MITIGATING SURGICAL DATA IMBALANCE WITH DUAL-PREDICTION VIDEO DIFFUSION MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Surgical video datasets are essential for scene understanding, enabling procedural modeling and intra-operative support. However, these datasets are often heavily imbalanced, with rare actions and tools under-represented, which limits the robustness of downstream models. We address this challenge with *SurgiFlowVid*, a sparse and controllable video diffusion framework for generating surgical videos of under-represented classes. Our approach introduces a dual-prediction diffusion module that jointly denoises RGB frames and optical flow, providing temporal inductive biases to improve motion modeling from limited samples. In addition, a sparse visual encoder conditions the generation process on lightweight signals (e.g., sparse segmentation masks or RGB frames), enabling controllability without dense annotations. We validate our approach on three surgical datasets across tasks including action recognition, tool presence detection, and laparoscope motion prediction. Synthetic data generated by our method yields consistent gains of 10–20% over competitive baselines, establishing *SurgiFlowVid* as a promising strategy to mitigate data imbalance and advance surgical video understanding methods.

1 INTRODUCTION

Robotic-assisted minimally invasive surgery (RAMIS) has become a cornerstone of modern surgical practice, offering patients reduced trauma, faster recovery, and fewer complications (Haidegger et al., 2022; Taylor et al., 2016). However, operating using an endoscopic video feed rather than direct vision introduces challenges such as limited depth perception, reduced haptic feedback, and altered hand–eye coordination. These limitations increase both the cognitive and technical demands placed on surgeons during procedures (Sørensen et al., 2016; Dagnino & Kundrat, 2024).

The emerging field of *Surgical Data Science* seeks to address these challenges by developing computational methods that leverage the video data generated during surgery. In particular, deep learning (DL) methods could be utilized to understand the surgical scene, thereby supporting intraoperative decisions and reducing the burden on surgeons. Surgical video datasets, therefore, play a central role in enabling tasks, including surgical phase and gesture recognition (Padoy et al., 2012; Funke et al., 2025; 2019a), instrument detection and segmentation (Nwoye et al., 2022b; Kolbinger et al., 2023), tool tracking (Schmidt et al., 2024), and skill assessment (Funke et al., 2019b; Hoffmann et al., 2024). However, despite recent efforts to release annotated datasets (Nasirihaghighi et al., 2025; Ayobi et al., 2024; Psychogyios et al., 2023; Wang et al., 2022), these resources remain heavily imbalanced, with rare actions, steps, or tool usages under-represented (see Fig. 1). Such skewed distributions limit the generalization of DL models. Common approaches such as class-sampling and augmentation can increase the frequency of these samples but do not contribute to the diversity of the dataset.

The data imbalance challenge in surgical datasets have motivated increasing interest in synthetic data generation. With the advent of diffusion models (DMs) (Ho et al., 2020; Dhariwal & Nichol, 2021a), synthetic surgical images have been successfully utilized to augment real datasets, thereby reducing imbalance and enhancing downstream performance Venkatesh et al. (2025b); Frisch et al. (2023); Nwoye et al. (2025). However, extending DMs to surgical video generation remains underexplored due to the substantial demands in data and compute. While recent progress in video synthesis is promising, controllability is especially critical in the surgical domain, where specific

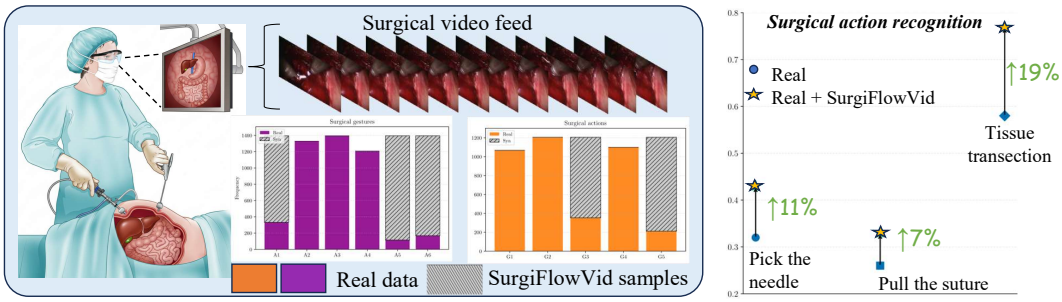

Figure 1: **Data challenge in the surgical domain**. During a laparoscopic procedure, the surgeon operates via the endoscopic video feed (video on the monitor). ML models can leverage these videos for providing guidance through surgical scene understanding. However, the datasets are *skewed* as shown in the bar plots. We aim to mitigate data imbalance with synthetic samples. The right plot shows improvements from adding samples generated from our approach (SurgiFlowVid).

tools and anatomical structures must appear in procedure-dependent contexts (e.g., laparoscopy vs. robotic surgery). Prior work often relies on dense per-frame segmentation masks to control video generation (Biagini et al., 2025; Sivakumar et al., 2025; Yeganeh et al., 2024; Iliash et al., 2024; Cho et al., 2024), but these require costly expert annotations that are rarely available. In practice, surgical datasets typically contain only sparse segmentation masks—or none at all—while under-represented classes are particularly scarce. This raises a critical question: *how can generative models improve learning for under-represented classes when only sparse or no conditional signals are available?*

To address this challenge, we propose *SurgiFlowVid* (**Surgi**cal **Flow**-inducted **Vid**eo generation), a diffusion-based framework designed to synthesize spatially and temporally coherent surgical videos of under-represented classes. We introduce a dual-prediction approach that jointly denoises RGB frames and optical flow maps, providing inductive biases to improve motion modeling from limited data. Beyond text prompts, SurgiFlowVid can condition directly on RGB frames or sparse segmentation masks, when available, via a visual encoder. While video DMs typically rely on heavy compute, our approach is tailored to the constrained settings common in healthcare, ensuring practical applicability. SurgiFlowVid generates diverse and coherent videos of under-represented classes, which we use to augment real datasets and evaluate the models across multiple datasets and downstream tasks. By tackling the challenges of data imbalance, our approach advances robust DL methods for surgical video understanding, contributing to the broader goal of improving surgical healthcare. We summarize our contributions as follows:

1. We address the critical challenge of data imbalance in surgical datasets by synthesizing video samples of under-represented classes with diffusion models, providing a principled way to augment real world datasets.
2. We introduce *SurgiFlowVid*, a surgical video diffusion framework equipped with a dual-prediction diffusion U-Net that leverages both RGB frames and optical flow to capture spatio-temporal relationships, even in the minimal available video samples of under-represented classes. In addition, a visual encoder enables conditioning on sparse conditional frames when available, removing the need for costly dense annotations.
3. We extensively evaluate the proposed framework, starting with an analysis of synthetic data attributes and extending to three surgical datasets across diverse surgical downstream tasks: action recognition, tool presence detection, and laparoscope motion prediction. The results show consistent performance gains of 10–20% over strong baselines, highlighting the effectiveness of our approach in advancing robust surgical video understanding models.

2 Related Work

Synthetic data in surgery 2D synthetic laparoscopic surgical images generated using GANs (Goodfellow et al., 2014) and diffusion models (DMs) (Dhariwal & Nichol, 2021b; Sohl-Dickstein et al., 2015) have been shown to enhance downstream tasks (Venkatesh et al., 2024; 2025b;

Frisch et al., 2023; Nwoye et al., 2025; Allmendinger et al., 2024; Martyniak et al., 2025; Pfeiffer et al., 2019). However, these approaches remain limited to static image generation and fail to capture the temporal context essential for surgical videos, which are the primary data source in real-time procedures. While diffusion models have also shown success in medical imaging domains such as MRI and CT (Dorjsembe et al., 2022; Khader et al., 2023; Zhao et al., 2025a), these modalities differ fundamentally from surgical video data.

Surgical Video Synthesis Although laparoscopic video synthesis has attracted increasing attention in recent years, its potential for addressing data imbalance in surgical tasks remains underexplored. Endora (Li et al., 2024) introduced unconditional video generation by incorporating semantic features from a DINO (Caron et al., 2021) backbone, while MedSora (Wang et al., 2024) proposed a framework based on a Mamba diffusion model. However, both approaches lacked controllability, which is crucial for generating task-specific videos that can mitigate data imbalance. Iliash et al. (2024) and SurGen (Cho et al., 2024) extended video generation by conditioning on pre-defined instrument masks to synthesize coherent surgical phases. Yet, these methods requires vast quantities of labeled real data ($\approx$ 200K videos), which restricts its applicability to well-studied procedures, such as cholecystectomy (Nwoye et al., 2022a; Twinanda et al., 2016), and prevents its generalization to less documented surgeries.

Other works, such as VISAGE (Yeganeh et al., 2024) and SG2VID Sivakumar et al., 2025, condition generation on action graphs which require curated datasets with detailed annotations and they are often unavailable for many surgical procedures. SurgSora (Chen et al., 2024a) instead conditions video synthesis on user-defined instrument trajectories, whereas Bora (Sun et al., 2024) leverages large language models (LLMs) to generate instruction prompts for controlling video generation. More recently, SurV-Gen (Venkatesh et al., 2025a) was proposed as a video diffusion framework for generating samples of rare classes. This method employs a rejection sampling strategy to filter out degenerate cases (poor consistency) of synthetic videos from a large candidate pool. Although there exists plethora of state-of-the-art video diffusion models for the natural domain (Rombach et al., 2022b; Yang et al., 2024a; Agarwal et al., 2025; Polyak et al., 2024), adapting them for the surgical domain is challenging due to the large amounts of curated video data and compute needed to train them. Additional related work is in the appendix (A).

Our approach, although closely related to SurV-Gen, introduces notable advantages: by incorporating optical flow as an inductive bias, we generate temporally coherent and plausible videos without the need for rejection sampling. Additionally, by conditioning on sparse segmentation masks or RGB frames, we achieve greater controllability and diversity in generating under-represented classes.

3 SURGIFLOWVID

Our goal is to alleviate data imbalance by generating spatially and temporally coherent surgical videos of under-represented classes, a task that is made difficult by the limited data available to model spatial and temporal dynamics accurately. To address this, we introduce *SurgiFlowVid*, which includes a multi-stage conditioninal training process built upon the SurV-Gen framework (Venkatesh et al., 2025a) with the following core modifications:

(i) *Dual-prediction diffusion U-Net:* we introduce a U-Net module that jointly predicts RGB frames and optical flow maps during training, enabling the model to capture temporal motion alongside spatial appearance which cannot be reliably inferred from RGB appearance alone.

(ii) *Sparse conditional guidance:* dense segmentation masks are rarely available in surgical datasets, and relying solely on text or label conditioning provides weak guidance. Instead, we design a sparse visual guidance encoder that conditions the diffusion process on either the available sparse segmentation masks or the RGB frames from the input video. Our model supports both text-based unconditional generation and conditional generation with sparse masks (if available), generating under-represented class samples with spatio-temporal consistency.

We first review SurV-Gen and follow it with explaining our approach. The overview of our approach is shown in Fig. 2.

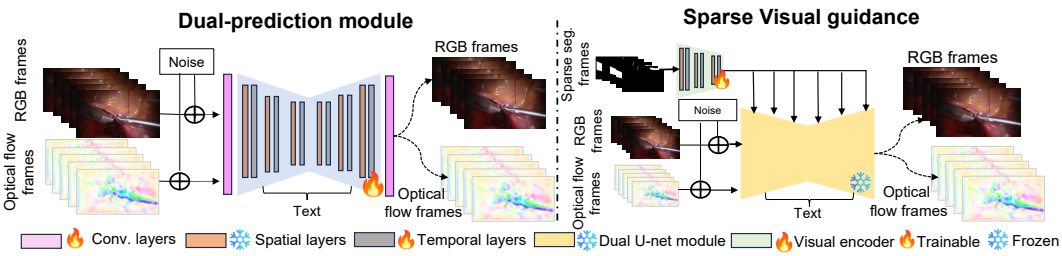

Figure 2: **SurgiFlowVid approach.**The dual-prediction diffusion U-Net module reconstructs both RGB and optical flow frames from noised inputs to capture spatio-temporal dynamics from limited data. Sparse visual encoder is trained with segmentation masks (if available) or RGB frames for conditional generation; optical flow is used only during training.

(i) Surgical Video Generation We build our framework on top of the SurV-Gen model, which follows a two-stage training strategy. In the first stage, Stable Diffusion (SD) (Rombach et al., 2022a) is adopted as the base text-to-image model, where the diffusion process is performed in the latent space. An image x_0 is first encoded into z_0 via an encoder $E(x_0)$, and during the forward diffusion process z_0 is iteratively perturbed as $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon$, with $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, and $\bar{\alpha}_t = \prod_{s=0}^{t} \alpha_s$, where β_t determines the noise strength. A denoising network $\epsilon_\theta(\cdot)$ is trained to reverse this process by minimizing the reconstruction loss

$$\mathcal{L} = \mathbb{E}_{E(x_0), y, \epsilon\sim\mathcal{N}(0,I), t}\left[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2\right], \tag{1}$$

where y denotes the text prompt associated with x_0. In the second stage, the fine-tuned spatial layers of the SD model is extended to operate directly on surgical video sequences. Temporal transformer blocks (Vaswani et al., 2017) are inserted after each spatial block while keeping spatial layers frozen, thereby focusing the training on temporal dynamics. Given a video tensor $v \in \mathbb{R}^{b\times c\times f\times h\times w}$, where b is the batch size, f the number of frames, h, w and c are the height, width and channel dimensions respectively, the temporal layers reshape v to $(bhw) \times f \times c$ and apply self-attention: $v_{\text{out}} = \text{Softmax}\left(\frac{QK^T}{\sqrt{c}}\right) V$ with $Q = v_{\text{in}} W_Q$, $K = v_{\text{in}} W_K$, and $V = v_{\text{in}} W_V$ as the query, key, and value projections. Cross-frame attention captures motion dynamics, but relying solely on it—or on text and label conditioning—is insufficient to model tool and tissue motion.

(ii) Dual-prediction module In our approach, we modify the U-net such that optical flow, p, is taken as an input along with input tensor, v. Given two consecutive frames $v_1, v_2 \in \mathbb{R}^{3\times H\times W}$, the optical flow is computed as $D_t(v_1, v_2) = (d_1, d_2)$, which encodes the pixel displacement at location (v_1, v_2). We convert D_t into an RGB image by computing a normalized magnitude $r(v_1, v_2)$ and angle θ:

$$r(v_1, v_2) = \frac{\sqrt{\widehat{d}_1^2 + \widehat{d}_2^2}}{\|D_t\|_{\max} + \varepsilon}, \qquad \theta(v_1, v_2) = \frac{1}{\pi}\,\text{atan2}\big(\widehat{d}_2,\, \widehat{d}_1\big),$$

where $\widehat{d}_1, \widehat{d}_2$ denote the normalized flow components and $\varepsilon > 0$ ensures numerical stability. The angle θ is mapped to a color, while the magnitude r attenuates this color to produce the RGB encoding resulting in the flow tensor $p^{c\times(f-1)\times h\times w}$. We define the *dual-prediction* diffusion U-Net by modifying its input and output layers to process RGB frames and optical flow jointly. Specifically, the first layer is adapted to accommodate both tensors, v and p, while the final layer is modified to predict both RGB and flow frames. These layers are trained together with the temporal attention layers using the loss function (L) defined as,

$$\mathcal{L} = \mathbb{E}_{E(x_0), y, \epsilon\sim\mathcal{N}(0,I), t}\Big[\|\epsilon - \epsilon_\theta(z_t, t, y)\|_2^2 + \lambda_p \|\epsilon - \epsilon_\theta(z_p, t, y)\|_2^2\Big], \tag{2}$$

where z_p is the noised optical flow frames and λ_p is a weighting parameter. The model jointly denoises each chunk of RGB and flow frames. We freeze the spatial layers in this stage and optical flow is used solely as a training-time signal.

(iii) Sparse visual guidance To incorporate conditional guidance, we extend the sparse condition encoder proposed in SparseCtrl (Guo et al., 2024), which propagates sparse signals (e.g., frames) across time using spatial and temporal layers to improve consistency between conditioned and unconditioned frames. In our framework, we integrate the dual-prediction U-net and redefine the *sparse visual encoder* as a lightweight module that encodes only the sparse conditional frames. The U-Net backbone is frozen, and only the visual encoder (SVE) is optimized using the loss in Eq. 2. By incorporating optical flow into the loss, we explicitly supervise both motion and structure, allowing the model to move beyond appearance propagation alone, thereby reducing data requirements and improving robustness. Formally, given sparse conditional signals $s_s \in \mathbb{R}^{3\times H\times W}$ (e.g., RGB frame or segmentation mask) and a binary mask $m \in \{0,1\}^{1\times H\times W}$ indicating whether a frame is conditioned, the sparse encoder input is constructed as $\hat{c} = [\, s_s \,\|\, m \,]$ where $\|$ denotes channel-wise concatenation. This design offers flexibility by enabling diverse conditional inputs to guide the generation process. At inference, sparse frames from the real dataset are selected and assigned specific temporal indices. The guidance module conditions generation by injecting the encoded sparse frames at their respective positions and using cross-attention to propagate their spatial structure to adjacent frames. Optical flow is not used during the sampling stage.

4 EXPERIMENTS

In this section, we outline our experimental setup, evaluation schemes and the downstream tasks we evaluate the generated synthetic datasets. We focus particularly on the under-represented classes, and generate videos of such classes to match their instances to the well represented ones. We define an under-represented class at the clip level as one with a high imbalance ratio (see Sec. C for details).

Datasets (i) **SAR-RARP50** consists of radical prostatectomy (robotic) videos from 50 patients, with a split of 35, 5, and 10 patients for training, validation, and test sets, respectively (Psychogyios et al., 2023). The annotated surgical actions include: picking up the needle (A1), positioning the needle (A2), pushing the needle through tissue (A3), pulling the needle (A4), cutting the suture (A5), tying the knot (A6), and returning the needle (A7). Since action A6 occurs only once in the test set, it is omitted from evaluation. The under-represented classes in this dataset are A1, A5, and A7. The primary task involves recognizing the surgical action at time t given a video clip. In addition, segmentation masks are available for nine classes collected at 1fps. Using these masks, we construct the task of surgical tool presence detection, where the objective is to identify which instruments are present in a given surgical video.
(ii) **GraSP** includes robotic prostatectomy procedures (Ayobi et al., 2024). It consists of 13 patients with a two-fold cross-validation setup, where five patients are held out for testing. The dataset contains annotations for 20 different surgical actions. For this study, we focus on a subset of five actions: pulling the suture (G1), tying the suture (G2), cutting the suture (G3), cutting between the prostate and bladder neck (G4), and identifying the iliac artery (G5). All classes except G5 are under-represented. Instrument annotations are also provided for six classes at every 35 secs making them sparse in nature.

(iii) **AutoLaparo** contains laparoscopic hysterectomy videos from 21 patients, with annotations describing the movements of the laparoscope (Wang et al., 2022). In total, it contains approximately 300 clips, each lasting 10 seconds, covering six motion types: up, down, left, right, zoom-in, and zoom-out. The laparoscope motion occurs precisely at the 5th second of each clip, enabling the formulation of two tasks. In the *online* recognition setting, only the first 5 seconds are provided to the model to predict the upcoming motion, which is particularly relevant for real-time applications. In the *offline* setting, the entire clip is available, and the task is to classify the laparoscope motion using full temporal context. These annotations can be used for developing automatic field-of-view control systems. We considered all movement classes to be under-represented.

Baselines For comparison, we evaluate against recent surgical video diffusion models. Endora (Li et al., 2024) is a fully transformer-based unconditional diffusion model, which we train separately on each minor class due to its lack of controllability. SurV-Gen (Venkatesh et al., 2025a) serves as a conditional baseline with both text and label guidance. We also include its rejection sampling (RS) strategy, which filters out degenerate generations and thus represents a strong reference baseline. In addition, we adapt the SparseCtrl (Guo et al., 2024) model, an effective conditional video diffusion

approach that generates videos conditioned on text and sparse conditional masks. We train the SurgiFlowVid model with only text conditioning and follow it by sparse segmentation and RGB frames. The SurV-Gen model acts as the ablation of our approach without the dual-prediction module. Additionally, we ablate the model without the SVE module. We maintain a patient specific test split and train the model only on the train split. Videos of 16-frames are generated at four frames-per-second aligning with the requirements of the downstream task. Together, these baselines span unconditional, conditional, and sparse conditional video diffusion approaches, providing a comprehensive reference for evaluating our method.

Evaluation scheme We systematically structure our experimental design into three parts to evaluate the role of synthetic data in addressing class imbalance.

(i) **Synthetic data attributes:** We analyze which attributes of synthetic data are essential for improving downstream performance. To this end, we conduct controlled experiments on the surgical action recognition task. First, we *duplicate* the training set and train for the same number of epochs to evaluate whether performance gains arise from true data diversity rather than simple repetition. Second, to assess the effects of *spatial* and *temporal* consistency, we simulate degraded data by applying elastic deformations and noise to video frames (disrupting spatial structure) and by shuffling frames (disrupting temporal order). Third, we evaluate the effect of *sparse conditioning* by constructing videos from only sparse frames and examining their impact on downstream performance.

(ii) **Class modeling:** We investigate whether synthetic data is more effective when all under-represented classes are modeled jointly or when each class is modeled separately.

(iii) **Downstream tasks:** We evaluate the effect of synthetic data on three surgical downstream applications: surgical action recognition, surgical tool presence detection, and laparoscope motion prediction.

Downstream models For surgical action (step) recognition, we employ the MViT-v2 (Li et al., 2022) model, which has shown strong performance on the SAR-RARP50 dataset and we report the averaged video-wise Jaccard index per class. The TAPIS model was used for the GraSP dataset, which incorporates an MViT backbone, and evaluate performance using mean average precision (mAP) averaged across videos, as described in Ayobi et al. (2024). For surgical tool presence detection, the Swin Transformer (base) (Liu et al., 2021) was opted in a multi-label classification setting, reporting the Dice score as the evaluation metric. Finally, for laparoscope motion recognition, we utilize a ResNet3D (Hara et al., 2017) model to classify motion categories from input clips, with mean F1 score as the metric. We apply inverse frequency balancing with video frame augmentation only on the real datasets during training. Especially, we add synthetic videos of under-represented to the real dataset and leave the well balanced classes undisturbed. Each model is run with three different seeds, and we report the mean and standard deviation across videos. These model choices ensure fair and robust state-of-the-art baselines for video understanding tasks. Please refer to the appendix for details on model training(D) and additional experiments and evaluations(B).

5 Results & Discussion

Synthetic data attributes Our evaluation of different synthetic data attributes for under-represented classes in the SAR-RARP50 dataset is in Table 1. Readers can refer to the suppl. for additional results (Sec. B.1). Merely duplicating the training set does not improve performance, as it fails to introduce additional sample diversity. Frame shuffling causes a slight decline in performance, underscoring the importance of temporal consistency in video-based tasks. Similarly, injecting noise into frames or conditioning only on sparse frames results in a more substantial drop of about 3–5%. Together, these findings reveal three key aspects: (i)

Table 1: **Attributes of synthetic** data experiment on the under-represented classes of the SAR-RARP50 dataset.

Method	A1	A5	A7
Real	$0.32_{\pm 0.19}$	$0.10_{\pm 0.04}$	$0.32_{\pm 0.15}$
Data duplicate	$0.32_{\pm 0.17}$	$0.11_{\pm 0.02}$	$0.32_{\pm 0.13}$
Frame shuffle	$0.30_{\pm 0.14}$	$0.06_{\pm 0.09}$	$0.30_{\pm 0.17}$
Sparse frame	$0.28_{\pm 0.14}$	$0.05_{\pm 0.05}$	$0.29_{\pm 0.10}$
Noisy frame	$0.29_{\pm 0.14}$	$0.04_{\pm 0.05}$	$0.29_{\pm 0.10}$

Table 2: **Surgical action recognition on the SAR-RARP50 dataset**. Under-represented classes are highlighted, and Jaccard index is reported. *Ic* denotes individual class modeling, and RS indicates rejection sampling. †denotes the ablation models. Addition of synthetic samples from SurgiFlowVid indicates comprehensive gains for the under-represented classes.

Training data	Cond. type		Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
	Text	Sparse mask							
Real	–	–	$0.32_{\pm 0.19}$	$0.66_{\pm 0.09}$	$0.78_{\pm 0.10}$	$0.61_{\pm 0.09}$	$0.10_{\pm 0.04}$	$0.32_{\pm 0.15}$	$0.46_{\pm 0.08}$
Real + Endora	–	–	$0.32_{\pm 0.14}$	$0.63_{\pm 0.05}$	$0.76_{\pm 0.07}$	$0.61_{\pm 0.11}$	$0.08_{\pm 0.04}$	$0.33_{\pm 0.10}$	$0.45_{\pm 0.05}$
Real + SurV-Gen (w/o RS) †	✓	–	$0.31_{\pm 0.19}$	$0.64_{\pm 0.07}$	$0.77_{\pm 0.06}$	$0.60_{\pm 0.10}$	$0.13_{\pm 0.10}$	$0.37_{\pm 0.18}$	$0.46_{\pm 0.03}$
Real + SurV-Gen (RS)	✓	–	$0.35_{\pm 0.12}$	$0.63_{\pm 0.02}$	$0.77_{\pm 0.03}$	$0.61_{\pm 0.08}$	$0.14_{\pm 0.09}$	$0.39_{\pm 0.15}$	$0.48_{\pm 0.06}$
Real + SparseCtrl	✓	RGB	$0.36_{\pm 0.17}$	$0.65_{\pm 0.06}$	$0.78_{\pm 0.07}$	$0.64_{\pm 0.11}$	$0.09_{\pm 0.07}$	$0.40_{\pm 0.12}$	$0.48_{\pm 0.04}$
Real + SparseCtrl	✓	Seg.	$0.36_{\pm 0.14}$	$0.61_{\pm 0.12}$	$0.77_{\pm 0.07}$	$0.63_{\pm 0.11}$	$0.16_{\pm 0.11}$	$0.38_{\pm 0.17}$	$0.49_{\pm 0.04}$
Real + SurgFlowVid †	✓	–	$0.43_{\pm 0.12}$	$0.65_{\pm 0.07}$	$0.77_{\pm 0.07}$	$0.63_{\pm 0.11}$	$0.11_{\pm 0.03}$	$0.35_{\pm 0.12}$	$0.49_{\pm 0.04}$
Real + SurgFlowvid	✓	RGB	$0.36_{\pm 0.17}$	$\mathbf{0.67_{\pm 0.06}}$	$0.78_{\pm 0.08}$	$\mathbf{0.65_{\pm 0.12}}$	$0.17_{\pm 0.10}$	$0.42_{\pm 0.12}$	$0.51_{\pm 0.04}$
Real + SurgFlowVid	✓	Seg.	$\mathbf{0.44_{\pm 0.18}}$	$0.66_{\pm 0.07}$	$\mathbf{0.79_{\pm 0.08}}$	$0.64_{\pm 0.04}$	$0.18_{\pm 0.09}$	$0.42_{\pm 0.15}$	$0.52_{\pm 0.04}$
Real + SurgFlowVid (*Ic*)	✓	–	$0.37_{\pm 0.16}$	$0.65_{\pm 0.04}$	$0.77_{\pm 0.07}$	$0.61_{\pm 0.10}$	$0.14_{\pm 0.03}$	$0.42_{\pm 0.18}$	$0.49_{\pm 0.06}$
Real + SurgFlowvid (*Ic*)	✓	RGB	$0.36_{\pm 0.14}$	$0.65_{\pm 0.03}$	$\mathbf{0.79_{\pm 0.15}}$	$0.64_{\pm 0.08}$	$\mathbf{0.20_{\pm 0.09}}$	$\mathbf{0.52_{\pm 0.12}}$	$\mathbf{0.53_{\pm 0.02}}$
Real + SurgFlowVid (*Ic*)	✓	Seg.	$0.41_{\pm 0.19}$	$0.63_{\pm 0.06}$	$0.77_{\pm 0.03}$	$0.62_{\pm 0.12}$	$0.10_{\pm 0.05}$	$0.38_{\pm 0.16}$	$0.48_{\pm 0.06}$

Table 3: **Surgical step recognition on the GraSP dataset**. The best scores are in **bold** and the mAP scores are reported. Considerable performance gains are noticed for our approach with the sparse RGB frames in comparison to solely using the real dataset.

Training data	Cond. type		Pull the suture	Tie the suture	Cut the suture	Cut btw. the prostate	Identify the iliac artery	Mean.
	Text	Sparse mask						
Real	–	–	$0.26_{\pm 0.03}$	$0.44_{\pm 0.01}$	$0.43_{\pm 0.06}$	$0.72_{\pm 0.07}$	$0.52_{\pm 0.08}$	$0.47_{\pm 0.03}$
Real + Endora	–	–	$0.26_{\pm 0.02}$	$0.39_{\pm 0.02}$	$0.40_{\pm 0.05}$	$0.70_{\pm 0.01}$	$0.51_{\pm 0.03}$	$0.45_{\pm 0.04}$
Real + SurV-Gen (w/o RS) †	✓	–	$0.30_{\pm 0.01}$	$0.43_{\pm 0.02}$	$0.41_{\pm 0.09}$	$0.71_{\pm 0.04}$	$0.57_{\pm 0.07}$	$0.48_{\pm 0.03}$
Real + SurV-Gen (RS)	✓	–	$0.30_{\pm 0.02}$	$0.44_{\pm 0.03}$	$0.42_{\pm 0.09}$	$0.73_{\pm 0.02}$	$0.58_{\pm 0.04}$	$0.49_{\pm 0.02}$
Real + SparseCtrl	✓	RGB	$0.27_{\pm 0.01}$	$0.43_{\pm 0.01}$	$0.40_{\pm 0.09}$	$0.71_{\pm 0.04}$	$0.55_{\pm 0.04}$	$0.46_{\pm 0.04}$
Real + SurgFlowVid †	✓	–	$0.30_{\pm 0.01}$	$0.43_{\pm 0.03}$	$0.44_{\pm 0.09}$	$0.69_{\pm 0.04}$	$0.60_{\pm 0.07}$	$0.49_{\pm 0.04}$
Real + SurgFlowvid	✓	RGB	$\mathbf{0.33_{\pm 0.01}}$	$\mathbf{0.48_{\pm 0.02}}$	$\mathbf{0.47_{\pm 0.01}}$	$\mathbf{0.74_{\pm 0.02}}$	$0.60_{\pm 0.05}$	$\mathbf{0.52_{\pm 0.04}}$
Real + SurgFlowVid (*Ic*)	✓	–	$0.31_{\pm 0.04}$	$0.41_{\pm 0.03}$	$0.42_{\pm 0.04}$	$0.72_{\pm 0.04}$	$\mathbf{0.61_{\pm 0.05}}$	$0.49_{\pm 0.03}$
Real + SurgFlowvid (*Ic*)	✓	RGB	$0.31_{\pm 0.01}$	$0.45_{\pm 0.01}$	$0.43_{\pm 0.03}$	$0.72_{\pm 0.02}$	$0.55_{\pm 0.05}$	$0.50_{\pm 0.02}$

synthetic data must not simply replicate the training set, but rather provide *data diversity*, (ii) maintaining *temporal consistency* is critical, and (iii) preserving *spatial structure* is essential to sustain downstream performance. Overall, this analysis underlines that downstream tasks inherently require both spatial and temporal consistency, and synthetic data must therefore satisfy both to be effective.

Surgical action recognition. (i) *SAR-RARP50*: The results of surgical action recognition task is reported in Tab. 2. The SurV-Gen model with rejection sampling achieves better performance on under-represented classes compared to using synthetic samples directly, suggesting that its gains stem primarily from the sampling strategy rather than the generative model itself. Synthetic samples from SparseCtrl improves scores across all three underrepresented classes. Our approach, *SurgiFlowVid*, even with text-only conditioning, yields performance improvements in two out of the three under-represented classes, with gains in the range of 3–11%. Adding conditional masks further enhances performance across all classes, with SurgiFlowVid conditioned on segmentation masks achieving improvements of 12%, 8%, and 10% were noticed with for the under-represented classes. Performance gains are also observed in well-balanced classes, which we attribute to the mutual dependencies among actions. For instance, augmenting data for the "picking the needle" class may indirectly benefit "positioning the needle" class, as the latter can often follow in the surgical workflow. Another noteworthy observation is that modeling each class individually produces a substantial improvement in mean performance, reaching 0.53 compared to 0.46 with real data alone. Particularly notable is the nearly 20% gain for A7, obtained with synthetic samples from SurgiFlowVid (RGB-frame) combined with individual-class training.

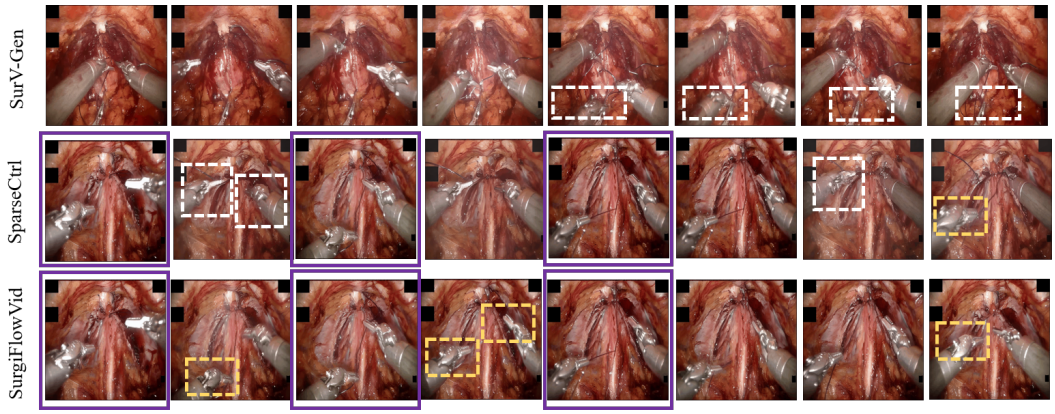

Figure 3: **Qualitative results** of the action "tie the suture." Purple boxes denote the sparse RGB conditioning frames. Suprious tools are generated in SurV-Gen (white box, row 1), while SparseCtrl alters tool types compared to the conditioning frames (white box, row 2), reflecting limited spatial consistency. SurgiFlowVid preserves both spatial and temporal structure, with consistent tools maintained across generated frames (yellow boxes, row 3).

(ii) *GraSP*: The effect of adding synthetic samples on the GraSP dataset is shown in Tab. 3 and the qualitative results are shown in Fig. 3. Incorporating samples from SurV-Gen yields small performance gains, whereas adding data generated by Endora (the unconditional baseline) or SparseCtrl with RGB-frame conditioning results in a decline in mAP score. By contrast, our method, SurgiFlowVid, achieves improvements in two of the four underrepresented classes even with text conditioning. Furthermore, with sparse segmentation masks SurgiFlowVid achieves performance gains across all under-represented classes. These results highlight the combined value of our dual-prediction and spar encoder modules, which enables the model to learn spatio-temporal relationships from limited data more effectively.

Table 4: **Surgical tool presence detection on SAR-RARP50 dataset**. Our approach with seg. conditioning outperforms the baseline across all seven tool categories.

Training data	Tool clasper	Tool wrist	Tool shaft	Suturing needle	Thread	Suction tool	Needle holder	Clamps	Catheter	Mean
Real	$0.85_{\pm 0.10}$	$0.84_{\pm 0.09}$	$0.88_{\pm 0.07}$	$0.70_{\pm 0.15}$	$0.75_{\pm 0.12}$	$0.69_{\pm 0.11}$	$0.66_{\pm 0.07}$	$0.44_{\pm 0.11}$	$0.46_{\pm 0.08}$	$0.69_{\pm 0.06}$
Real + SparseCtrl(Seg)	$0.87_{\pm 0.11}$	$0.83_{\pm 0.05}$	$\mathbf{0.89_{\pm 0.06}}$	$0.73_{\pm 0.12}$	$0.80_{\pm 0.13}$	$\mathbf{0.79_{\pm 0.10}}$	$0.74_{\pm 0.09}$	$0.69_{\pm 0.08}$	$0.50_{\pm 0.12}$	$0.74_{\pm 0.03}$
Real + SurgFlowVid(Seg)	$\mathbf{0.88_{\pm 0.09}}$	$\mathbf{0.85_{\pm 0.07}}$	$0.88_{\pm 0.10}$	$\mathbf{0.75_{\pm 0.11}}$	$\mathbf{0.81_{\pm 0.09}}$	$0.78_{\pm 0.15}$	$\mathbf{0.75_{\pm 0.04}}$	$\mathbf{0.73_{\pm 0.10}}$	$\mathbf{0.59_{\pm 0.05}}$	$\mathbf{0.79_{\pm 0.04}}$

Table 5: **Surgical tool presence detection on GraSP dataset**. Combining synthetic data from SurgiFlowVid yields marked improvements in dice scores.

Training data	Bipolar forceps	L.needle driver	Mono curved scissors	Prograsp forceps	Suction inst.	Clip applier	Laparoscopic inst.	Mean
Real	$0.94_{\pm 0.01}$	$0.56_{\pm 0.03}$	$0.95_{\pm 0.02}$	$0.72_{\pm 0.02}$	$0.71_{\pm 0.03}$	$0.34_{\pm 0.09}$	$0.56_{\pm 0.04}$	$0.68_{\pm 0.10}$
Real + SparseCtrl(Seg)	$\mathbf{0.95_{\pm 0.02}}$	$0.56_{\pm 0.02}$	$0.97_{\pm 0.01}$	$0.75_{\pm 0.03}$	$\mathbf{0.74_{\pm 0.07}}$	$0.35_{\pm 0.02}$	$\mathbf{0.60_{\pm 0.05}}$	$0.70_{\pm 0.04}$
Real + SurgFlowVid(Seg)	$0.94_{\pm 0.01}$	$\mathbf{0.58_{\pm 0.02}}$	$\mathbf{0.98_{\pm 0.01}}$	$\mathbf{0.78_{\pm 0.01}}$	$0.73_{\pm 0.04}$	$\mathbf{0.37_{\pm 0.03}}$	$\mathbf{0.60_{\pm 0.02}}$	$\mathbf{0.72_{\pm 0.02}}$

Surgical tool presence detection The results of the surgical tool presence detection task on the GraSP and SAR-RARP50 datasets are shown in Tab. 4 and Tab.5, respectively. Overall, the addition of synthetic samples from generative models leads to consistent performance improvements. This trend can be explained by the fact that the generated surgical videos naturally increase the occurrence of individual tools within the training set. On SAR-RARP50, our approach, SurgiFlowVid,

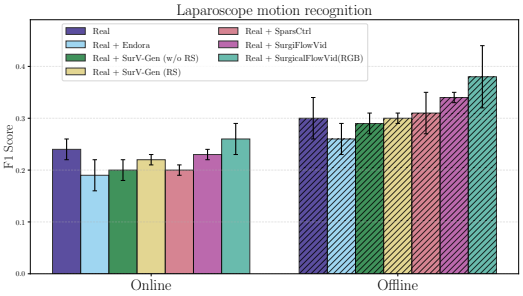

Figure 4: **Laparoscope motion prediction** in *online* (left) and *offline* (right) fashion on the AutoLapro dataset.

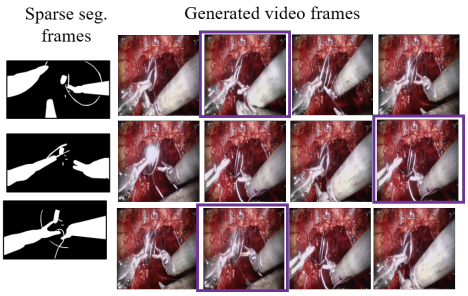

Figure 5: Tool types match the sparse seg. frames, but their position shifts, causing a failure case.

Table 6: Complexity analysis between models

Method	Trainable params. (M)	Video resolution	GPU mem (GB)	Inf. time(sec)
Endora	675	128×128	22	7.85s
SurV-Gen	435	256×256	48	6.55s
SurgiFlowVid	437	512×512	50	7.53s
SparseCtrl	453	256×256	50	10.20s
SurgiFlowVid + SVE	456	512×512	52	10.45s

Table 7: Sensitivity to different flow estimators

Method	A1	A5	A7
RAFT	$0.43_{\pm 0.12}$	$0.11_{\pm 0.03}$	$0.35_{\pm 0.12}$
UniMatch	$0.39_{\pm 0.14}$	$0.10_{\pm 0.02}$	$0.32_{\pm 0.10}$

Table 8: Sensitivity of λ_p

λ_p	A1	A5	A7
0.25	$0.32_{\pm 0.10}$	$0.09_{\pm 0.05}$	$0.31_{\pm 0.11}$
0.50	$0.43_{\pm 0.12}$	$0.11_{\pm 0.03}$	$0.35_{\pm 0.12}$
0.75	$0.44_{\pm 0.11}$	$0.13_{\pm 0.02}$	$0.33_{\pm 0.09}$
1.00	$0.40_{\pm 0.09}$	$0.10_{\pm 0.04}$	$0.31_{\pm 0.10}$

achieves a 10-point improvement over using only real data, compared to a 5-point gain from SparseCtrl. Notably, SparseCtrl's reliance on sparse conditioning yields only limited benefits, improving performance for a single under-represented class out of four. These findings further underscore the importance of generating videos with coherent spatio-temporal context for downstream tool detection models to perform effectively. For the GraSP dataset, improvements with synthetic samples are more subtle. SparseCtrl yields modest gains, while SurgiFlowVid achieves a 6% improvement for the prograsp forceps class and an overall 4% improvement across the dataset. *Together, these results highlight that SurgiFlowVid not only improves rare-class detection but also strengthens overall tool recognition performance.*

Laparoscope motion prediction Fig. 4 presents the results of laparoscope motion detection on the AutoLaparo dataset. Among the baselines, SurV-Gen (RS) achieves better performance than Endora, while SparseCtrl with RGB-frame conditioning performs best on the online recognition task. Our approach, SurgiFlowVid, already outperforms SurV-Gen with text-only conditioning, and the RGB-mask conditioned version surpasses all baselines. Similar trends are observed for the offline recognition task, where both F1 scores are higher compared to the online setting. This suggests that providing a longer temporal context enables the downstream model to classify laparoscope motion more accurately. Overall, these findings demonstrate that SurgiFlowVid can effectively adapt to smaller datasets while offering substantial benefits for developing automatic field-of-view control systems. *This highlights the practical utility of our method in developing real-time surgical assistance systems.*

Complexity analysis In comparison to Endora, both SurV-Gen and our method require fewer trainable parameters due to their two-stage training design (Tab. 6). Notably, our approach generates videos at 512×512 pixels, while SurV-Gen and SparseCtrl generate at 256×256 pixels and Endora at 128×128 pixels, with only a 2M-parameter increase from the optical-flow branch. To keep training efficient, we pre-compute optical flow offline. We do not compute the optical flow during sampling for each samples. The overhead is small relative to the substantial performance gains ($+12\%$) achieved for under-represented classes. With continued advances in GPU capabilities, these costs are expected to decline further.

Sensitivity analysis To evaluate the sensitivity of our approach to the choice of optical-flow estimator, we replaced RAFT with UniMatch (Xu et al., 2023). As shown in Tab. 8, RAFT yields slightly higher performance, which we attribute to its multi-stage refinement process. Overall, the differences are small, indicating that our method is compatible with different flow models. We also

varied the parameter λ_p to assess its influence on the training objective (Eq. 2). The best performance was observed for values between 0.5 and 0.75 (Tab 7). Based on preliminary experiments conducted on a smaller subset, we selected $\lambda_p = 0.65$ for all experiments.

Limitations While our approach demonstrates strong performance gains for under-represented classes, it has certain limitations. In this work, we focus on short temporal–context surgical tasks and therefore generate clips of about four seconds. For longer-horizon tasks such as surgical phase recognition, our framework can be extended with autoregressive generation strategies similar to FlowVid (Liang et al., 2024), requiring only minimal modifications to the training setup. Additionally, the sparsity of segmentation frames can occasionally result in incorrect tool positioning (see Fig. 5). Our method shows better instrument overlap (Sec. B.9), but additional gains can be achieved by incorporating tool kinematics or test-time correction, which we leave for future work.

6 Conclusion

In this work, we addressed the critical challenge of data imbalance in surgical datasets by generating synthetic video samples of under-represented classes with our proposed framework, *SurgiFlowVid*. The framework generates spatially and temporally coherent videos through a dual-prediction diffusion U-Net that jointly models RGB frames and optical flow, while a sparse visual encoder enables controllable generation using only the limited conditional signals typically available in surgical datasets. Extensive experiments across three datasets and downstream tasks—surgical action recognition, tool presence detection, and laparoscope motion prediction—demonstrate consistent improvements over strong baselines. By bridging advances in machine learning with the needs of surgical data science, this work helps address the scarcity of data on rare events and moves toward more robust surgical video understanding models.

7 Reproducibility Statement

All the information such as models, hyper-parameters and datasets needed to reproduce this work has been included in the appendix.

8 Ethics Statement

All datasets used in this study are publicly available. Any additional internal datasets were used only after obtaining the necessary institutional data use approvals. Large Language Models (LLMs) were used exclusively to refine the manuscript, including correcting grammatical errors and improving readability. No LLMs were used for data analysis, interpretation, or the generation of results.

References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Simeon Allmendinger, Patrick Hemmer, Moritz Queisner, Igor Sauer, Leopold Müller, Johannes Jakubik, Michael Vössing, and Niklas Kühl. 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*, pp. 31–46. Springer, 2024.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Nicolás Ayobi, Santiago Rodríguez, Alejandra Pérez, Isabela Hernández, Nicolás Aparicio, Eugénie Dessevres, Sebastián Peña, Jessica Santander, Juan Ignacio Caicedo, Nicolás Fernández, et al. Pixel-wise recognition for holistic surgical scene understanding. *arXiv preprint arXiv:2401.11174*, 2024.

Samir Brahim Belhaouari, Ashhadul Islam, Khelil Kassoul, Ala Al-Fuqaha, and Abdesselam Bouzerdoum. Oversampling techniques for imbalanced data in regression. *Expert systems with applications*, 252:124118, 2024.

Diego Biagini, Nassir Navab, and Azade Farshad. Hierasurg: Hierarchy-aware diffusion model for surgical video generation. *arXiv preprint arXiv:2506.21287*, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Hila Chefer, Uriel Singer, Amit Zohar, Yuval Kirstain, Adam Polyak, Yaniv Taigman, Lior Wolf, and Shelly Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models. *arXiv preprint arXiv:2502.02492*, 2025.

Tong Chen, Shuya Yang, Junyi Wang, Long Bai, Hongliang Ren, and Luping Zhou. Surgsora: Decoupled rgbd-flow diffusion model for controllable surgical video generation. *arXiv preprint arXiv:2412.14018*, 2024a.

Wuxing Chen, Kaixiang Yang, Zhiwen Yu, Yifan Shi, and CL Philip Chen. A survey on imbalanced learning: latest research, applications and future directions. *Artificial Intelligence Review*, 57(6): 137, 2024b.

Joseph Cho, Samuel Schmidgall, Cyril Zakka, Mrudang Mathur, Dhamanpreet Kaur, Rohan Shad, and William Hiesinger. Surgen: Text-guided diffusion model for surgical video generation. *arXiv preprint arXiv:2408.14028*, 2024.

Emanuele Colleoni and Danail Stoyanov. Robotic instrument segmentation with image-to-image translation. *IEEE Robotics and Automation Letters*, 6(2):935–942, 2021.

Emanuele Colleoni, Dimitris Psychogyios, Beatrice Van Amsterdam, Francisco Vasconcelos, and Danail Stoyanov. Ssis-seg: Simulation-supervised image synthesis for surgical instrument segmentation. *IEEE Transactions on Medical Imaging*, 41(11):3074–3086, 2022.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 113–123, 2019a.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical data augmentation with no separate search. *arXiv preprint arXiv:1909.13719*, 2(4):7, 2019b.

Giulio Dagnino and Dennis Kundrat. Robot-assistive minimally invasive surgery: trends and future directions. *International Journal of Intelligent Robotics and Applications*, 8(4):812–826, 2024.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021a.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021b.

Zolnamar Dorjsembe, Sodtavilan Odonchimed, and Furen Xiao. Three-dimensional medical image synthesis with denoising diffusion probabilistic models. In *Medical imaging with deep learning*, 2022.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 203–213, 2020.

Yannik Frisch, Moritz Fuchs, Antoine Sanner, Felix Anton Ucar, Marius Frenzel, Joana Wasielica-Poslednik, Adrian Gericke, Felix Mathias Wagner, Thomas Dratsch, and Anirban Mukhopadhyay. 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.

Isabel Funke, Sebastian Bodenstedt, Florian Oehme, Felix von Bechtolsheim, Jürgen Weitz, and Stefanie Speidel. Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. In *International conference on medical image computing and computer-assisted intervention*, pp. 467–475. Springer, 2019a.

Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019b.

Isabel Funke, Dominik Rivoir, Stefanie Krell, and Stefanie Speidel. Tunes: A temporal u-net with self-attention for video-based surgical phase recognition. *IEEE Transactions on Biomedical Engineering*, 2025.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *European Conference on Computer Vision*, pp. 330–348. Springer, 2024.

Tamas Haidegger, Stefanie Speidel, Danail Stoyanov, and Richard M Satava. Robot-assisted minimally invasive surgery—surgical robotics in the data age. *Proceedings of the IEEE*, 110(7): 835–846, 2022.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 3154–3160, 2017.

William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in neural information processing systems*, 35: 27953–27965, 2022.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

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

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022a.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022b.

Hanna Hoffmann, Isabel Funke, Philipp Peters, Danush Kumar Venkatesh, Jan Egger, Dominik Rivoir, Rainer Röhrig, Frank Hölzle, Sebastian Bodenstedt, Marie-Christin Willemer, et al. Aixsuture: vision-based assessment of open suturing skills. *International Journal of Computer Assisted Radiology and Surgery*, 19(6):1045–1052, 2024.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Ivan Iliash, Simeon Allmendinger, Felix Meissen, Niklas Kühl, and Daniel Rückert. Interactive generation of laparoscopic videos with diffusion models. In *MICCAI Workshop on Deep Generative Models*, pp. 109–118. Springer, 2024.

Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023.

Fiona R Kolbinger, Franziska M Rinner, Alexander C Jenke, Matthias Carstens, Stefanie Krell, Stefan Leger, Marius Distler, Jürgen Weitz, Stefanie Speidel, and Sebastian Bodenstedt. Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise–an experimental study. *International Journal of Surgery*, 109(10):2962–2974, 2023.

Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International conference on medical image computing and computer-assisted intervention*, pp. 230–240. Springer, 2024.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4804–4814, 2022.

Feng Liang, Bichen Wu, Jialiang Wang, Licheng Yu, Kunpeng Li, Yinan Zhao, Ishan Misra, Jia-Bin Huang, Peizhao Zhang, Peter Vajda, et al. Flowvid: Taming imperfect optical flows for consistent video-to-video synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8207–8216, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

Sabina Martyniak, Joanna Kaleta, Diego Dall Alba, Michal Naskrket, Szymon Plotka, and Przemyslaw Korzeniowski. Simuscope: Realistic endoscopic synthetic dataset generation through surgical simulation and diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4268–4278. IEEE, 2025.

Sahar Nasirihaghighi, Negin Ghamsarian, Leonie Peschek, Matteo Munari, Heinrich Husslein, Raphael Sznitman, and Klaus Schoeffmann. Gynsurg: A comprehensive gynecology laparoscopic surgery dataset. *arXiv preprint arXiv:2506.11356*, 2025.

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, 2022a.

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, 2022b.

Chinedu Innocent Nwoye, Rupak Bose, Kareem Elgohary, Lorenzo Arboit, Giorgio Carlino, Joël L Lavanchy, Pietro Mascagni, and Nicolas Padoy. Surgical text-to-image generation. *Pattern Recognition Letters*, 190:73–80, 2025.

Nicolas Padoy, Tobias Blum, Seyed-Ahmad Ahmadi, Hubertus Feussner, Marie-Odile Berger, and Nassir Navab. Statistical modeling and recognition of surgical workflow. *Medical image analysis*, 16(3):632–641, 2012.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 119–127. Springer, 2019.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Dimitrios Psychogyios, Emanuele Colleoni, Beatrice Van Amsterdam, Chih-Yang Li, Shu-Yu Huang, Yuchong Li, Fucang Jia, Baosheng Zou, Guotai Wang, Yang Liu, et al. Sar-rarp50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge. *arXiv preprint arXiv:2401.00496*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022a.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022b.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

Mabrouka Salmi, Dalia Atif, Diego Oliva, Ajith Abraham, and Sebastian Ventura. Handling imbalanced medical datasets: review of a decade of research. *Artificial intelligence review*, 57(10):273, 2024.

Adam Schmidt, Omid Mohareri, Simon DiMaio, Michael C Yip, and Septimiu E Salcudean. Tracking and mapping in medical computer vision: A review. *Medical Image Analysis*, 94:103131, 2024.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Ssharvien Kumar Sivakumar, Yannik Frisch, Ghazal Ghazaei, and Anirban Mukhopadhyay. Sg2vid: Scene graphs enable fine-grained control for video synthesis. *arXiv preprint arXiv:2506.03082*, 2025.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Stine Maya Dreier Sørensen, Mona Meral Savran, Lars Konge, and Flemming Bjerrum. Three-dimensional versus two-dimensional vision in laparoscopy: a systematic review. *Surgical endoscopy*, 30(1):11–23, 2016.

Weixiang Sun, Xiaocao You, Ruizhe Zheng, Zhengqing Yuan, Xiang Li, Lifang He, Quanzheng Li, and Lichao Sun. Bora: Biomedical generalist video generation model. *arXiv preprint arXiv:2407.08944*, 2024.

Russell H Taylor, Arianna Menciassi, Gabor Fichtinger, Paolo Fiorini, and Paolo Dario. Medical robotics and computer-integrated surgery. In *Springer handbook of robotics*, pp. 1657–1684. Springer, 2016.

Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pp. 402–419. Springer, 2020.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

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.

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.

Danush Kumar Venkatesh, Isabel Funke, Micha Pfeiffer, Fiona Kolbinger, Hanna Maria Schmeiser, Marius Distler, Jürgen Weitz, and Stefanie Speidel. Mission Balance: Generating Under-represented Class Samples using Video Diffusion Models . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15970. Springer Nature Switzerland, September 2025a.

Danush Kumar Venkatesh, Dominik Rivoir, Micha Pfeiffer, Fiona Kolbinger, and Stefanie Speidel. Data augmentation for surgical scene segmentation with anatomy-aware diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2280–2290. IEEE, 2025b.

Zhenbin Wang, Lei Zhang, Lituan Wang, Minjuan Zhu, and Zhenwei Zhang. Optical flow representation alignment mamba diffusion model for medical video generation. *arXiv preprint arXiv:2411.01647*, 2024.

Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 486–496. Springer, 2022.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13941–13958, 2023.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024a.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024b.

Yousef Yeganeh, Rachmadio Lazuardi, Amir Shamseddin, Emine Dari, Yash Thirani, Nassir Navab, and Azade Farshad. Visage: Video synthesis using action graphs for surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 146–156. Springer, 2024.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Can Zhao, Pengfei Guo, Dong Yang, Yucheng Tang, Yufan He, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. *arXiv preprint arXiv:2508.05772*, 2025a.

Shifang Zhao, Long Bai, Kun Yuan, Feng Li, Jieming Yu, Wenzhen Dong, Guankun Wang, Mobarakol Islam, Nicolas Padoy, Nassir Navab, et al. Rethinking data imbalance in class incremental surgical instrument segmentation. *Medical Image Analysis*, pp. 103728, 2025b.

Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

The appendix is structured as follows:

A Extended Related Work

Video diffusion models Diffusion-based video generation methods have recently demonstrated strong efficiency and scalability by operating in continuous latent spaces (Ho et al., 2020; Rombach et al., 2022a). Early work by (Ho et al., 2022b) extended pixel-space diffusion to videos using probabilistic DMs, while (Harvey et al., 2022) proposed generating sparse frames with interpolation, though limited to low-resolution synthetic datasets. Large-scale efforts such as Make-A-Video (Singer et al., 2022) and Imagen Video (Ho et al., 2022a) employ cascaded super-resolution pipelines built on DALLE-2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022), respectively, but require billions of parameters and massive compute resources. Stable Video Diffusion (Blattmann et al., 2023) has been widely adopted in the natural image/video community, while several closed-source systems—such as MovieGen (Polyak et al., 2024), Pika[1], and Gen (Runway)[2], Veo[3]—achieve high-quality generation conditioned on diverse modalities ranging from text to depth maps.

On the open-source side, AnimateDiff (Guo et al., 2023) and SparseCtrl (Guo et al., 2024) extend image diffusion models to videos, while OpenSora (Zheng et al., 2024) represents a large community-driven effort to replicate Sora[4]. The CogVideo family (Yang et al., 2024b; Hong et al., 2022) introduces expert transformer architectures for video synthesis and has been adopted in prior surgical applications (Biagini et al., 2025; Iliash et al., 2024). However, CogVideo is a 5B parameter model requiring vast datasets and heavy compute, making it impractical for limited surgical data where overfitting is a risk. We inspire our approach from the more recently proposed methods such as FlowVid (Liang et al., 2024) and VideoJam (Chefer et al., 2025). FlowVid proposed a flow warped video-to-video generation framework, wherein optical flow was used to maintain the structure of objects between frames during translation. This framework trained on a corpus of 10M videos. The primary application of this work differs from ours such that we intend to generate

[1] https://pika.art/login
[2] https://runwayml.com/
[3] https://deepmind.google/models/veo/
[4] https://openai.com/sora/

new videos with conditional signals. Secondly, VideoJam explored video prediction with a DiT-based architecture (Peebles & Xie, 2023), but its 30B parameter model was trained on 100M videos, produces only 256×256 outputs, and lacks controllability—an essential requirement for surgical applications.

In contrast, our work targets the surgical domain under constrained compute budgets, focusing on the critical issue of data imbalance. We build upon small-scale surgical video diffusion models and introduce a sparse, controllable framework tailored to generate under-represented surgical classes. To the best of our knowledge, we are the first to introduce a conditional video diffusion framework to mitigate the data imbalance issue for surgical application. While future work could explore scaling to larger models, our approach demonstrates a practical pathway toward improving surgical video understanding in realistic healthcare settings.

Data imbalance The presence of rare classes is a common challenge in real-world datasets. In classification, oversampling is frequently used to mitigate this issue by sampling under-represented classes more often during training (Belhaouari et al., 2024). Standard augmentation methods such as horizontal flipping, random resizing, and cropping are widely used, while regional dropout methods (Zhong et al., 2020) randomly remove image regions to improve robustness and generalization. More advanced strategies, including RandAugment (Cubuk et al., 2019b) and AutoAugment (Cubuk et al., 2019a), apply diverse pixel-level operations (e.g., rotation, shear, translation, color jitter) through either random selection or learned policies. Other approaches combine multiple images, such as Mixup (Zhang et al., 2017), which blends both pixel values and labels, and CutMix (Yun et al., 2019), which replaces patches from one image with regions from another, maximizing pixel efficiency while mixing labels. These augmentation strategies have been specifically proposed for image classification tasks. Readers can refer to (Chen et al., 2024b) for a detailed survey.

Within the surgical domain, class imbalance is particularly prevalent due to challenges in data collection (e.g., reliance on single-center data), the rarity of specific surgical events, and ethical or legal restrictions on data sharing (Salmi et al., 2024; Maier-Hein et al., 2017). Such imbalance often degrades the performance of downstream models. While augmentation and re-sampling strategies have been shown to improve medical imaging tasks (Salmi et al., 2024), surgical video understanding tasks lacks dedicated augmentation approaches. Prior attempts have used synthetic data, for instance via image-to-image translation, to complement real datasets for only surgical instrument segmentation tasks (Colleoni et al., 2022; Colleoni & Stoyanov, 2021; Zhao et al., 2025b). In this work, we establish a strong baseline for real datasets by combining curated image-level augmentation techniques with inverse frequency balancing, which up-weights under-represented classes. We use this strategy only during the training of real datasets. To directly assess the utility of synthetic data as a complementary augmentation strategy, we merge generated videos with real data without applying further augmentations.

B Additional results

B.1 Synthetic data attributes

The results on different aspects of synthetic data for the SAR-RARP50 dataset are presented in Tab. 9. Performance remains unchanged when the training data is merely duplicated, a trend consistent across most classes. In contrast, perturbations to either the spatial or temporal structure of the videos result in clear performance degradation. This behavior aligns with the role of the downstream model, which relies on both spatial structure (e.g., the arrangement of organs) and temporal dynamics (e.g., tissue motion and single or multi-tool interactions) to classify an action. Notably, the action class "cutting the suture," which is already highly imbalanced, suffers a substantial drop in performance when frame-level noise is introduced. Similar results were noticed for the GraSP dataset (Tab. 10). Interestingly we noticed for shuffling the frames lead to a small improvement in scores for two of the under-represented classes. This results could also be attributed to the downstream model architecture difference between the TAPIS model and the plain MViT model. However, overall these findings highlight that synthetic data cannot simply replicate training samples, nor can it exhibit spatial or temporal inconsistencies, if it is to provide meaningful benefits for downstream tasks.

Table 9: **Attributes of synthetic** data experiment on the SAR-RARP50 dataset. Merely replicating the training data does not lead to any improvement in performance. The degradation of the spatial or temporal structure leads to a decline in downstream model performance.

Training data	Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
Real	$0.32_{\pm 0.19}$	$0.66_{\pm 0.09}$	$0.78_{\pm 0.10}$	$0.61_{\pm 0.09}$	$0.10_{\pm 0.04}$	$0.32_{\pm 0.15}$	$0.46_{\pm 0.08}$
Data duplication	$0.32_{\pm 0.17}$	$0.60_{\pm 0.03}$	$0.78_{\pm 0.08}$	$0.61_{\pm 0.10}$	$0.10_{\pm 0.03}$	$0.31_{\pm 0.11}$	$0.45_{\pm 0.06}$
Frame shuffle	$0.30_{\pm 0.19}$	$0.63_{\pm 0.08}$	$0.74_{\pm 0.11}$	$0.60_{\pm 0.08}$	$0.06_{\pm 0.09}$	$0.30_{\pm 0.17}$	$0.43_{\pm 0.04}$
Sparse frame	$0.28_{\pm 0.14}$	$0.60_{\pm 0.07}$	$0.70_{\pm 0.04}$	$0.59_{\pm 0.09}$	$0.05_{\pm 0.05}$	$0.29_{\pm 0.10}$	$0.42_{\pm 0.03}$
Noisy frame	$0.29_{\pm 0.14}$	$0.62_{\pm 0.07}$	$0.76_{\pm 0.04}$	$0.60_{\pm 0.09}$	$0.04_{\pm 0.05}$	$0.29_{\pm 0.10}$	$0.43_{\pm 0.02}$

Table 10: **Attributes of synthetic** data experiment on the GraSP dataset.

Training data	Pull the suture	Tie the suture	Cut the suture	Cut btw.the prostate	Identify iliac artery	Mean.
Real	$0.26_{\pm 0.03}$	$0.44_{\pm 0.01}$	$0.43_{\pm 0.06}$	$0.72_{\pm 0.07}$	$0.52_{\pm 0.08}$	$0.46_{\pm 0.08}$
Data duplication	$0.25_{\pm 0.02}$	$0.44_{\pm 0.02}$	$0.43_{\pm 0.05}$	$0.71_{\pm 0.06}$	$0.52_{\pm 0.04}$	$0.46_{\pm 0.04}$
Frame shuffle	$0.27_{\pm 0.04}$	$0.40_{\pm 0.02}$	$0.42_{\pm 0.01}$	$0.69_{\pm 0.03}$	$0.53_{\pm 0.04}$	$0.46_{\pm 0.02}$
Sparse frame	$0.24_{\pm 0.02}$	$0.38_{\pm 0.03}$	$0.40_{\pm 0.02}$	$0.68_{\pm 0.02}$	$0.48_{\pm 0.01}$	$0.43_{\pm 0.02}$
Noisy frame	$0.20_{\pm 0.04}$	$0.35_{\pm 0.05}$	$0.34_{\pm 0.06}$	$0.66_{\pm 0.03}$	$0.46_{\pm 0.05}$	$0.40_{\pm 0.04}$

B.2 Additional Surgical action dataset

We further evaluated surgical action recognition on the GynSurg dataset (Nasirihaghighi et al., 2025), which consists of laparoscopic gynecological procedures with four annotated actions: coagulation (P1), needle passing (P2), suction/irrigation (P3), and transection (P4). The classes P3 and P4 are under-represented. Each action is provided as short 3-second video clips, making the dataset well-suited for action recognition. Importantly, this dataset differs substantially from SAR-RARP50 and GraSP in terms of anatomy, environment, tool usage, and camera motion, allowing us to demonstrate the generalizability of our approach across diverse surgical settings. We adopt the MViTv2 model as the downstream architecture.

Results are reported in Fig. 6. Synthetic samples from SparseCtrl improve performance by 8–9% for the under-represented classes. In contrast, our method with text conditioning achieves consistent gains across all four classes, raising the average Jaccard score to 0.72 compared to 0.66 with real data only. Conditioning with RGB frames yields further improvements of nearly 20 points for P3 and P4. These results highlight the advantage of combining dual-prediction with sparse visual encoding to generate synthetic videos that preserve both spatial and temporal consistency.

B.3 Model architecture

We further analyzed the impact of synthetic data using a different architecture for action recognition on SAR-RARP50. Since the MViT model is purely transformer-based, we tested whether synthetic samples introduce any architectural bias by comparing against X3D (Feichtenhofer, 2020), a lightweight 3D convolutional model with only 3M parameters (vs. 30M for MViT). The evaluation setup remained identical to previous experiments. The results are shown in Tab. 11. Compared to Tab. 2, the mean Jaccard score with real data dropped to 0.38 for X3D (vs. 0.46 for MViT), as expected given the smaller capacity of X3D.

Synthetic data from SparseCtrl led to modest improvements, while SurgiFlowVid with text conditioning provided only subtle gains. However, consistent with trends in Tab. 2, adding sparse RGB or segmentation masks as conditional signals in SurgiFlowVid yielded considerable improvements across the under-represented classes. Similar trends were noticed when we performed individual class modelling with the results shown in Tab. 12. These findings suggest that performance gains

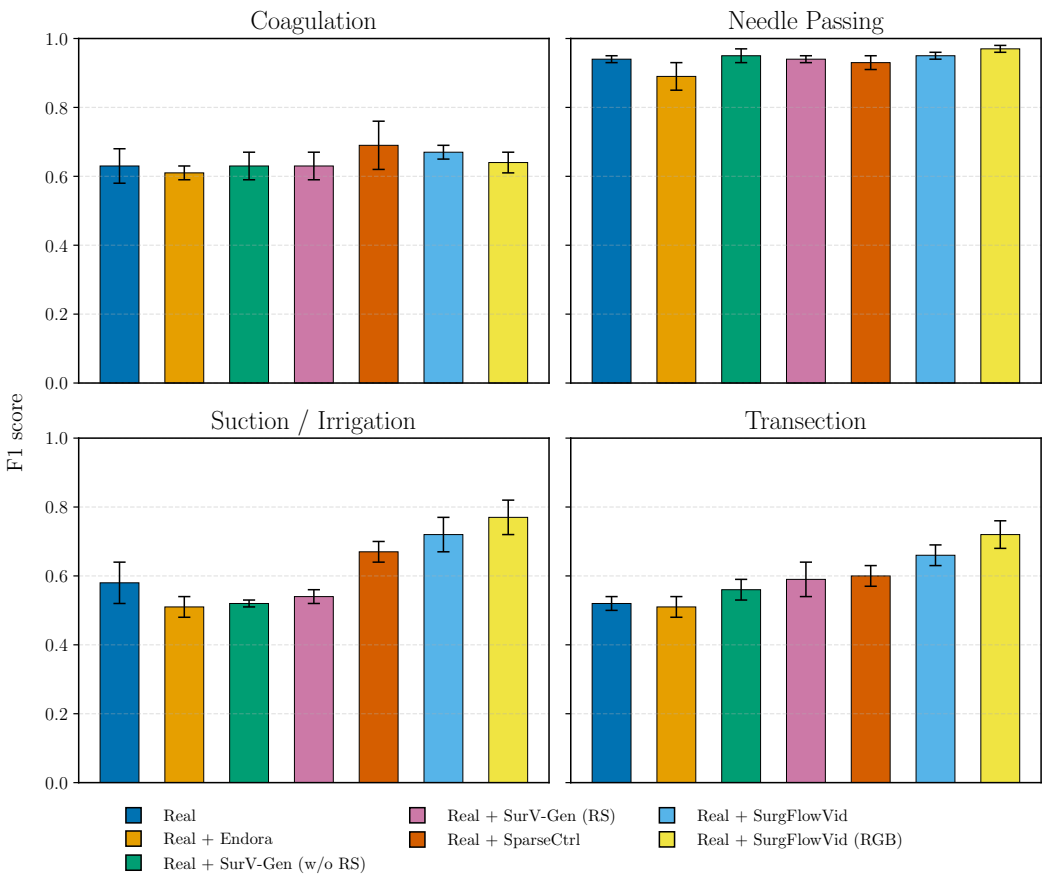

Figure 6: **Surgical action recognition** results on the GynSurg dataset, reported using the F1 score. The under-represented classes are "Suction" and "Transection". The addition of synthetic samples for both the balanced classes shows smaller improvements. However, the synthetic video samples from our approach (SurgiFlowVid) with text conditioning improves performance for both under-represented classes, while sparse RGB frame conditioning yields gains of up to 20 points in comparison to using only the real dataset.

Table 11: **Influence of model architecture**. The surgical action recognition task on the SAR-RARP50 dataset using X3D model. The Jaccard index is reported. Best and second-best scores are highlighted in blue and green, respectively. Under-represented classes are indicated with shade. We notice similar trends to Tab. 2, where the addition of samples from our approach leads to performance gains for all the under-represented classes.

Training data	Cond. type		Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
	Text	Sparse mask							
Real	–	–	$0.22_{\pm 0.01}$	$0.54_{\pm 0.08}$	$0.75_{\pm 0.07}$	$0.51_{\pm 0.13}$	$0.10_{\pm 0.02}$	$0.20_{\pm 0.12}$	$0.38_{\pm 0.06}$
Real + Endora	–	–	$0.19_{\pm 0.04}$	$0.53_{\pm 0.02}$	$0.75_{\pm 0.05}$	$0.50_{\pm 0.10}$	$0.09_{\pm 0.05}$	$0.18_{\pm 0.04}$	$0.38_{\pm 0.06}$
Real + SurV-Gen (w/o RS)	✓	–	$0.22_{\pm 0.10}$	$0.54_{\pm 0.04}$	$0.75_{\pm 0.02}$	$0.51_{\pm 0.08}$	$0.11_{\pm 0.09}$	$0.19_{\pm 0.08}$	$0.39_{\pm 0.07}$
Real + SurV-Gen (RS)	✓	–	$0.23_{\pm 0.11}$	$0.54_{\pm 0.06}$	$0.74_{\pm 0.07}$	$0.52_{\pm 0.11}$	$0.10_{\pm 0.09}$	$0.23_{\pm 0.16}$	$0.39_{\pm 0.06}$
Real + SparseCtrl	✓	RGB	$0.34_{\pm 0.17}$	$0.60_{\pm 0.07}$	$0.77_{\pm 0.08}$	$0.58_{\pm 0.09}$	$0.08_{\pm 0.05}$	$0.23_{\pm 0.16}$	$0.43_{\pm 0.03}$
Real + SparseCtrl	✓	Seg.	$0.33_{\pm 0.14}$	$0.58_{\pm 0.06}$	$0.75_{\pm 0.07}$	$0.57_{\pm 0.13}$	$0.09_{\pm 0.03}$	$0.28_{\pm 0.17}$	$0.43_{\pm 0.04}$
Real + SurgFlowVid	✓	–	$0.34_{\pm 0.13}$	$0.58_{\pm 0.06}$	$0.75_{\pm 0.05}$	$0.55_{\pm 0.13}$	$0.18_{\pm 0.09}$	$0.29_{\pm 0.12}$	$0.45_{\pm 0.04}$
Real + SurgFlowvid	✓	RGB	$0.30_{\pm 0.19}$	$0.58_{\pm 0.06}$	$0.74_{\pm 0.07}$	$0.58_{\pm 0.10}$	$0.10_{\pm 0.08}$	$0.26_{\pm 0.17}$	$0.43_{\pm 0.02}$
Real + SurgFlowVid	✓	Seg.	$0.39_{\pm 0.12}$	$0.60_{\pm 0.05}$	$0.76_{\pm 0.08}$	$0.56_{\pm 0.12}$	$0.13_{\pm 0.05}$	$0.35_{\pm 0.12}$	$0.47_{\pm 0.02}$

from synthetic data are not biased toward a specific architecture; instead, both transformer- and convolution-based models benefit from the spatial and temporal consistency encoded in synthetic videos. For the GraSP dataset, we opted to use the TAPIS model as proposed in (Ayobi et al., 2024) as this model performed in par with other convolutional architectues.

Using the features extracted from downstream models, temporal models are trained to enhance action recognition further. However, the reported performance improvements were minimal (Funke et al., 2025), and we therefore did not pursue such experiments in this study. Future work could explore this direction in greater depth, focusing on identifying which features from synthetic data are most beneficial for improving the generation process. Additionally, incorporating temporal learning strategies on top of these features may provide further gains for surgical action recognition tasks.

Table 12: **Influence of model architecture**. The surgical action recognition task on the SAR-RARP50 dataset using X3D model with *individual class modelling*. The Jaccard index is reported. We notice smaller gains for the action "cut the suture" (see Tab. 11) by modeling each of the under-represented classes separately.

Training data	Cond. type		Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
	Text	Sparse mask							
Real	–	–	$0.22_{\pm 0.01}$	$0.54_{\pm 0.08}$	$0.75_{\pm 0.07}$	$0.51_{\pm 0.13}$	$0.10_{\pm 0.02}$	$0.20_{\pm 0.12}$	$0.38_{\pm 0.06}$
Real + SurV-Gen (RS)	✓	–	$0.25_{\pm 0.12}$	$0.54_{\pm 0.03}$	$0.76_{\pm 0.09}$	$0.51_{\pm 0.09}$	$0.10_{\pm 0.13}$	$0.24_{\pm 0.18}$	$0.40_{\pm 0.05}$
Real + SparseCtrl	✓	RGB	$0.30_{\pm 0.16}$	$0.59_{\pm 0.07}$	$0.75_{\pm 0.06}$	$0.57_{\pm 0.11}$	$0.10_{\pm 0.09}$	$0.21_{\pm 0.12}$	$0.42_{\pm 0.03}$
Real + SparseCtrl	✓	Seg.	$0.30_{\pm 0.17}$	$0.57_{\pm 0.04}$	$0.76_{\pm 0.07}$	$0.57_{\pm 0.09}$	$0.20_{\pm 0.05}$	$0.37_{\pm 0.10}$	$0.46_{\pm 0.01}$
Real + SurgFlowvid	✓	RGB	$0.40_{\pm 0.16}$	$0.56_{\pm 0.02}$	$0.75_{\pm 0.04}$	$0.56_{\pm 0.16}$	$0.23_{\pm 0.13}$	$0.35_{\pm 0.15}$	$0.48_{\pm 0.02}$
Real + SurgFlowVid	✓	Seg.	$0.39_{\pm 0.11}$	$0.59_{\pm 0.04}$	$0.77_{\pm 0.03}$	$0.55_{\pm 0.10}$	$0.15_{\pm 0.06}$	$0.40_{\pm 0.10}$	$0.48_{\pm 0.05}$

B.4 VIDEO METRICS

We assess the temporal performance of the model using *Segmental F1@K* score. This metric penalizes both out-of-order predictions and over-segmentation. Segmental F1@K quantifies the temporal overlap between predicted and ground-truth segments, while being less sensitive to small boundary shifts caused by annotation noise. The metric is defined as,

$$\text{SegmentalF1@K} = \frac{2 \times (\text{Pr} \times \text{Rc})}{(\text{Pr} + \text{Rc})}, \tag{3}$$

where Pr and Rc denotes precision and recall. A prediction is considered a true positive (TP) if the IoU exceeds the threshold $T = K/100$; otherwise, it is counted as a false positive (FP). The results of the recognition task are shown in Tab.13 and Tab.14. Compared to using only the real dataset, the addition of synthetic samples leads to smaller improvements in overall performance. The addition of either RGB or segmentation conditioning lead to a similar scores of 0.37 and 0.36 respectively. Overall, the synthetic samples from SurgiFlowVid prove very beneficial for both the balanced and the under-represented classes.

Table 13: **Surgical action recognition** on the SAR-RARP50 dataset. Segmental F1 scores are reported.

Training data	Cond. type		Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
	Text	Sparse mask							
Real	–	–	$0.28_{\pm 0.17}$	$0.40_{\pm 0.16}$	$0.62_{\pm 0.18}$	$0.41_{\pm 0.14}$	$0.09_{\pm 0.08}$	$0.22_{\pm 0.18}$	$0.32_{\pm 0.06}$
Real + Endora	–	–	$0.23_{\pm 0.13}$	$0.38_{\pm 0.06}$	$0.55_{\pm 0.09}$	$0.41_{\pm 0.10}$	$0.09_{\pm 0.09}$	$0.21_{\pm 0.08}$	$0.31_{\pm 0.08}$
Real + SurV-Gen (w/o RS)	✓	–	$0.26_{\pm 0.12}$	$0.40_{\pm 0.04}$	$0.55_{\pm 0.08}$	$0.41_{\pm 0.06}$	$0.12_{\pm 0.09}$	$0.23_{\pm 0.12}$	$0.33_{\pm 0.04}$
Real + SurV-Gen (RS)	✓	–	$0.27_{\pm 0.14}$	$0.40_{\pm 0.15}$	$0.58_{\pm 0.19}$	$0.42_{\pm 0.18}$	$0.20_{\pm 0.13}$	$0.23_{\pm 0.18}$	$0.35_{\pm 0.07}$
Real + SparseCtrl	✓	RGB	$0.32_{\pm 0.20}$	$0.41_{\pm 0.16}$	$0.57_{\pm 0.17}$	$0.44_{\pm 0.15}$	$0.10_{\pm 0.09}$	$0.25_{\pm 0.11}$	$0.35_{\pm 0.03}$
Real + SurgFlowVid	✓	–	$0.27_{\pm 0.14}$	$0.40_{\pm 0.16}$	$0.57_{\pm 0.16}$	$0.43_{\pm 0.13}$	$0.13_{\pm 0.08}$	$0.16_{\pm 0.07}$	$0.33_{\pm 0.04}$
Real + SurgFlowvid	✓	RGB	$0.31_{\pm 0.17}$	$0.43_{\pm 0.17}$	$0.59_{\pm 0.16}$	$0.45_{\pm 0.10}$	$0.15_{\pm 0.04}$	$0.31_{\pm 0.12}$	$0.37_{\pm 0.03}$

Table 14: **Surgical action recognition** on the SAR-RARP50 dataset. Segmental F1 scores are reported. for seg. frame conditioning.

Training data	Cond. type		Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
	Text	Sparse mask							
Real	–	–	$0.28_{\pm 0.17}$	$0.40_{\pm 0.16}$	$0.62_{\pm 0.18}$	$0.41_{\pm 0.14}$	$0.09_{\pm 0.08}$	$0.22_{\pm 0.18}$	$0.32_{\pm 0.06}$
Real + SparseCtrl	✓	Seg	$0.33_{\pm 0.19}$	$0.43_{\pm 0.14}$	$0.60_{\pm 0.19}$	$0.44_{\pm 0.15}$	$0.12_{\pm 0.10}$	$0.20_{\pm 0.10}$	$0.35_{\pm 0.05}$
Real + SurgFlowvid	✓	Seg	$0.30_{\pm 0.14}$	$0.42_{\pm 0.16}$	$0.58_{\pm 0.14}$	$0.43_{\pm 0.13}$	$0.13_{\pm 0.08}$	$0.32_{\pm 0.11}$	$0.36_{\pm 0.02}$

B.5 Ablation on sparse frames

We conducted an ablation study to examine the effect of the number of sparse RGB frames used during generation. We hypothesized that too few frames would provide insufficient controllability, while too many would replicate training data, reducing diversity. To test this, we varied the number of conditioning frames $(1, 3, 5, 10, 12)$ and generated videos, comparing their performance against models trained solely on real data. Results are shown in Fig. 7 (all minor classes modeled jointly) and Fig. 8 (each class modeled separately). A consistent trend across both settings is that using only one frame yields performance similar to the real-only baseline, indicating limited consistency and, in some cases, degenerate generations. Conversely, conditioning on 12 of the 16 frames produced results close to the real dataset baseline, as little additional diversity was introduced. Based on these findings, we adopted a strategy of sampling 3–5 random frames from the real dataset as conditional inputs. These experiments were initially conducted with the X3D model, and the same frame distribution was subsequently applied across all experiments, including the SparseCtrl baseline.

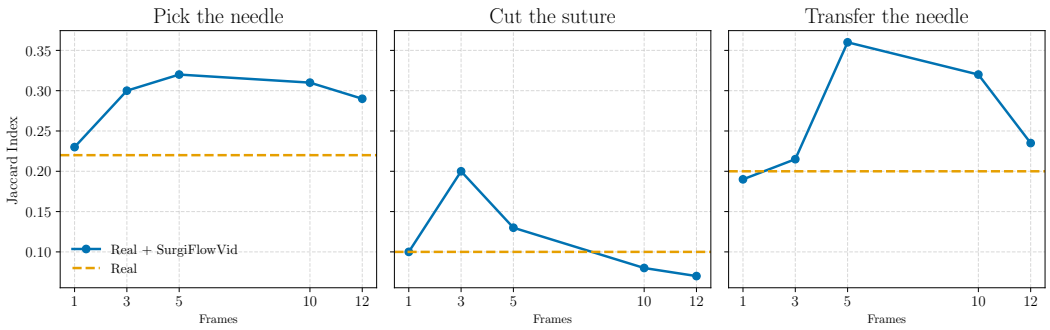

Figure 7: **Frame ablation**. The ablation on the number of sparse RGB frames on the SAR-RARP50 dataset. The results consists of using a X3D model with all the minor classes modeled together.

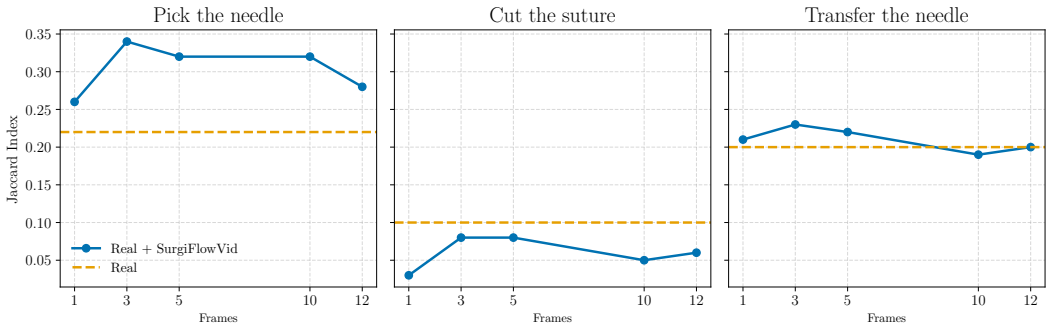

Figure 8: **Frame ablation**. The ablation on the number of sparse RGB frames on the SAR-RARP50 dataset. The results consists of using a X3D model with all the minor classes modeled separately.

B.6 Model Analysis

In this section, we analyze the model in terms of the video generation cost. The results are shown in Tab. 15. In comparison to Endora, both SurgV-Gen and our approach have lesser number of training parameters as the training is conducted in different stages. Our method, SurgiFlowVid is capable of generating videos at the resolution of 512×512 pixels whereas the baselines, SurV-Gen and SparseCtrl generates videos at 256×256 pixels and Endora at 128×128 pixels. We also train our approach at 512×512 pixels. Our framework is capable of training at lower resolutions but we opted to train them at higher resolutions as it could be helpful for the downstream task. There exists certain organs or auxillary tool structures which appears to be very small in shape. Generating videos at higher resolution can benefit these downstream models to learn these spatial structures effectively. We noticed the benefits for the classification of *catheter* and *clamps* in SAR-RARP50 dataset with synthetic videos from SurgiFlowVid (see Tab. 4). However, an analysis on the video resolution for the downstream task could shed more insights and we leave that for future work. As we generate videos at higher resolution, our approach requires a small overhead in terms of training and sampling times. We believe with the innovations in high performant GPUs these costs could be lowered drastically.

Table 15: **Model analysis**. The various parameters of the different baselines. SVE denotes the sparse visual encoder in our approach. The inference time was measured on a A100-80GB GPU.

Method	Trainable params. (M)	Video resolution	Sampling steps	Inf. time(sec)
Endora	675	128×128	50	7.85s
SurV-Gen	435	256×256	50	6.55s
SurgiFlowVid	437	512×512	50	7.53s
SparseCtrl	453	256×256	30	10.20s
SurgiFlowVid + SVE	456	512×512	30	10.45s

Table 16: **Image quality metrics**. The CLIP image score of different methods are reported here. Higher is better.

Method	SAR-RARP50			GynSurg		GraSP			
	A1	A5	A7	P3	P4	G1	G2	G3	G4
Endora	70.30	66.85	73.65	69.43	70.12	57.09	68.10	74.41	60.72
SurV-Gen	75.30	70.22	78.85	71.30	75.83	62.15	73.10	68.32	62.15
SurgiFlowVid	74.46	76.08	78.25	72.95	66.76	68.20	70.10	72.15	65.27

B.7 Image Quality Analysis

As our goal is to mitigate data imbalance, we focused primarily on generating videos of under-represented classes and evaluating them on the downstream task. We consider this approach as an effective way to directly measure the effectiveness and the usefulness of the synthetic videos. In this section, we evaluate the quality of the generated videos with the CLIP (Hessel et al., 2021) image and the LPIPS (Zhang et al., 2018) score. Both these metrics evaluate the quality of the generated frames using features from pre-trained models on large-scale natural images. The results are shown in Tab. 16 and Tab. 17. We compare our approach, SurgiFlowVid with text conditioning against Endora and SurV-Gen. We do not compute these scores for SparseCtrl or sparse visual encoder using our approach, as there already exists frames from the real dataset. The image quality varied between different classes and we did not notice a co-relation between these scores to the downstream model performance. Hence, these values should be interpreted with caution given that they are computed with pre-trained weights from models not trained on surgical images/videos.

Table 17: **Image quality metrics**. The LPIPS score of different methods are reported here. Lower is better.

Method	SAR-RARP50			GynSurg		GraSP			
	A1	A5	A7	P3	P4	G1	G2	G3	G4
Endora	0.70	0.53	0.59	0.54	0.49	0.63	0.66	0.65	0.63
SurV-Gen	0.68	0.54	0.57	0.51	0.56	0.57	0.67	0.71	0.74
SurgiFlowVid	0.66	0.56	0.52	0.49	0.50	0.51	0.60	0.74	0.72

B.8 Motion diversity

We quantify the intra-set video diversity using the VJEPA embeddings for the generated videos. Particularly, we measure the diversity within the generated videos using the video-level embeddings from the pre-trained VJEPA (Assran et al., 2025) model. The mean pairwise cosine distance (scores) is computed between the embeddings and the scores of the real dataset serve as the upper bound. A higher value indicates larger diversity within the generated videos.

Table 18: **Motion diversity**. The cosine distance between the video embedding from the VJEPA2 model is reported. Higher value indicates better diversity.

Method	SAR-RARP50				GraSP		
	A1	A5	A7	G1	G2	G3	G4
Real	0.121	0.115	0.109	0.102	0.113	0.131	0.117
Endora	0.043	0.061	0.045	0.064	0.079	0.056	0.065
SurV-Gen	0.056	0.102	0.048	0.056	0.094	0.067	0.072
SurgiFlowVid	0.113	0.110	0.101	0.091	0.083	0.106	0.110

As shown in the Tab. 18, SurgiFlowVid consistently produces higher scores which supports our claim that the proposed method generates diverse videos for challenging, under-represented classes while maintaining realism.

B.9 Mask overlap of generated videos

In Fig. 5, we observe that videos generated with sparse segmentation frames sometimes exhibit a drift in tool position relative to the provided masks. We attribute this behaviour to the high sparsity typical of surgical datasets. To further quantify this effect, we compare the mask overlap between the generated frames and the ground-truth masks. For this analysis, we train a SegFormer (Xie et al., 2021) model in a binary segmentation setting and evaluate videos produced by both SurgiFlowVid and SparseCtrl. As shown in Tab. 19, IoU scores are low for both models, reflecting the difficulty of the task, but our method leads by +8 points over SparseCtrl. This indicates that the dual-prediction design of SurgiFlowVid better preserves tool position. Nonetheless, there remains room for improvement, for example, by incorporating the strategies discussed in the limitations section.

Table 19: **Mask overlap of generated videos**.

Method	IOU ↑
SparseCtrl	0.34
SurgiFlowVid	0.42

B.10 Video augmentation baselines

In this work, we use synthetic videos as a form of data augmentation to reduce class imbalance in the real dataset. To compare against strong augmentation strategies, we conducted additional experiments using VideoMix-Spatial and TubeMix for action recognition on the SAR-RARP50 dataset. The results in Tab. 20 indicate that neither approach yielded substantial improvements, underscor-

ing the need for more targeted augmentation methods such as the generative strategy proposed here. Overall, SurgiFlowVid achieves the highest performance on the under-represented classes.

Table 20: **video augmentation baselines** experiment on the SAR-RARP50 dataset. The addition of synthetic data proves the most useful in comparison to other data augmentations.

Training data	Pick the needle	Position the needle	Push the needle	Pull the needle	Cut the suture	Return the needle	Mean.
Real + VideoMix	$0.33_{\pm 0.14}$	$0.64_{\pm 0.02}$	$0.77_{\pm 0.02}$	$0.61_{\pm 0.04}$	$0.10_{\pm 0.03}$	$0.30_{\pm 0.13}$	$0.45_{\pm 0.03}$
Real + TubeMix	$0.31_{\pm 0.17}$	$0.65_{\pm 0.03}$	$0.78_{\pm 0.05}$	$0.59_{\pm 0.02}$	$0.09_{\pm 0.05}$	$0.27_{\pm 0.16}$	$0.45_{\pm 0.04}$
Real + SurgiFlowVid	$0.44_{\pm 0.18}$	$0.66_{\pm 0.07}$	$0.79_{\pm 0.08}$	$0.64_{\pm 0.04}$	$0.18_{\pm 0.09}$	$0.42_{\pm 0.12}$	$0.52_{\pm 0.04}$

B.11 LAPAROSCOPE MOTION

In addition to the F1 score, we also computed the balanced accuracy as an additional metric. Fig. 9 shows the results on the laparoscope motion prediction task. Similar to the results seen in Fig. 4, the overall scores are higher for the the offline recognition.

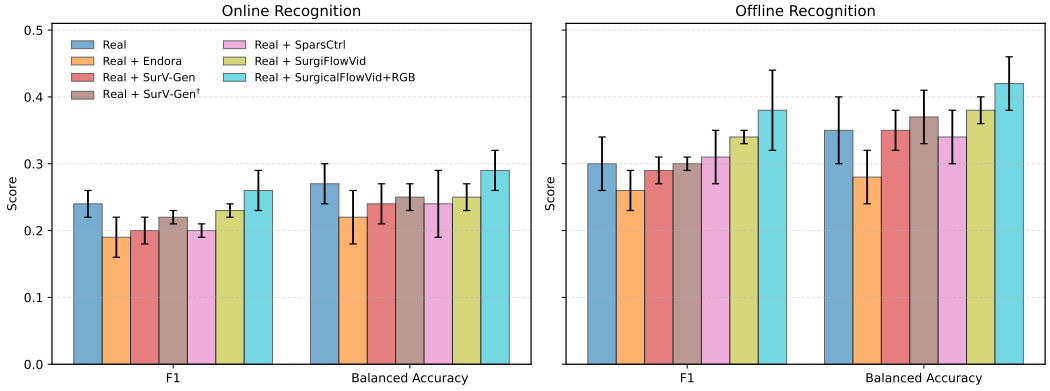

Figure 9: Laparoscope motion prediction on the Autolaparo dataset. Bars show mean score with standard deviation (error bars).

C DATASET

SAR-RARP50: The dataset consists actions annotated at 10 fps. Our initial experiments indicated this temporal frame to be very fine and hence we chose to sample the frames at 5 fps. The annotations for the surgical tools were available at 1 fps making it sparse in nature. For the sparse conditional generation, we randomly samples video frames in the range $3 - 5$ and place them in a different temporal order than the real dataset, so as to create the synthetic data as diverse as possible. For the sparse segmentation conditioning, we opted to include a minimum of 4 frames in the 16 frames video clips during training and sampling time.
GraSP: This dataset consists of annotations at both 30 and 1 fps temporal windows. As 1fps was very coarse in nature, we opted to sample frames at 5 fps from the 30 fps annotations. The segmentation annotations were available at every 35 seconds making them very sparse in nature. Based on dataset analysis, we noticed that creating video clips with at least one segmentation frame as conditioning for the under-represented samples were very challenging. Hence, we opted out of segmentation frames conditioning for the sparse visual encoder in our experiments. However, for the surgical tool presence detection task, we sampled a minimum of 4 frames around the available segmentation frame and used it as the conditioning to generate videos for this task.

The details on the addition of synthetic samples are shown in Tab. 21. We compute the imbalance ratio as the number of clips for the well balanced (most represented) class divided by the number of

clips for the other classes. For the SAR-RARP50 dataset, we chose the classes with the ratio higher than 2. Similarly, for the GRaSP and GynSurg datset, the ratio was chosen as 1.5 respectively.

Table 21: **Dataset details**. The values in the table include the total number of video clips from the training set. We add only synthetic samples to the under-represented classes to match and balance the instances with the well balanced classes.

Dataset	Step/action class	Data points in real dataset	Added syn. samples	Imbalance ratio
SAR-RARP50	Pick the needle	332	900	4.20
	Position the needle	1329	-	1.04
	Push the needle	1395	-	1.00
	Pull the needle	1208	-	1.15
	Cut the suture	115	1100	12.13
	Return the needle	168	1100	8.30
GraSP	Pull the suture	992	1600	2.82
	Tie the suture	712	1800	3.93
	Cut the suture	1213	1300	2.30
	Cut btw. the prostate	1616	1000	1.73
	Identify iliac artery	2800	-	1.00
GynSurg	Coagulation	690	-	1.29
	Needle passing	869	-	1.00
	Suction/Irrigation	267	550	3.35
	Transection	168	650	5.33

D MODEL TRAINING

D.1 DIFFUSION IMAGE PRE-TRAINING

We build upon the SurV-Gen model (Venkatesh et al., 2025a), which was initially proposed to generate synthetic samples of under-represented classes to mitigate data imbalance in surgical datasets. The framework adopts a multi-stage training procedure. In the first stage, frames are extracted from the training split of surgical videos and a 2D Stable Diffusion (SD) model (Rombach et al., 2022a) is trained. We follow the same pipeline with several modifications. Training the spatial SD directly on the limited frames from the downstream task datasets can result in overfitting, reduced diversity of generated frames, or potential data leakage. This phenomenon was observed in SurV-Gen, where synthetic augmentation yielded only marginal improvements without rejection sampling.

To address this issue, we curated an in-house dataset comprising video recordings from different surgical procedures. The dataset consists of approximately 7000 clips, each ranging from 6 to 8 minutes in length. From this collection, we extracted $\sim$ 4000 frames to train the 2D component of the model. We initialized training from the SD-v1.5 checkpoint, pre-trained on the large-scale LAION-5B dataset (Schuhmann et al., 2022), which provided a strong initialization compared to training from scratch. The model was fine-tuned for 3000 steps using the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of $1e^{-4}$, a batch size of 2, and gradient checkpointing enabled. Due to computational constraints, frames were resized from their original resolution of 1048×2048 to 512×512. For text conditioning, we employed simple prompts such as *"An image of a surgical procedure"*, with embeddings generated using the CLIP text encoder (Radford et al., 2021). This fine-tuned SD model served as the base 2D diffusion prior for any subsquent 2D diffusion models. We fine-tune this model on the downstream datasets before video diffusion training. The spatial priors are learnt during this stage.

D.2 DIFFUSION VIDEO PRE-TRAINING

Next, we focus on the video training stage. In the SurV-Gen approach, the spatial layers are frozen and only the temporal attention layers are trained during the second stage. In contrast, our framework trains the temporal layers jointly with both RGB and optical flow frames. To further improve

temporal modeling, we investigated a video pre-training strategy inspired by previous works on video diffusion models (Rombach et al., 2022b; Polyak et al., 2024). Our hypothesis is that temporal motion priors, such as the movement of tools, tissue motions andpartially tool tissue interactions can be better learned by training on the unconditional internally curated dataset, which contains diverse anatomical structures, varying illumination conditions, different endoscope motions, and a wide range of surgical tools and tool interactions. This dataset introduces substantial variability that more closely reflects real-world surgical scenarios.

To test this, we extended SurV-Gen and trained it in two ways, keeping the training recipe unchanged (i.e., only the temporal attention layers are updated). First, we trained SurV-Gen directly on the SAR-RARP50 dataset, where the 2D SD backbone was also trained on frames extracted from the same dataset. Second, we replaced the 2D SD backbone with our fine-tuned 2D model and pre-trained the temporal layers on the curated dataset of $\sim$ 7000 videos. For this, we created overlapping subsets of 3000, 5000, and 7000 videos, each containing at least 1500 new clips. The pre-trained temporal layers were then fine-tuned on SAR-RARP50.

This pre-training strategy is expected to accelerate learning of spatio-temporal representations from the limited SAR-RARP50 data. We then generated synthetic samples of under-represented classes using label guidance, following SurV-Gen, and evaluated their impact on downstream action recognition performance. The results are shown in Fig. 10.

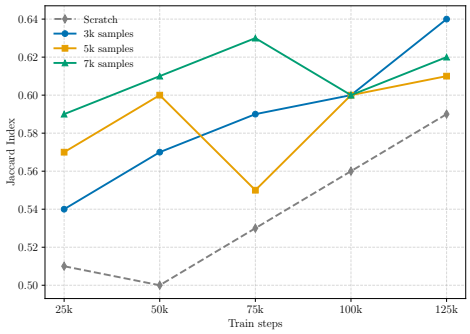

Figure 10: The results on video pre-training.

We analyzed only the three under-represented classes and report the weighted average Jaccard index of these classes. We notice that the pre-training strategy leads to higher recognition scores in comparison to using only the real dataset for the same number of training steps. We noticed smaller dips in performance for the 5k and 7k samples, which could be attributed to a distributional shift to the SAR-RARP50 dataset. On the other hand, we noticed a continuous improvement in jaccard scores for the 3k samples. Overall, these results indicate that the pre-training strategy leads to learning the spatio-temporal relationships better, such that when minimal data is available, the model can learn faster. Based on these results, we used the 2D spatial SD model and temporal attention layers pre-trained on our internal dataset as the starting checkpoints for the SurgiFlowVid training scheme.

D.3 SurgiFlowVid training

Based on these results we opted to use the temporal layers trained on our internal dataset as the pre-trained model. This offers the advantage that, the SurgiFlowVid training time reduces and also we can avoid the over-fitting of the dataset given the fact that there exists only limited training data from the downstream datasets. We fine-tune the pre-trained temporal attention layers using our proposed dual-prediction U-net module. The optical flow frames are extracted using the RAFT model (Teed & Deng, 2020). For SurgiFlowVid training, we extract clips of 16 frames at a frame rate of 5 for all the datasets. The hyperparameter details are mentioned in Tab. 22.

D.4 Downstream model training

For the action recognition task (SAR-RARP50), we used the MviT-v2 model from the SlowFast library[5]. We downsampled the videos to 224×384 pixels for training with a temporal resolution of 5 fps. Image augmentations such as PCA jitter, RGB scale shift, brightness and contrast shift, random flipping with scale cropping was used along with inverse frequency balancing during the training on the real data. For additional details on the model, readers can refer to SlowFast repo. We

[5] https://github.com/facebookresearch/SlowFast

followed the similar recipe for the GynSurg dataset. The model were trained for 150 epochs with a learning rate of $1e^4$ with the best model being chosen using a validation dataset.

For the GraSP dataset, we used the similar settings from the TAPIS model[6]. It is to be noted that we do not compare the values directly to the work from (Ayobi et al., 2024) on the GraSP dataset. This is due to the fact that the results reported from the TAPIS model have been obtained directly using the test set as the selection criteria during training. We create a separate validation set from the training set which we use as the selection criteria of the trained model. The test set is clearly separated during the training of both diffusion and downstream models to avoid any data leakage. For the combined training of real and synthetic videos, we opted for a simple and easier strategy than rejection sampling as proposed in (Venkatesh et al., 2025a). We sampled a batch of data points such that 25% of this batch consists of synthetic videos. We chose this method as it works on the fly during training and the time and effort in rejecting synthetic samples are drastically reduced.

For the surgical tool presence detection task, we used the Swin transformer model. The videos were resized to a resolution of 384×384 during training with augmentations such as RGB channel shift, scaled cropping and temporal shift. We trained the models using binary cross entropy loss with weighted sampling to include the imbalance in the surgical tools.

E Qualitative Results

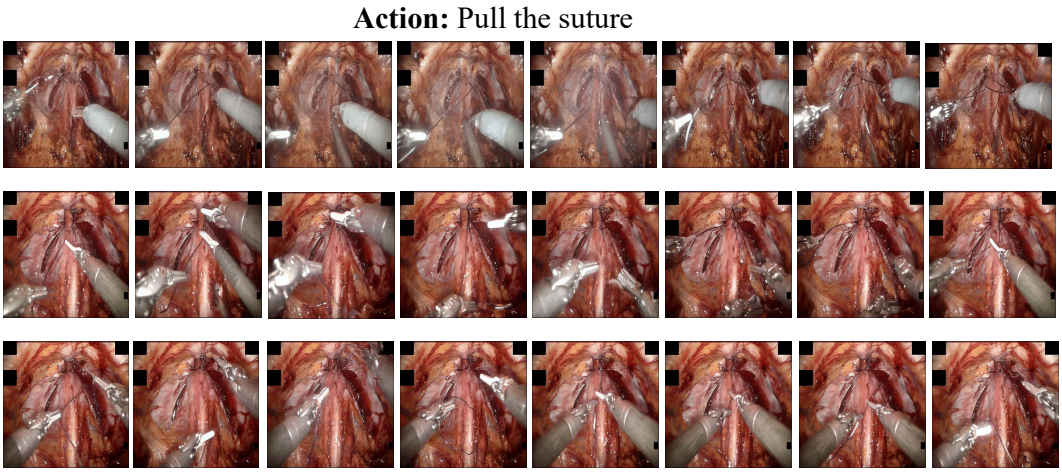

Figure 11: Results from SurgiFlowVid with text conditioning on GraSP dataset.

[6] https://github.com/BCV-Uniandes/GraSP/tree/main/TAPIS

Hyperparameter	Image fine-tuning	Video-pretraining	SurgiFlowVid training
Datatset			
No. of samples	4000	7000	Train split of the dataset
Resolution	512×512	256×256 & 512×512	512×512
Video length	-	16 frames	16 frames
Sample rate	-	5	4-5
Context length	-	16	16
Model params			
Pre-trained model	SDv-1.5	Pre-trained on internal	Pre-trained on internal
Params frozen	-	Spatial layers	Spatial layers
Temporal layers			
Depth	-	2	2
Temporal resolution	-	$[1, 2, 4, 8]$	$[1, 2, 4, 8]$
Head channels	-	16	16
No. of heads	-	8	8
Position encoding	-	sinusoidal	sinusoidal
PE dim	-	24	24
Cross attention dim	-	32	32
Act.function	-	GeLU	GeLU
Training params			
Optimizer	AdamW	AdamW	AdamW
Learning rate	$1e^{-4}$	$1e^{-5}$	$1e^{-5}$
Lr warm steps	500	5000	5000
Lr scheduler	cosine	cosine	cosine
β_1	0.9	0.9	0.94
β_2	0.999	0.999	0.995
Weight decay ω	$1e^{-2}$	-	-
Train steps	3000	125k	$75 - 125$k
Train timestep			
Diffusion step	1000	1000	1000
Noise schedule	linear	linear	linear
β_0	$1e{-4}$	0.00085	0.00085
β_T	0.02	0.012	0.012
Sampling params			
Sampler	DDPM	DDIM	DDIM
Steps	-	50	50 (30 for SVE)
CFG scale	6.5	5.5	5.0
Device requirements			
GPU-type	A100-40GB	H200-80GB	H200-140GB
No. of gpus	1	1	1

Table 22: Hyperparameters for training the 2D and the temporal attention layers of the diffusion model. SVE denotes *Sparse visual encoder* used for conditional generation.

Action: Tie the suture

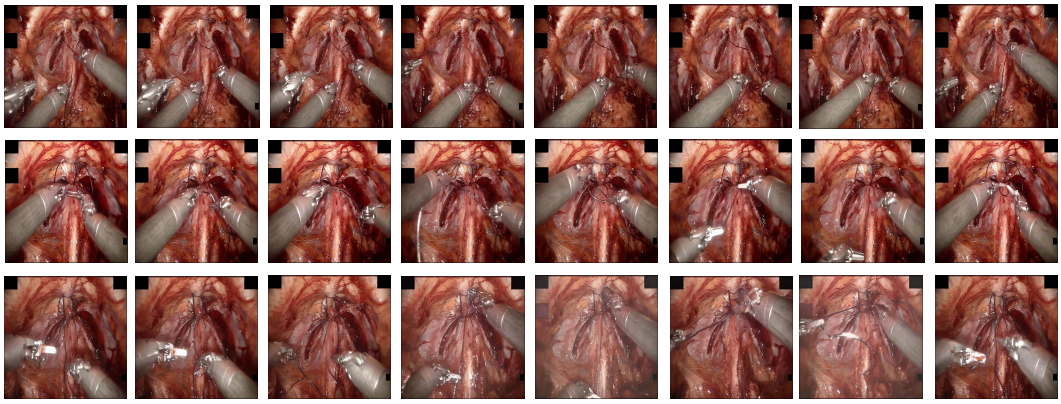

Figure 12: Results from SurgiFlowVid with text conditioning on GraSP dataset.

Action: Cut the suture or tissue

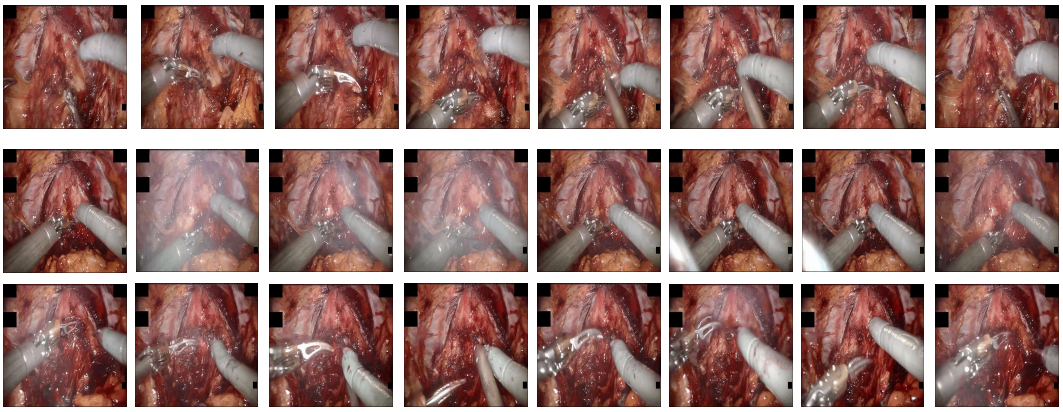

Figure 13: Results from SurgiFlowVid with text conditioning on GraSP dataset. In the 2nd row, we notice the presence of smoke as the tissue is cauterized using the tool.

Action: Cut btw. prostate & bladder

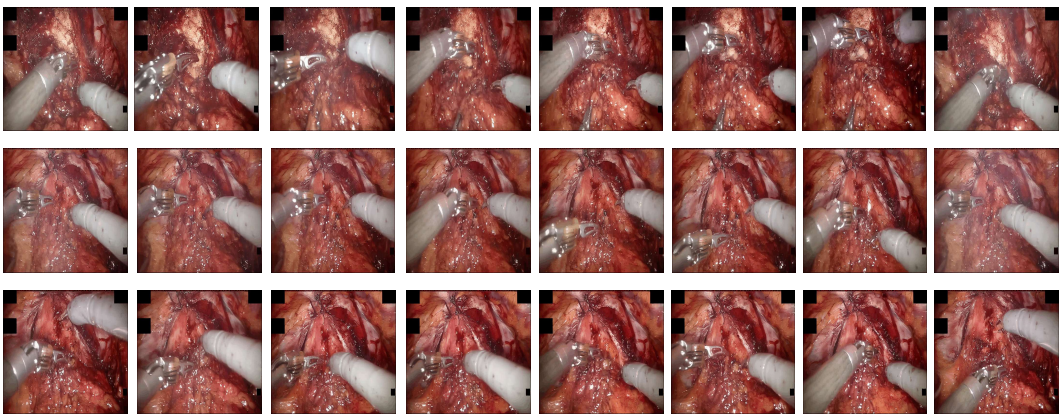

Figure 14: Results from SurgiFlowVid with text conditioning on GraSP dataset.

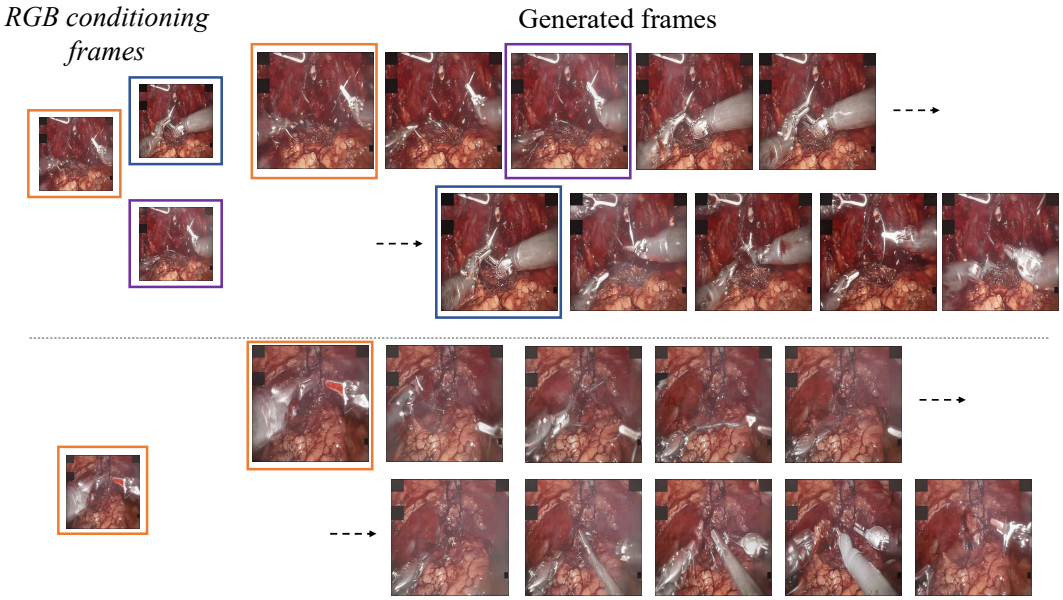

Figure 15: Results from SurgiFlowVid with RGB conditioning on GraSP dataset.The frames on the left indicate the sparse conditioning frames and the left frames indicate the generated video frames. The coloured boxes show the position of the corresponding condition frame. The dotted arrow indicates the next subsequent frames. The action corresponds to *pull the suture*.

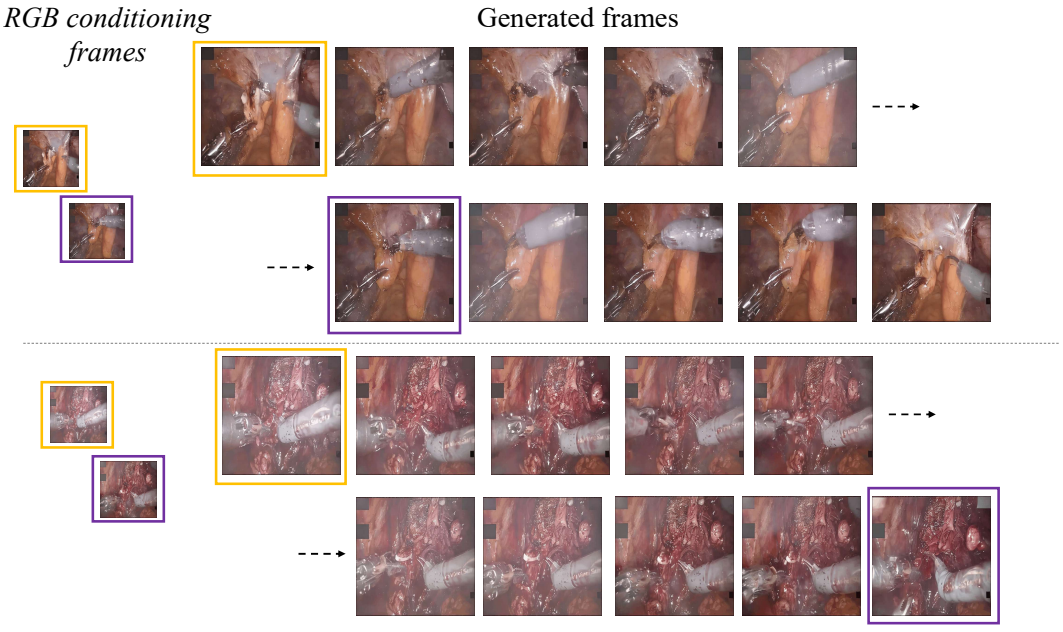

Figure 16: Results from SurgiFlowVid with RGB frame conditioning on GraSP dataset. The action corresponds to *cut the tissue*.