International Journal of Computer Assisted Radiology and Surgery (2025) 20:1421–1429
https://doi.org/10.1007/s11548-025-03397-y

ORIGINAL ARTICLE

SurGrID: controllable surgical simulation via Scene Graph to Image Diffusion

Yannik Frisch[1,4] · Ssharvien Kumar Sivakumar[1] · Çağhan Köksal[2] · Elsa Böhm[4] · Felix Wagner[4] · Adrian Gericke[4] · Ghazal Ghazaei[3] · Anirban Mukhopadhyay[1]

Received: 12 March 2025 / Accepted: 14 April 2025 / Published online: 21 May 2025
© The Author(s) 2025

Abstract

Purpose Surgical simulation offers a promising addition to conventional surgical training. However, available simulation tools lack photorealism and rely on hard-coded behaviour. Denoising Diffusion Models are a promising alternative for high-fidelity image synthesis, but existing state-of-the-art conditioning methods fall short in providing precise control or interactivity over the generated scenes.

Methods We introduce SurGrID, a Scene Graph to Image Diffusion Model, allowing for controllable surgical scene synthesis by leveraging Scene Graphs. These graphs encode a surgical scene's components' spatial and semantic information, which are then translated into an intermediate representation using our novel pre-training step that explicitly captures local and global information.

Results Our proposed method improves the fidelity of generated images and their coherence with the graph input over the state of the art. Further, we demonstrate the simulation's realism and controllability in a user assessment study involving clinical experts.

Conclusion Scene Graphs can be effectively used for precise and interactive conditioning of Denoising Diffusion Models for simulating surgical scenes, enabling high-fidelity and interactive control over the generated content.

Keywords Controllable surgical simulation · Scene Graph · Denoising Diffusion Models

Introduction

The conventional methodology for training novice surgeons follows the Halstedian apprenticeship model, in which trainees are deemed competent after completing a minimum number of surgical procedures under supervision [1]. Nevertheless, the Halstedian approach raises ethical concerns over using patients for training and has been associated with increased complication rates and poor outcomes [2]. However, when paired with simulation-based training, it offers a promising way to overcome regulatory and ethical complexities while establishing a safe and controlled environment where novice surgeons can refine their skills without the potentially severe consequences of real-world failures. It has also been shown that surgeons who have received simulation-based training demonstrate higher overall performance and make fewer errors during their initial surgeries than those who only received conventional training [3].

In addition to using phantom or cadaver specimens for training simulation, numerous studies have shown the effectiveness of virtual reality simulators for training [3, 4]. However, these tools are computer graphics-based, with manually defined rules for rendering logic. Therefore, new surgery techniques and edge cases must be manually programmed, but still fall short in replicating the complexity and variability of human anatomy and the subtle nuances of real-life surgical procedures [5]. Recent works propose

Yannik Frisch and Ssharvien Kumar Sivakumar contributed equally to this work

✉ Yannik Frisch
yannik_phil.frisch@tu-darmstadt.de

✉ Ssharvien Kumar Sivakumar
ssharvien_kumar.sivakumar@tu-darmstadt.de

1 TU Darmstadt, Fraunhoferstr. 5, 64297 Darmstadt, Germany

2 TU Munich, Arcisstr. 21, 80333 Munich, Germany

3 Carl Zeiss AG, Kistlerhofstr. 75, 81379 Munich, Germany

4 Universitätsmedizin Mainz, Langenbeckstr. 1, 55131 Mainz, Germany

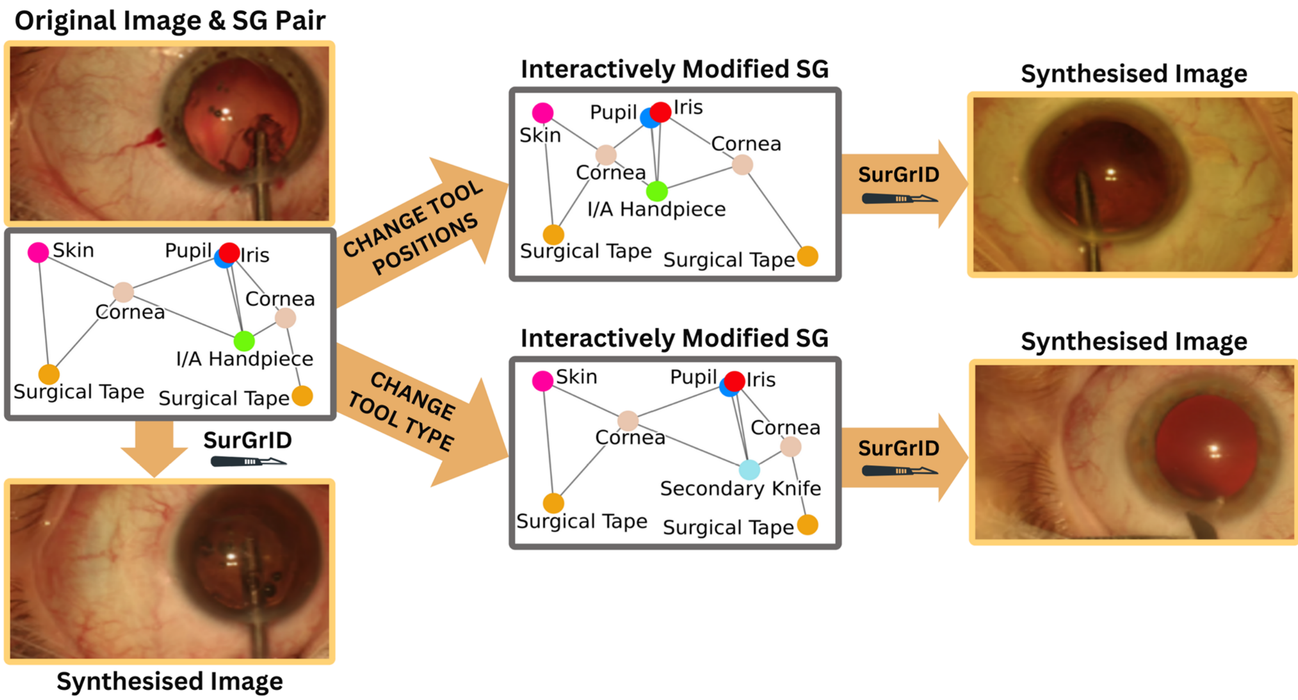

Fig. 1 Concept of SurGrID. SurGrID conditions image synthesis on Scene Graphs (SGs) for precise control over the anatomy/tool type, size, and position. In the bottom left, SurGrID generates a new image spatially identical to the original image using only the ground-truth SG. This demonstrates that SGs can effectively encode a surgical scene's spatial and semantic information. On the right, we interactively modify the SG to control anatomy/tool position (top) and type (bottom)

using Denoising Diffusion Models (DDMs) for photorealistic interactive simulation to address it. However, these methods either rely on text prompts to guide generation [6], which offers limited spatial control, or use mask prompts [5], which are not easily interactable and require continuous mask adjustments.

Recognising this gap, we propose SurGrID, a compact yet interpretable and precise conditioning of DDMs with Scene Graphs (SGs). SurGrID is trained on surgical videos of actual patients and synthesises new surgical scenes with high controllability and excellent fidelity. At the core of our method lies the adoption of SG representations as compact and human-readable encodings of surgical scenes [7–9]. The hierarchical and relational nature of SGs makes them an ideal candidate for simulating dynamic and interactive surgical environments. We encode precise spatial information, such as the size and position of the anatomies and tools, into the node features of SGs obtained from a surgical scene, allowing fine-tuned control over the synthesis, as shown in Fig. 1.

To synthesise images that accurately reflect the SG, we must acquire an intermediate representation that can align between the SG and images [10, 11]. To this end, we propose pre-training a graph encoder on the (image, segmentation, graph) triplet that captures local and global information from the surgical scene. We focus on predicting fine-grained details of the anatomy and tools for local information by masking them randomly and reconstructing masked regions based on the SG input. We employ a discriminative approach to learn global information, which captures the layout and interaction between anatomy and tools. This approach enforces the encoder to cluster representations of compliant SGs and masks in the latent space.

We then leverage the obtained graph embeddings for conditionally training a DDM [12] to synthesise realistic surgical images while allowing precise control over the generated content through graph conditioning. The ability to generate diverse and realistic images from such structured representations holds immense potential for surgical simulation, offering a novel avenue for creating varied and high-fidelity surgical scenes. To the best of our knowledge, this manuscript is the first to propose SG to Image Diffusion for precisely controllable surgical simulation.

Contributions

- We show that SGs can encode surgical scenes in a human-readable format. We propose a novel pre-training step that encodes global and local information from (image, mask, SG) triplets. The learned embeddings are employed to condition graph to Image Diffusion for high-quality and precisely controllable surgical simulation.

- We evaluate our generative approach on scenes from cataract surgeries using quantitative fidelity and diversity measurements, followed by an extensive user study involving clinical experts.

Related work

Simulation by generative models typically conditions the synthesis on simple inputs such as class labels [13, 14], reference images [15, 16], or text queries [17]. However, none of these techniques provides a highly controllable and precise conditioning mechanism suitable for complex, high-risk scenarios such as surgical simulation. We postulate that Scene Graph conditioning can serve as an accurate conditioning mechanism.

Translating SG representations into realistic images has gained interest within the computer vision community. To achieve this task, Johnson et al. [18] use the intermediate representation of a *scene layout*, which they predict from input SGs and translate into an image using a Cascaded Refinement Network. With the rise of Denoising Diffusion Models [12, 19, 20] as a competing generative method, Yang et al. [10] have laid foundational principles for SG to image synthesis with diffusion models, utilising Masked Contrastive Pre-Training to obtain graph embeddings for conditioning. SceneGenie [21] guides the denoising process through information from CLIP embeddings [22] for input text queries together with scene layout and segmentation mask predictions from the text query. R3CD [23] introduces SG-Transformers to improve the global and local information in graph embeddings and uses a contrastive loss to improve relation-aware synthesis. Mishra et al. [11] introduce adversarial pre-training of a graph encoder, aligning graph embeddings with CLIP image embeddings [22] and removing the need for an intermediate scene layout representation.

These methods operate under the *assumption that the image embeddings corresponding to a particular SG will be distant from those associated with entirely different SGs*. However, this assumption falls short in surgical simulation, rendering these approaches ineffective when applied directly. For instance, neighbouring frames in a surgical video may depict different tools and have distinct SGs. Yet, they occupy a similar position in the image embedding space due to similarly looking anatomy. As a result, frames from the same video tend to cluster together in the image embedding space. However, we would like frames with similar surgical tools and positions to be close in the embedding space, regardless of the video they come from. The existing methods further do not address the need for precise control over positioning and sizing, which is essential for surgical simulation.

Method

This section presents our approach for translating Scene Graphs into intermediate representations that capture both local and global information. Subsequently, we outline the conditional diffusion model used for SG to image generation.

Pre-training Scene Graph encoder

We leverage a surgical segmentation dataset $\mathcal{D}$ consisting of (image, mask) pairs $(x, m) \in \mathcal{D}$ to extract SGs from ground-truth segmentation masks. Details on this conversion can be found in Supplementary Section A. The approaches in subsequent sections can seamlessly be included in datasets with already available SGs.

For the graph encoder E_G, we employ a series of stacked graph neural network (GNN) layers [24] to process the input graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Each GNN layer updates node representations by aggregating information from their neighbours, effectively capturing local graph topologies:

$$h_v^{(l+1)} = \mathrm{GNN}^{(l)}\left(h_v^{(l)}, \bigoplus_{u \in \mathcal{N}(v)} h_u^{(l)}\right) \tag{1}$$

where $h_v^{(l)}$ is the feature vector of node v at layer l, $N(v)$ represents the neighbours of v, and $\bigoplus$ denotes a differentiable, permutation-invariant aggregation function, such as sum, mean, or max. The final node representations are aggregated through mean pooling to form a graph-level latent representation z_G.

We propose pre-training two distinct graph encoders, as visualised in Fig. 2: E_G^{loc}, which focuses on capturing local information, hence embedding fine-grained details of the anatomy and surgical tools. E_G^{glob} concentrates on retaining global information, focusing on the overall structural alignment based on compositions and interactions within the surgical scene. The overall pre-training occurs in the latent space, where embeddings for the image, z_x, and segmentation masks, z_m, are obtained using a VQ-GAN model [25]. These models are trained separately and referred to as E_m and E_x in Fig. 2. Supplementary Section B analyses and motivates the pre-training mechanisms using t-SNE visualisations of the embedding spaces.

For the local information, we randomly select a class from an image x's paired mask and apply rectangular masking of that class's bounding box in the image. Both the full image x and the masked image x^r are encoded as $z_x = E_x(x)$ and $z_x^r = E_x(x^r)$. To train E_G^{loc}, the task is framed as predicting

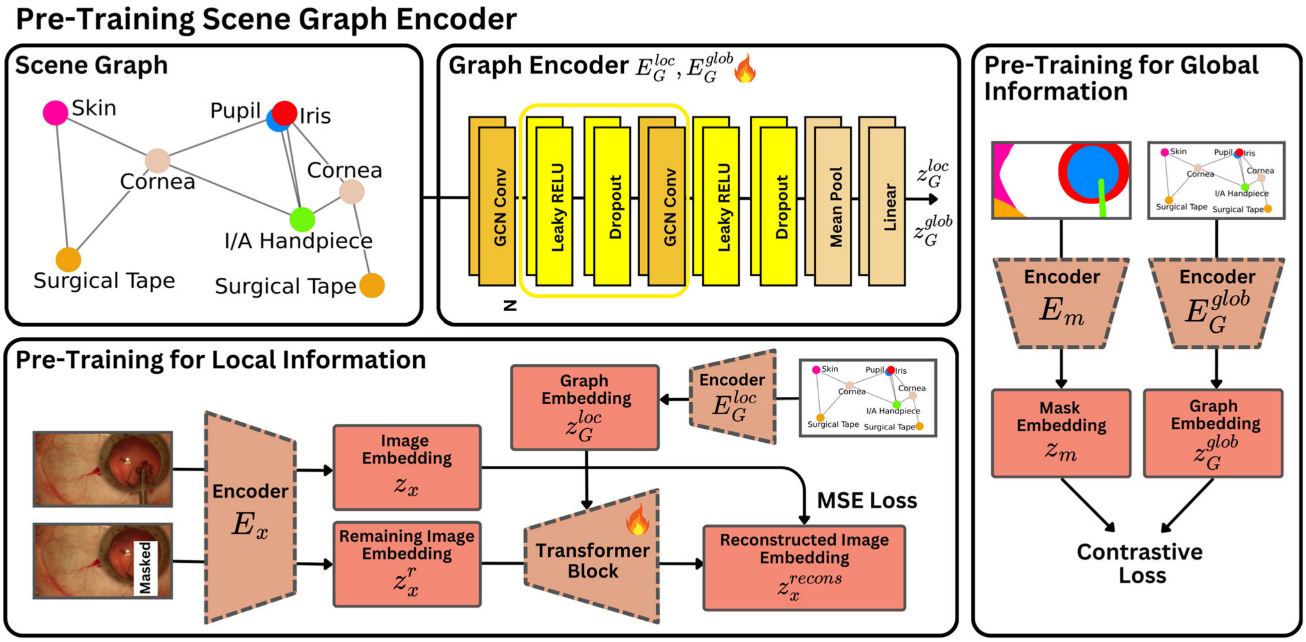

Fig. 2 SurGrID workflow for pre-training graph encoders. To capture local and global information, we pre-train two separate graph encoders composed of stacked GNN layers (top centre). E_G^{loc} is trained to encode local information by randomly masking objects in the image and predicting the masked area using the SG (bottom). E_G^{glob} is trained to encode global information by aligning the graph embedding space with the segmentation mask embedding space (right)

z_x based on the available z_x^r and its corresponding SG representation $z_G^{\text{loc}} = E_G^{\text{loc}}(\mathcal{G})$. The graph encoder is trained in conjunction with a transformer-based decoder, d_θ, to minimise the reconstruction loss:

$$\mathcal{L}_{\text{local}} = \mathbb{E}_{(x,\mathcal{G})\sim D} \left\| z_x - d_\theta \left(z_x^r, z_G^{loc} \right) \right\|_2^2 \tag{2}$$

To capture the global information of the surgical scene, we train E_G^{glob} by aligning its graph embeddings, $z_G^{\text{glob}} = E_G^{\text{glob}}(\mathcal{G})$, with the segmentation mask embeddings, z_m, inspired by CLIP's training approach [22]. This alignment must not be performed in the image embedding space, as neighbouring frames with different tools in a surgical video often share a large portion of the anatomical features, causing them to cluster in the embedding space. Our goal, however, is to ensure that frames featuring the same tool, even across different videos and possibly different textural features, are grouped in the embedding space. To achieve this, we include segmentation mask embeddings, as they capture high-level information about the overall scene layout and are a more detailed representation of the SG. This enables them to accurately represent the SG in the embedding space, bringing similar SGs closer together. For training, we employ a contrastive loss:

$$\mathcal{L}_{\text{global}} = \mathbb{E}_{(\mathcal{G},m^+,\{m_i^-\}_{i=1}^k)\sim D} \left[-\log \frac{\exp\left(z_G \cdot z_{m^+}\right)}{\exp\left(z_G \cdot z_{m^+}\right) + \sum_i \exp\left(z_G \cdot z_{m_i^-}\right)} \right] \tag{3}$$

where m^+ represents the compliant segmentation mask to $\mathcal{G}$, while $m_i{}^-$ are the negative targets, which do not comply with $\mathcal{G}$ and are randomly sampled from $\mathcal{D}$.

SG-conditioned Denoising Diffusion Model

For image synthesis, we combine the standard diffusion model setup outlined in DDM [12] with Classifier-Free Guidance (CFG) [26]. The conditioning $c = concat(z_G^{\text{loc}}, z_G^{\text{glob}})$ is derived by fusing the embeddings from the two graph encoders $E_G^{\text{loc}}, E_G^{\text{glob}}$. We then train a parameterised denoising model ϵ_θ to approximate denoising steps $p(x_{t-1}|x_t, c)$ by minimising:

$$\mathcal{L}_{\text{DDM}} = \mathbb{E}_{t,x_0,\epsilon} \left[\|\epsilon - \epsilon_\theta(x_t, t, c)\|^2 \right] \tag{4}$$

where $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim \mathcal{N}(0, I)$, and $\bar{\alpha}_t$ is the variance schedule of the diffusion process. By then sampling z_T from a Gaussian distribution $p(z_T) \sim \mathcal{N}(0, I)$, we can synthesise new samples by iteratively applying the denoising network.

CFG substitutes the predicted noise $\epsilon_\theta(z_t, t, c)$ at each diffusion time step t as:

$$\epsilon'_\theta(x_t, t, c) := (1+\omega)\epsilon_\theta(x_t, t, c) - \omega\epsilon_\theta(x_t, t) \tag{5}$$

where ω is a scaling parameter, which we empirically set to 2.0, and $\epsilon_\theta(x_t, t)$ is trained by randomly dropping the conditioning with probability 0.2.

Experiments and results

This section outlines our experimental setup and provides quantitative and qualitative evaluations of the high-quality surgical samples generated from Scene Graphs. We conclude with a visual assessment user study involving clinical experts.

Setup and datasets

We evaluate our approach on the *CaDIS* dataset [27], a semantic segmentation dataset for cataract surgery videos. Cataract surgery is one of the most frequently performed procedures worldwide [28], where clouded lenses are removed and replaced with artificial ones. The dataset captures 4670 frames from 25 videos which are at least 0.3s apart. They have a resolution of 960 × 540 pixels. In setting II, as defined in *CaDIS* [27], the dataset includes segmentation masks for 17 classes divided into four anatomical labels, ten tool labels, and three miscellaneous labels.

We split the examples based on the 25 available videos, allocating 19 for training, 3 for validation, and 3 for testing, resulting in 3550, 534, and 586 (image, mask, graph) triplets, respectively. Our model was trained on an NVIDIA RTX 4090 GPU with an image size of 128 × 128. The SG representations, model weights, and code to reproduce our results will be provided at https://github.com/MECLabTUDA/SurGrID upon acceptance.

Quantitative image analysis

To evaluate the objective image quality of generated samples, we synthesise examples from ground-truth SGs and assess the FID and KID scores [29] against the same amount of real samples. Further, we evaluate the diversity of generated samples using the LPIPS diversity metric [30].

To also assess how effectively the synthesised images adhere to the conditioning provided by the SG, we employ the Mask R-CNN model [31] pre-trained on *CaDIS* [27] for object detection, which predicts bounding boxes (BBs) along with class labels. This process essentially reverses the workflow by converting images back into SGs. By assessing the BBs with the IoU metric (at a 50% threshold) and evaluating the predicted object classes using the F1 score (at 50% IoU threshold), we can assess the accuracy of the synthesised anatomies and tools, along with their corresponding positions and sizes.

As an additional baseline, we deploy a latent diffusion model (LDM) [20] conditioned on CLIP embeddings [22] of NLP strings. These strings resemble a combination of triplets, which we construct from spatial relations such as "left of," "right of," "above," "below," "inside," and "surrounding." The strings are then embedded using the CLIP text encoder, and the LDM is conditionally trained on the resulting embeddings and otherwise analogously to the other methods.

As presented in Table 1, our method not only surpasses the baselines regarding FID and KID scores, reflecting improved image quality but also adheres more closely to the conditioning criteria, as shown by the higher BB IoU and F1 scores.

Ablation studies

We present ablation scores for our method trained exclusively on embeddings containing either only global or local information in the top part of Table 2. Combining both embeddings significantly improves the method's performance, highlighting the necessity to include both types of information in the generative process. The bottom part of Table 2 additionally displays ablation results on the guidance scale ω.

Qualitative results

Besides quantitative evaluations, we visually assess the quality of our generated images and how well changes to the input graph translate to the synthesised images. As shown in Fig. 3 our generative model can synthesise high-quality images with coherent visual semantics following the SG inputs.

The first column presents qualitative results from altering the positions of tools and anatomy within the SG, with the synthesised images accurately reflecting these positional adjustments. In the second column, the SG modifications focus solely on changing tool types. The generated images maintain a consistent spatial position while only adapting the type of displayed tool. Even though the tools can be displayed with different angles in the image, the centroid coordinates used for conditioning remain accurate. The third column progressively removes primary surgical tools, while the fourth column removes auxiliary tools such as retractors and surgical tape. In Appendix Section E, we have outlined limitations and future work. However, in most cases, the syn-

Table 1 Quantitative fidelity and diversity results

Method	FID (↓)	KID (↓)	LPIPS (↑)	BB IoU@0.5 (↑)	F1@0.5 (↑)
Real Image	–	–	0.599	0.636	0.585
Sg2Im [18]	94.9	0.096	0.416	0.305	0.197
LDM [20] (CLIP-cond.)	38.6	0.033	0.455	0.310	0.169
SGDiff [10] ($\omega = 2.0$)	42.0	0.033	0.449	0.372	0.224
SurGrID ($\omega = 2.0$)	**26.6**	**0.019**	**0.456**	**0.549**	**0.424**

The bold values in Table represent the best obtained scores on the respective metrics

Table 2 Guidance scale and embedding pre-training ablation

Method	FID (↓)	KID (↓)	LPIPS (↑)	BB IoU@0.5 (↑)	F1@0.5 (↑)
Real Image	–	–	0.599	0.636	0.585
SurGrID (Local only)	93.2	0.083	0.313	0.247	0.169
SurGrID (Global only)	120.2	0.111	0.536	0.199	0.128
SurGrID ($\omega = 2.0$)	**26.6**	**0.019**	**0.456**	**0.549**	**0.424**
SGDiff [10] ($\omega = 1.0$)	41.7	0.033	0.449	0.368	0.225
SGDiff [10] ($\omega = 2.0$)	42.0	0.033	0.449	0.372	0.224
SGDiff [10] ($\omega = 3.0$)	42.1	0.034	0.449	0.368	0.223
SGDiff [10] ($\omega = 4.0$)	43.0	0.036	0.448	0.366	0.222
SGDiff [10] ($\omega = 5.0$)	42.2	0.033	0.449	0.367	0.222
SurGrID ($\omega = 1.0$)	**26.1**	**0.019**	0.455	**0.551**	0.423
SurGrID ($\omega = 2.0$)	26.6	**0.019**	0.456	0.549	**0.424**
SurGrID ($\omega = 3.0$)	47.3	0.041	0.436	0.465	0.314
SurGrID ($\omega = 4.0$)	81.7	0.074	0.452	0.348	0.203
SurGrID ($\omega = 5.0$)	114.0	0.114	**0.464**	0.300	0.162

The bold values in Table represent the best obtained scores on the respective metrics

thesised images reliably show high fidelity and adhere to the input SG.

Clinical expert assessment

We also conduct a visual assessment user study involving three ophthalmologists with at least two years of experience to evaluate the quality of our generative model. For simple and consistent statistical evaluation, we provide them with a graphical user interface (GUI) as displayed in Appendix Section D. The GUI allows for loading ground-truth graphs along with the ground-truth image. The graph's nodes can be moved, deleted, or have their class changed. We instruct our participants to load four different ground-truth graphs and sequentially perform the following actions on each: First, participants are instructed to generate a batch of four samples from the ground-truth SG without modifications. We request them to score the samples' realism and coherence with the graph input using a Likert scale of 1 to 7. Here, 1 represents *not realistic/coherent at all*, and 7 indicates *totally realistic/coherent*. Second, the participants are requested to spatially move nodes in the canvas and again judge the synthesised samples. Third, participants change the class of one of the instrument nodes and judge the generated images. Lastly, participants are instructed to remove one of the instruments or miscellaneous classes and judge the synthesised image a final time. The study's average results are summarised in Table 3. Additionally, Section C in Appendix lists their summarised answers for the reasoning behind their ratings.

On average, participants took around 30 min to complete the user study. They were allowed unlimited time to assess the synthesised images, enabling them to thoroughly inspect for any subtle artefacts that might not be immediately noticeable. Despite the extended review time, participants consistently found the synthesised images realistic, with an average score of 5.82, and coherent, with an average score of 5.70, with the changes they had made to the SG.

Conclusions

We present Controllable Surgical Simulation via Scene Graph to Image Diffusion (SurGrID), the first controllable surgical simulator based on Scene Graph to Image Diffusion. We demonstrate the potential of SGs to encode semantic and spatial information of surgical scenes and as informative conditioning for synthesising new, unseen images

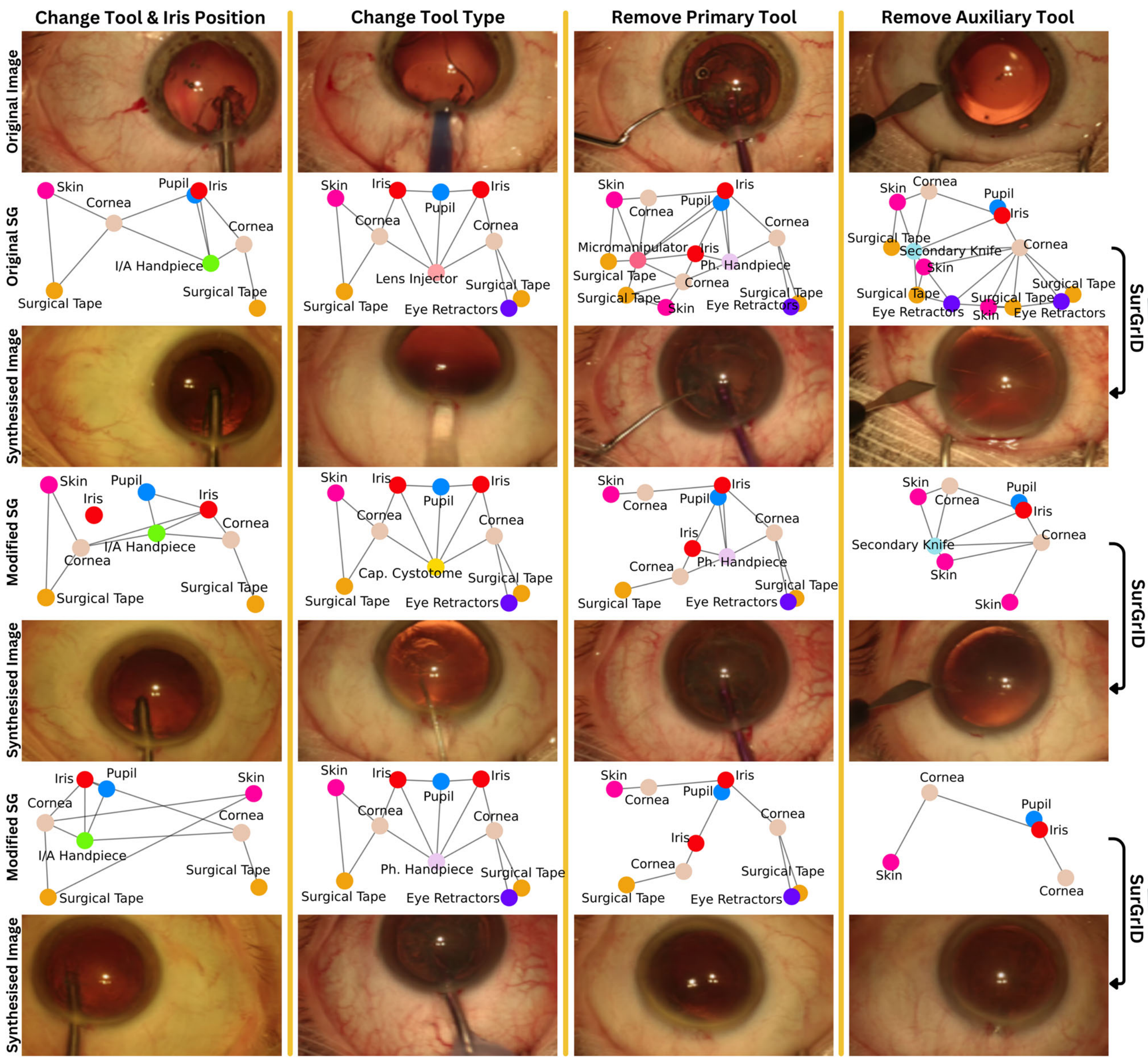

Fig. 3 Scene Graph to image generation. The top row displays the original image, followed by the corresponding SG in the second row and synthesised images from the original SGs in the third row. Readers can judge the spatial coherence of the synthesised images in the third row by comparing them to the original image. Rows four and six present modified SGs; underneath them are synthesised images conditioned on these SGs

Table 3 The abbreviation "Real." refers to the realism of the synthesised image, while "Coh." refers to the coherence with the input condition

User	Synth. from GT		Spatial Modif.		Tool Modif.		Tool Remov.	
	Real	Coh	Real	Coh	Real	Coh	Real	Coh
P1	6.5 ± 0.5	6.5 ± 1.0	6.3 ± 0.9	6.3 ± 0.9	5.3 ± 1.2	4.5 ± 1.9	6.3 ± 0.9	5.5 ± 2.3
P2	5.3 ± 0.9	5.3 ± 0.5	4.5 ± 0.5	4.3 ± 2.0	5.3 ± 0.9	5.8 ± 0.9	5.5 ± 1.2	5.5 ± 1.9
P3	6.3 ± 0.9	6.3 ± 0.9	6.5 ± 1.0	5.5 ± 0.5	6.0 ± 0.8	6.8 ± 0.5	6.3 ± 0.5	6.5 ± 0.5
Average	6.0 ± 0.9	6.0 ± 0.9	5.8 ± 1.2	5.3 ± 1.4	5.5 ± 1.0	5.7 ± 1.4	6.0 ± 0.9	5.8 ± 1.6

with Denoising Diffusion Models. We also demonstrate that by interactively modifying the SGs, changes are directly reflected in the generated image, enabling precise control over the generation process. This overcomes the limitations of text prompts, which lack precision, and segmentation masks, which are difficult to modify and interact with. In our user study, surgeons verified the generated images to be realistic and coherent with the changes made to the Scene Graphs, highlighting SurGrID's substantial potential for realistic surgical simulation controlled by scene graphs. Our method surpasses state-of-the-art techniques in image quality and coherence to the graph input, which we demonstrate quantitatively and qualitatively. This paves the way for photorealistic surgical simulations that are trainable from real surgical videos while retaining the high controllability needed for surgical simulation.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11548-025-03397-y.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been partially funded by the German Federal Ministry of Education and Research as part of the Software Campus programme (research grant 01|S23067).

Code availability Code will be published upon acceptance.

Declarations

Conflict of interest The authors declare no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Lee R, Raison N, Lau WY, Aydin A, Dasgupta P, Ahmed K, Haldar S (2020) A systematic review of simulation-based training tools for technical and non-technical skills in ophthalmology. Eye 34(10):1737–1759
2. Kwong A, Law SK, Kule RR, Nouri-Mahdavi K, Coleman AL, Caprioli J, Giaconi JA (2014) Long-term outcomes of resident-versus attending-performed primary trabeculectomy with mitomycin c in a united states residency program. Am J Ophthalmol 157(6):1190–1201
3. Thomsen ASS, Bach-Holm D, Kjærbo H, Højgaard-Olsen K, Subhi Y, Saleh GM, Park YS, La Cour M, Konge L (2017) Operating room performance improves after proficiency-based virtual reality cataract surgery training. Ophthalmology 124(4):524–531
4. Staropoli PC, Gregori NZ, Junk AK, Galor A, Goldhardt R, Goldhagen BE, Shi W, Feuer W (2018) Surgical simulation training reduces intraoperative cataract surgery complications among residents. Simul Healthc 13(1):11–15
5. Iliash I, Allmendinger S, Meissen F, Kühl N, Rückert D (2024) Interactive generation of laparoscopic videos with diffusion models. In: MICCAI workshop on deep generative models, Springer, pp 109–118
6. Cho J, Schmidgall S, Zakka C, Mathur M, Shad R, Hiesinger W (2024) Surgen: text-guided diffusion model for surgical video generation. arXiv:2408.14028
7. Holm F, Ghazaei G, Czempiel T, Özsoy E, Saur S, Navab N (2023) Dynamic scene graph representation for surgical video. In: ICCV workshop, pp 81–87
8. Murali A, Alapatt D, Mascagni P, Vardazaryan A, Garcia A, Okamoto N, Mutter D, Padoy N (2023) Latent graph representations for critical view of safety assessment. IEEE Trans Med Imaging 43(3):1247–1258
9. Köksal Ç, Ghazaei G, Holm F, Farshad A, Navab N (2024) Sangria: surgical video scene graph optimization for surgical workflow prediction. arXiv:2407.20214
10. Yang L, Huang Z, Song Y, Hong S, Li G, Zhang W, Cui B, Ghanem B, Yang M-H (2022) Diffusion-based scene graph to image generation with masked contrastive pre-training. arXiv:2211.11138
11. Mishra R, Subramanyam A (2024) Scene graph to image synthesis: integrating clip guidance with graph conditioning in diffusion models. arXiv:2401.14111
12. Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. NeurIPS 33:6840–6851
13. Müller-Franzes G, Niehues JM, Khader F, Arasteh ST, Haarburger C, Kuhl C, Wang T, Han T, Nolte T, Nebelung S et al (2023) A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Sci Rep 13(1):12098
14. Frisch Y, Fuchs M, Sanner A, Ucar FA, Frenzel M, Wasielica-Poslednik J, Gericke A, Wagner FM, Dratsch T, Mukhopadhyay A (2023) Synthesising rare cataract surgery samples with guided diffusion models. In: MICCAI, Springer, pp 354–364
15. Kim B, Ye JC (2022) Diffusion deformable model for 4d temporal medical image generation. In: MICCAI, Springer, pp 539–548
16. Fuchs M, Sivakumar SKR, Schöber M, Woltering N, Eich M-L, Schweizer L, Mukhopadhyay A. Harp: Unsupervised histopathology artifact restoration. In: MIDL
17. Allmendinger S, Hemmer P, Queisner M, Sauer I, Müller L, Jakubik J, Vössing M, Kühl N (2024) Navigating the synthetic realm: harnessing diffusion-based models for laparoscopic text-to-image generation. In: AI for Health Equity and Fairness: Leveraging AI to address social determinants of health, Springer, Switzerland, pp 31–46
18. Johnson J, Gupta A, Fei-Fei L (2018) Image generation from scene graphs. In: CVPR, pp 1219–1228
19. Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. Adv Neural Inf Process Syst 34:8780–8794
20. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: CVPR, pp 10684–10695
21. Farshad A, Yeganeh Y, Chi Y, Shen C, Ommer B, Navab N (2023) Scenegenie: Scene graph guided diffusion models for image synthesis. In: ICCV workshop, pp 88–98
22. Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al. (2021) Learning transferable visual models from natural language supervision. In:

International conference on machine learning. PMLR, pp 8748–8763
23. Liu J, Liu Q (2024) R3cd: scene graph to image generation with relation-aware compositional contrastive control diffusion. AAAI 38:3657–3665
24. Wu Z, Pan S, Chen F, Long G, Zhang C, Philip SY (2020) A comprehensive survey on graph neural networks. IEEE Trans Neural Netw Learn Syst 32(1):4–24
25. Esser P, Rombach R, Ommer B (2020) Taming transformers for high-resolution image synthesis. 2021 IEEE. In: CVPR, vol 10
26. Ho J, Salimans T (2022) Classifier-free diffusion guidance. arXiv:2207.12598
27. Grammatikopoulou M, Flouty E, Kadkhodamohammadi A, Quellec G, Chow A, Nehme J, Luengo I, Stoyanov D (2021) Cadis: Cataract dataset for surgical rgb-image segmentation. Med Image Anal 71:102053
28. Allen D, Vasavada A (2006) Cataract and surgery for cataract. BMJ 333(7559):128–132
29. Bińkowski M, Sutherland DJ, Arbel M, Gretton A (2018) Demystifying mmd gans. arXiv:1801.01401
30. Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018) The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR, pp 586–595
31. He K, Gkioxari G, Dollár P, Girshick R (2017) Mask r-cnn. In: ICCV, pp 2961–2969

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.