translation to generate data for surgical applications. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):985–993, 2024. 2

[51] Tom Vercauteren, Mathias Unberath, Nicolas Padoy, and Nassir Navab. Cai4cai: The rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE*, 108(1):198–214, 2020. 1

[52] Xudong Wang, Trevor Darrell, Sai Saketh Rambhatla, Rohit Girdhar, and Ishan Misra. Instancediffusion: Instance-level control for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6232–6242, 2024. 5

[53] Zihao Wang, Yingyu Yang, Yuzhou Chen, Tingting Yuan, Maxime Sermesant, Hervé Delingette, and Ona Wu. Mutual information guided diffusion for zero-shot cross-modality medical image translation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. 2

[54] Zihao Wang, Yingyu Yang, Yuzhou Chen, Tingting Yuan, Maxime Sermesant, Hervé Delingette, and Ona Wu. Mutual information guided diffusion for zero-shot cross-modality medical image translation. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. 2

[55] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *International Conference on Medical image computing and computer-assisted intervention*, pages 35–45. Springer, 2022. 2

[56] Junde Wu, Rao Fu, Huihui Fang, Yu Zhang, Yehui Yang, Haoyi Xiong, Huiying Liu, and Yanwu Xu. Medsegdiff: Medical image segmentation with diffusion probabilistic model. In *Medical Imaging with Deep Learning*, pages 1623–1639. PMLR, 2024. 2

[57] Julian Wyatt, Adam Leach, Sebastian M. Schmon, and Chris G. Willcocks. Anoddpm: Anomaly detection with denoising diffusion probabilistic models using simplex noise. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 649–655, 2022. 2

[58] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 5

[59] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1395–1403, 2015. 5

[60] Yijun Yang, Huazhu Fu, Angelica Aviles-Rivero, Carola-Bibiane Schönlieb, and Lei Zhu. *DiffMIC: Dual-Guidance Diffusion Network for Medical Image Classification*, pages 95–105. 10 2023. 2

[61] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023. 5

[62] Paolo Zaffino, Sara Moccia, Elena De Momi, and Maria Francesca Spadea. A review on advances in intra-operative imaging for surgery and therapy: imagining the operating room of the future. *Annals of Biomedical Engineering*, 48(8):2171–2191, 2020. 1

[63] Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant, 2024. 2

[64] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5, 6

[65] Tianjun Zhang, Yi Zhang, Vibhav Vineet, Neel Joshi, and Xin Wang. Controllable text-to-image generation with gpt-4. *arXiv preprint arXiv:2305.18583*, 2023. 5

[66] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. Migc: Multi-instance generation controller for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6818–6828, 2024. 5

[67] Muzaffer Özbey, Onat Dalmaz, Salman U. H. Dar, Hasan A. Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023. 2