scientific data

Check for updates

OPEN

DATA DESCRIPTOR

TEMSET-24K: Densely Annotated Dataset for Indexing Multipart Endoscopic Videos using Surgical Timeline Segmentation

Muhammad Bilal[1,11] ✉, Mahmood Alam[1], Deepashree Bapu[2], Stephan Korsgen[2], Neeraj Lal[2,3], Simon Bach[2,3], Amir M. Hajiyavand[3], Muhammed Ali[2], Kamran Soomro[4], Iqbal Qasim[5], Paweł Capik[4], Aslam Khan[6], Zaheer Khan[4], Hunaid Vohra[7], Massimo Caputo[7,8], Andrew D. Beggs[2,3], Adnan Qayyum[9], Junaid Qadir[10] & Shazad Q. Ashraf[1,2,3,11] ✉

Indexing endoscopic surgical videos is vital in surgical data science, forming the basis for systematic retrospective analysis and clinical performance evaluation. Despite its significance, current video analytics rely on manual indexing, a time-consuming process. Advances in computer vision, particularly deep learning, offer automation potential, yet progress is limited by the lack of publicly available, densely annotated surgical datasets. To address this, we present TEMSET-24K, an open-source dataset comprising 24,306 trans-anal endoscopic microsurgery (TEMS) video microclips. Each clip is meticulously annotated by clinical experts using a novel hierarchical labeling taxonomy encompassing "phase, task, and action" triplets, capturing intricate surgical workflows. To validate this dataset, we benchmarked deep learning models, including transformer-based architectures. Our in silico evaluation demonstrates high accuracy (up to 0.99) and F1 scores (up to 0.99) for key phases like "Setup" and "Suturing." The STALNet model, tested with ConvNeXt, ViT, and SWIN V2 encoders, consistently segmented well-represented phases. TEMSET-24K provides a critical benchmark, propelling state-of-the-art solutions in surgical data science.

Background & Summary

Over 300 million surgical procedures are performed worldwide annually[1]. While surgery is a crucial healthcare intervention, it also carries significant risks, with surgical complications currently ranking as the third leading cause of global mortality[2]. Surgical adverse events result in major quality-of-life (QoL) issues for patients, and methods that critically evaluate intra-operative events have significant potential to drive up surgical standards and reduce morbidity. A key method for enhancing surgical standards involves the use of high-resolution endoscopic surgical videos (ESV). These videos capture minimally invasive surgeries (MIS) with high-definition visual records at 60 frames per second, producing two simultaneous full HD streams. This results in over 50GB of data for a single uncompressed video, with even greater volumes as surgeries lengthen, or 4K resolution technology is adopted. This poses significant challenges when attempting to create adequate storage capacity in Secure Digital Environments (SDEs), such as those recently implemented by the National Health Service (NHS), UK[3]. Despite the storage and energy costs, the value of ESV files in capturing surgical details is significant, especially at scale. Reduction in storage requirements without loss of vital information will inevitably lead to significant energy and cost savings in line with the plan to reduce the NHS Carbon Footprint to zero by 2040[4].

[1]Birmingham City University, Birmingham, United Kingdom. [2]University Hospitals Birmingham, Birmingham, United Kingdom. [3]University of Birmingham, Birmingham, United Kingdom. [4]University of the West of England, Bristol, United Kingdom. [5]University of Hertfordshire, Hatfield, Hertfordshire, United Kingdom. [6]University of Bradford, Bradford, United Kingdom. [7]University of Bristol, Bristol, United Kingdom. [8]School of Medical, University of Auckland, Auckland, New Zealand. [9]Information Technology University, Lahore, Pakistan. [10]Qatar University, Doha, Qatar. [11]These authors contributed equally: Muhammad Bilal, Shazad Q. Ashraf. ✉e-mail: muhammad.bilal@bcu.ac.uk; S.ashraf.2@bham.ac.uk

Apart from the storage and management challenges, another major stumbling block hindering surgical scene understanding is the lack of richly annotated, comprehensive datasets. A meticulously assembled large dataset is invaluable for training machine learning models to recognise objects like instruments and anatomical structures in the surgical field of view and to understand procedural phases, tasks, and intra-operative actions. Such capabilities in scene synthesis, facilitated by automated algorithms, are vital for elucidating the intricacies of surgical workflows[5] and evaluating surgeon performance[6]. This underscores the importance of developing high-quality representative ESV datasets in a SDE to advance surgical data science and create state-of-the-art (SOTA) vision tools for clinical use.

Recent advancements in video-based analysis (VBA) using AI-driven computer vision techniques present substantial opportunities for enhancing surgical scene understanding through more scalable and robust methodologies[7]. Tailoring these VBA approaches specifically for ESV is crucial for demonstrating their efficacy in surgical data science and their potential application in real-world clinical settings. At the core of surgical scene understanding is the segmentation of surgical timelines, which involves analysing video sequences to categorise diverse surgical elements—ranging from phases and tasks to activities and adverse events. Unlike object segmentation, which focuses on image-level analysis, timeline segmentation operates at the video level, presenting a volumetric and "moving object" challenge far more complex than natural scenes of stationary objects. Moreover, the high similarity between different surgical phases, variability in surgeon styles, inconsistent labelling, ambiguous workflow transitions, and the scarcity of annotated training data exacerbate the complexity. These challenges hinder the development of reliable digital tools for practical and widespread clinical use.

Additionally, manually reviewing extensive ESV files is time-consuming and inefficient for human clinical experts. If done systematically, this can take up significant time that could be used for other clinical tasks. Consequently, creating digital solutions capable of conducting comprehensive and accurate evaluations of ESV clips becomes essential to propel advancements in the field. This study aims to: (a) establish a systematic methodology for curating a high-quality, "densely" annotated ESV dataset, (b) assess the performance of cutting-edge video analysis models for surgical timeline segmentation, and (c) validate the most effective model for indexing ESV files to enhance search capabilities. This paper outlines strategies for transitioning from laboratory in-silico models to clinical applications, aiming to harness AI's potential to enhance interventional care to drive up surgical standards. To integrate SOTA methodologies in surgical data science, the whole pathway from video recording to annotation and analysis must be digitised. In summary, we make the following salient contributions:

1. We present timeline annotation taxonomy for TEMS procedures capturing five phases, 12 tasks, and 21 actions for enabling end-to-end surgical timeline segmentation using machine learning.
2. We put forward `TEMSET-24K`—a densely annotated dataset of 24,306 TEMS microclips, each labeled using our proposed timeline taxonomy, and enriched with metadata including action type (bleeding event), remaining surgical time, and other contextual attributes.
3. We share our endoscopic video review (`EVR`) Python library with the surgical data science community to perform the necessary pre-processing required for curating and managing large multipart surgical video datasets in other surgical specialities.
4. We implement and evaluate STALNet for surgical timeline segmentation in ESV using state-of-the-art encoders—`ConvNeXt`, `ViT`, and `SWIN V2`—to demonstrate the effectiveness of our proposed taxonomy and multi-target formulation, benchmarking the `TEMSET-24K` dataset for surgical video indexing.

Related Work. This section discusses prior work related to timeline analysis in surgical videos and state-of-the-art VBA methods, highlighting the potential for integrating timeline recognition with VBA to enhance the performance and generalisability of surgical data science solutions.

Timeline Analysis in Surgical Scenes. Timeline analysis in surgical videos involves breaking down surgical procedures into distinct phases, tasks, and actions to provide a comprehensive understanding of the surgical workflow. Detailed workflow specifications capture all surgical nuances using phase/task/action triplets, which are essential for designing intelligent systems in the clinical operating room. These systems can provide context-aware decision support, monitor and optimise surgical operations, and offer early alerts for potential deviations and anomalies[8,9].

Numerous studies have focused on surgical workflow analysis to identify missing activities in distinct phases, ensuring that surgeons complete necessary tasks before moving to the next phase[10]. Techniques for identifying surgical phases include data from sensors on tool tracking systems[11], binary signals from instrument usage[12], and surgical robots[13]. However, obtaining these signals typically requires additional hardware or time-consuming manual annotation, which could increase the workload associated with the surgical process[14].

Recent studies have focused on deriving the workflow solely from routinely collected endoscopic videos during surgery[15]. Automatic workflow recognition from surgical videos eliminates the need for additional equipment[16]. Notable studies include the development of EndoNet, a convolutional neural network (CNN) architecture designed to recognise surgical phases using only visual information from cholecystectomy procedures[17]. Other studies have employed temporal CNN models and transformer-based models for phase recognition in surgical activities[18,19]. For instance, Funke *et al.* proposed a temporal model, TUNeS, which integrates self-attention into a convolutional U-Net architecture to enhance surgical phase recognition[20].

Emergence of Video-Based Analytics. VBA involves meticulously breaking down and examining video content to extract important insights and intra-operative key events, transforming visual streams into semantically meaningful representations that can be easily analysed at scale[21,22]. Understanding surgical scenes requires

consideration of the temporal dimension, making VBA crucial for providing an accurate understanding of surgical processes by examining both spatial and temporal features[23].

Real-time VBA can significantly enhance surgical care, particularly for minimally invasive techniques, by providing context-aware intra-operative decision support using AI models that swiftly and accurately extract knowledge from real-time video data. This situational guidance can improve surgical outcomes by aiding in applications such as calculating surgery duration, recording important events, assessing surgical skills, and providing intra-operative assistance[24–26]. However, timeline labels in most research often lack the detail required for realistic clinical tasks, providing only coarse-grained information that fails to encompass surgical phases, tasks, and discrete actions needed for objective assessment and benchmarking of surgical performance[27,28].

Additionally, deep learning methods for surgical timeline segmentation often require large volumes of annotated data, which remain scarce in surgical domains[29]. To address this gap, Valderrama *et al*. introduced the PSI-AVA dataset[30], which offers comprehensive annotations for activity recognition in prostatectomy videos. The dataset supports short-term tasks such as atomic action recognition and includes 11 phases, 20 steps, 7 instrument types, and 16 action classes. The authors also proposed the TAPIR model, leveraging transformers for action recognition.

Similarly, Ayobi *et al*. presented the GraSP dataset and the TAPIS model—another transformer-based approach—designed to support multilevel understanding of surgical activities, including long-term temporal tasks, instrument segmentation, and atomic action detection[31]. ESAD[32] is another endoscopic surgery dataset, comprising 16 hours of radical prostatectomy videos annotated with 46,300 action instances across 21 categories. CholecT50[33] focuses on laparoscopic procedures and includes 50 videos with around 100,900 annotated frames, structured into 100 triplet classes defined by combinations of 6 instruments, 10 verbs, and 15 targets.

Other notable datasets include Cholec80[17], EndoScapes[34], and CholecTrack20[35], each offering significant value within their respective domains. However, most of these datasets face notable limitations. Many are limited in size, restricting their utility for training deep learning models. Some, such as CholecT50, do not publish their test sets, complicating model validation and reproducibility. Others may contain incomplete or inconsistent annotations, limiting their generalisability and performance when used in clinical AI applications[36].

By combining insights from timeline analysis and VBA, our study aims to further advance the understanding and evaluation of surgical procedures through the development of high-quality, deeply annotated ESV dataset and the establishment of robust AI models for surgical timeline segmentation.

Methods

Transanal Endoscopic Microsurgery (TEMS) Overview. The dataset described in this paper comprises recordings of TEMS procedures performed on patients with early rectal cancer or large pre-cancerous polyps[37]. During the TEMS procedure, an operating scope is inserted trans-anally into the rectum. This is a stable and flexible platform that enables surgical access from the anorectal junction to the most cephalic aspect of the rectum. The top of the rectum is roughly 15cm from the anal verge (bottom of the anal canal). Most of the rectal lumen can be reached with this TEMS operating scope. The surgeon adjusts the scope to reach and remove the tumour, manoeuvring it as needed. The procedure begins with a `setup` phase, which includes preparing the scope, instruments, and the surgical site. The rectum is inflated with carbon dioxide to a preset pressure, and faecal debris and fluid are removed with a suction device to obtain clear views. The main phase involves dissecting the tumour, removing the specimen, and closing the rectal wall defect. Surgeons use a `clockface` analogy to navigate around the lesion site, facilitating precise removal. Dissection may be partial (mucosa and submucosa) or full thickness (deeper muscle tissues).

During dissection, multiple small events like surgical "smoke" fogging the lens, lens wash, tissue cauterisation, tissue retraction, fluid aspiration, and bleeding may occur. These are inter-related, for example tissue cauterisation results in surgical smoke that fogs up the operating scope and is normally managed by scope wash to clean the camera lens and aspiration of any fluid in the field of view. Bleeding is controlled with diathermy instruments and aspiration. Various instruments are used based on surgical needs. After dissection, the specimen is removed through the scope for histological analysis. In the closure phase, the surgeon uses a running suture to close the rectal wall defect. This involves handling the needle, driving it through the rectal wall, and pulling the suture to fully close the defect. Figure 1 illustrates the key steps of the surgical workflow for the entire TEMS procedure.

Patient Cohort. This study included fully de-identified videos from patients with a clinical diagnosis of rectal polyps or cancer. Pre-operatively, patients underwent standard clinical staging, including optical endoscopy, biopsy, endo-rectal ultrasound, magnetic resonance imaging, and computed tomography. These cases were discussed in a cancer multidisciplinary meeting before elective surgery was offered. A team of four specialist colorectal surgeons, all Fellows of the Royal College of Surgeons (FRCS), performed TEMS using a Richard Wolf trans-anal operating platform.

Ethical Statement and Data Compliance. The study is registered as a clinical audit with the University Hospitals Birmingham (UHB), conforming to local ethical standards, under the Clinical Audit Registration Management System (CARMS) number 20648. The audit title was "Automated Surgical Timeline Prediction in Endoscopic Videos Using Computational AI Techniques" and was signed off at a divisional level and reviewed by the UHB Information Governance team (reference IG937). The justification for the project was to improve the ability of surgeons to retrospectively review operations using reliable timestamps. The surgical video analysis was performed by clinicians using scripts designed and shared by computational data scientists. Informed consent was obtained from all patients before recording fully de-identified surgical videos. Specifically, all patients signed institutional consent forms that gave permission to share a fully anonymised video on an open-access platform

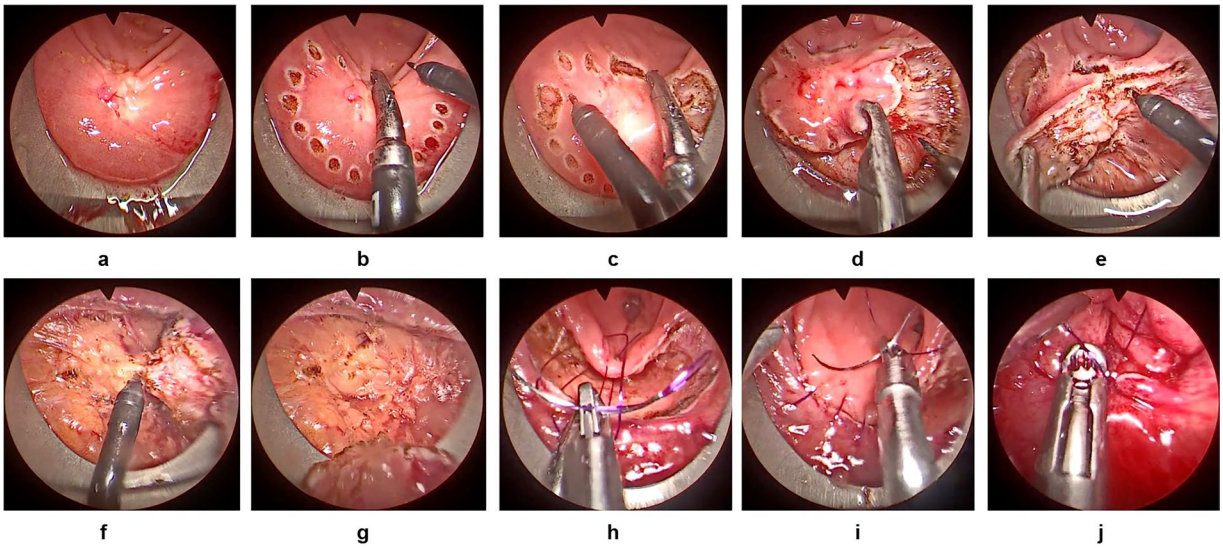

Fig. 1 TEMS surgical workflow. A typical surgical flow from landmarking of the rectal polyp to dissection, lesion removal and closure of the rectal wall defect. The key milestones of a TEMS procedure are detailed in images a-j: [**a**] Baseline lesion in view after setup; [**b**] Application of landmark "dots" to outline the lesion; [**c**] Dissection of the wall through the mucosa and muscle; [**d**,**e**] Circumferential removal of the lesion; [**f**,**g**] Final removal and extraction of the specimen; [**h**,**i**] Closure of the rectal wall defect with a suture; and [**j**] Application of a metal clip to secure the suture and ensure complete closure.

that can be seen by medical professionals, researchers or members of the general public. The videos have already been used to train junior surgical trainees in the West Midlands (UK) region. In accordance with NHS ethical standards and the UK General Data Protection Regulation (GDPR), the routinely collected ESV dataset underwent a full anonymisation procedure to ensure the removal of any identifiable information to protect patient privacy and ensure confidentiality. Rigorous measures were implemented to review each video by at least two clinicians to ensure that patient identifiers were not accidentally captured or disclosed. Surgical scenes that extended beyond the abdominal cavity, capturing the surgical team or hospital surroundings, were removed by the surgical team. These segments, typically occurring when the camera was temporarily extracted from the endo-luminal cavity for cleaning purposes, were replaced by blank frames while preserving patient privacy and maintaining the surgical procedure's overall chronological sequence and duration.

Data Capture & Sharing. We used the Operating Room (OR) visualisation system for data acquisition comprising a stereo endoscopic 50-degree scope and eyepiece attached to the Karl Storz Image 1 Hub HD (high-definition) camera system. Multipart HD videos were recorded and archived using the Karl Storz AIDA™ system, which contains an intelligent export manager that automatically saves surgical video files during surgery. These files were stored on encrypted NHS hospital-based hard drives. All patient information was removed to ensure no metadata containing patient-related information was shared or made accessible to the project teams.

Co-Creation of Dense Taxonomy Labels for Timeline Segmentation. In our effort to develop a comprehensive taxonomy for annotation, the project task group worked with specialist surgeons to define a representative surgical workflow. This collaborative co-creation process was essential for capturing the intricate details necessary to describe various downstream clinical tasks, in order to facilitate precise and detailed video labelling.

To achieve this, we structured the labels into `phase`, `task`, and `action` "triplets". This hierarchical framework allowed for a detailed end-to-end breakdown of the surgical procedure:

- **Phase**: Represents "high level" activities encompassing a series of surgical tasks (for example setup of the TEMS scope or dissection phase).
- **Task**: Mid-level activities within a phase that encompass specific tasks (for example, dissection phase may involve landmarking the dissection plane around the tumour or mucosal dissection).
- **Action**: The most granular unit activity within a task, (for example, dissection, retraction, lens wash, identification of bleeding, haemostasis or aspiration of fluid).

For the TEMS procedure, we identified five key high-level phases: "Setup", "Dissection", "Specimen Removal", "Closure of Defect", and "Scope Removal". Each phase consists of multiple tasks, which in turn are made up of specific actions. This "triplet" structure ensures that every aspect of the surgery is captured in detail, facilitating accurate and structured analysis. (See Fig. 2 for detailed specification of our proposed TEMS surgical workflow taxonomy.)

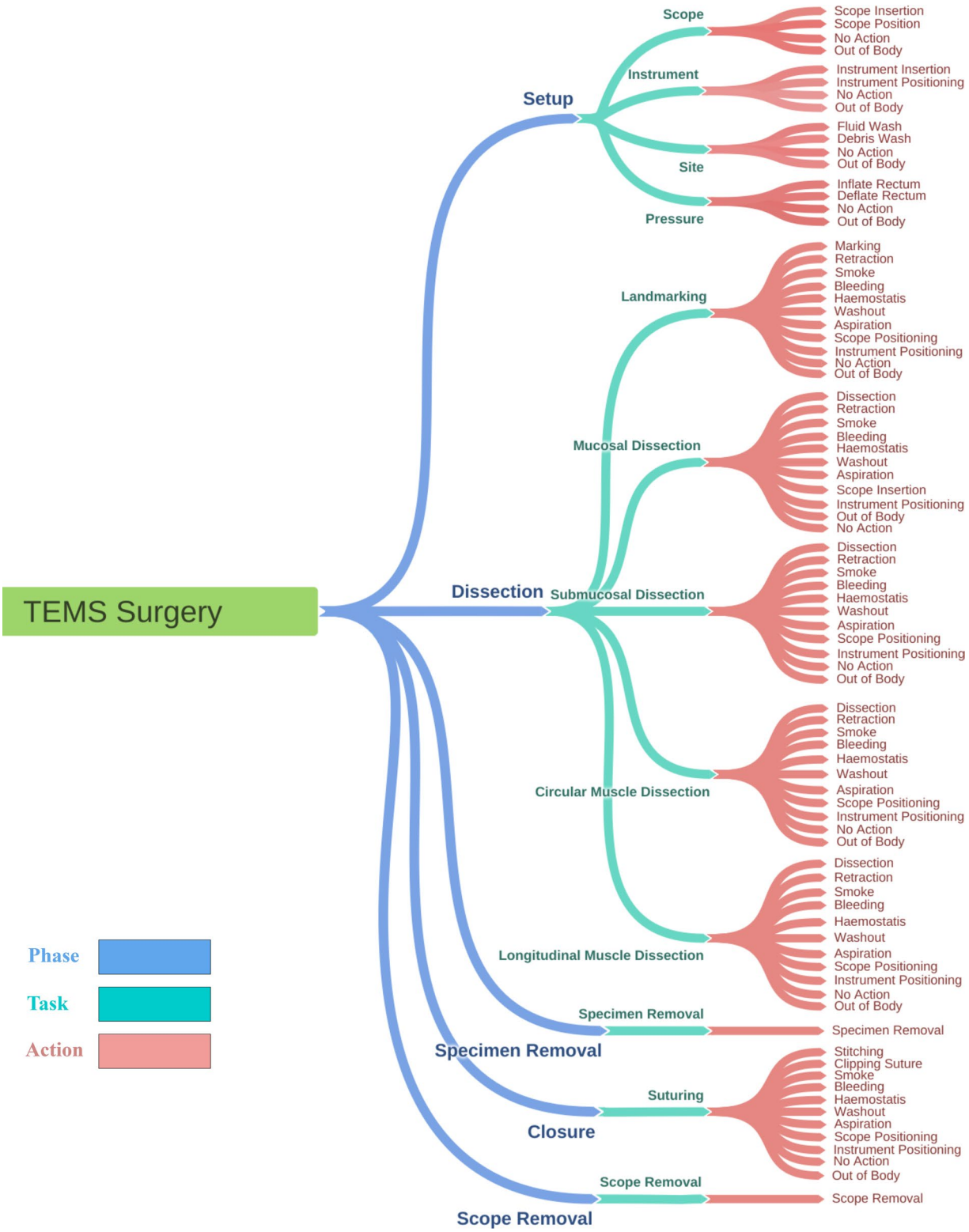

Fig. 2 Proposed Taxonomy of TEMS Surgical Workflow. The TEMS operation can be split into three levels: [**A**] High level activity phase (such as Set-up, Dissection, Specimen Removal, Closure and Scope Removal), [**B**] Task based activities (such as scope insertion, instrument movement, site wash and pressure increase), [**C**] Small unit tasks (such as tissue marking, tissue retraction, smoke identification, bleeding identification and haemostasis).

This structured approach not only helps in the detailed documentation of the procedure but also enhances the ability to extract key events within an operation. This can be subsequently used to perform post-operative clinical assessments, either at an individual surgeon level or in comparison of techniques in a cohort of surgeons.

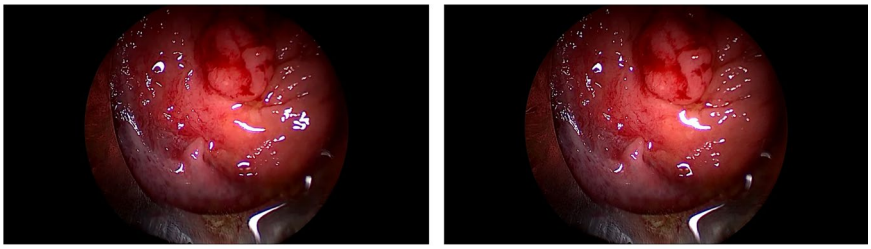

Fig. 3 Side-by-side comparison of videos before and after pre-processing. Panels (**a**) and (**b**) show image quality before and after pre-processing, respectively. This shows that despite reduction in the size of the ESV file by a factor of 10 (from 1GB to 0.1GB), there was no loss in quality.

This approach ensures that every critical operative action is accounted for and can be analysed in a constructive manner for measuring incremental performance improvements. This enables extremely "dense" data to be extracted from VBAs and analysed at scale. Large video libraries can be interrogated with the approaches defined above to understand changes in individual surgical performance over time as well as comparing surgeons against their peers.

Our proposed triplet structure was co-designed to reflect both procedural flow and anatomical specificity during TEMS. This hierarchical structure enables a multi-resolution view of the surgical workflow—from broad procedural phases to the most granular operative actions. While some actions (e.g., washout, dissection, retraction) recur across multiple tasks, this repetition is intentional and reflects the practical reality of surgical practice, where the same action may serve very different functions depending on its anatomical and procedural context. For example, the action "Washout" during the `task` "Mucosal Dissection" typically involves clearing superficial debris (e.g., blood, mucus or faeces) from the mucosal layer, whereas the same action "Washout" in `task` "Circular Muscle Dissection" is often used to manage deeper cautery-induced bleeding and visual occlusion in the muscularis propria layer of the rectal wall. Though the action label remains the same, its clinical intent, anatomical site, and operative complexity differ substantially. Likewise, the action "Dissection" in the triplet "Dissection.Submucosal Dissection.Dissection" versus "Dissection.Longitudinal Muscle Dissection. Dissection" implies distinct surgical phases that require subtly differing tissue-handling skills, which are contextually disambiguated by the full triplet.

This design choice ensures that surgical behaviours are not only recorded but also meaningfully classified, enabling accurate temporal segmentation and clinically relevant performance review. It allows the dataset to represent shared surgical events (e.g., bleeding, smoke) that occur across anatomical layers while preserving their context. Furthermore, this structure supports detailed clinical analyses—such as comparing the frequency of bleeding or washout events across tasks, or assessing task efficiency within specific phases. The taxonomy has been iteratively validated against expert-defined intraoperative workflows, ensuring consistency and clinical relevance. In this way, the triplet-based classification provides both strong clinical significance and the granularity needed for scalable, high-resolution surgical analysis.

Infrastructure Setup, Video Annotation, and Exporting Labels. Our ESV dataset is extracted from multipart videos of TEMS surgeries, with uncompressed original videos of 10.34 gigabytes (GB) in size and covering procedures lasting up to 6 hours. We evaluated several labeling platforms and found that LS (Label Studio version 1.12.1) is the most suitable for video annotation, despite some limitations. LS[38] is a secure web-based annotation tool supporting text, photos, videos, audio, and sequence data. While it offers extensive features, certain constraints affected its usability for our research. First, LS requires video alongside audio tracks for timeline segmentation functionality, but no audio was included in this dataset construction for obvious privacy reasons. Second, the default video upload limit for one file in LS is 250 megabytes, which can be extended but affects annotation interface performance adversely. Lastly, managing multipart ESV files within a single project is challenging, as LS treats each surgery as a separate project, complicating data organisation.

To address these issues, we inserted blank audio tracks into the raw videos using `FFmpeg` (https://ffmpeg.org/). We also compressed the bitrate from 13 Mbps to 1 Mbps, significantly reducing file size to enable smoother uploads and lower server load—without compromising visual quality. Figure 3a,b illustrates a surgical scene before and after compression, showing minimal loss in visual fidelity.

After preparing the videos, LS was installed on a server in a `Docker` container and made accessible for labelling by using a `ngrok` tunnelling platform. ESV files were uploaded into LS across several projects. For each project, the LS interface was customised for video-based timeline annotation. The inclusion of empty audio tracks enabled the use of the LS timeline component for segment-based labelling, significantly aiding efficient clinical annotation of large multipart ESV files. Secure logins, allowed different surgeons to perform initial segmentation, which was then reviewed and finalised by a panel of clinical domain experts to ensure consistent labelling. Finally, we exported the labels from LS in JSON format, with each multipart ESV file generating one JSON file for timeline segmentation.

Post-Processing of Annotations to Generate ML-ready Dataset. We developed a systematic approach to transform the densely annotated multipart ESV files into a machine learning-ready dataset. The raw ESV footage contains approximately 3 million frames. Due to the computational challenge of handling such a

large dataset, we adopted an intelligent data sampling strategy to curate 24,306 microclips. This involved identifying semantically distinct anchoring frames—using cosine distance similarity—that capture significant changes and meaningful transitions in the surgical scene.

This approach reduced the number of anchoring points per video to an average of approximately ~ 550, compared to the original $\sim 15K$ frames. Each anchoring frame guided the generation of a corresponding microclip: a 30-second retrospective window of surgical video leading up to that point, thereby preserving the temporal dynamics of the operation. Each microclip was downsampled from 125 fps to 30 fps (yielding approximately 900 frames), and collectively they form a large-scale, temporally rich dataset—not a static image collection. This structured granularity distinguishes TEMSET-24K[39] from prior datasets such as CholecT50[33], which, while impactful, offers sparse annotations and lacks a publicly released validation split. Moreover, TEMSET-24K is released in a machine learning-ready format, with timeline-consistent clip naming, transition-aware labels, and metadata such as "time-to-finish" for surgical forecasting tasks.

Importantly, these microclips are not isolated snapshots but curated temporal segments, specifically designed for training models to segment fine-grained surgical timeline. Advanced `FFmpeg` features were used to uniquely name each microclip by combining the surgery ID, ESV file name, and timestamp, ensuring chronological order and reproducibility. These microclips are stored as `.mp4` files in the `microclips` folder.

Next, we implemented a range-based query method to map each microclip to its corresponding label from the Label Studio-exported JSON files. This mapping is stored in the `timeline_labels.csv` file, which initially contains the columns `filename` and `timeline_label_raw`. To enhance the annotation metadata, we added columns such as `surgery_name`, `video_name`, and `timestamp`.

We identified label overlaps at action transitions, where `timeline_label_raw` contained dual labels for boundary frames. To resolve this, we created a new `timeline_label` column by selecting the trailing label and introduced a `transition` column to capture such transitions. For multi-target modelling, we split `timeline_label` into three columns: `timeline_phase_label`, `timeline_task_label`, and `timeline_action_label`, each representing a different level of the TEMS taxonomy. We also added a `valid` column to indicate the suggested validation set, allowing researchers to benchmark algorithms using this dataset.

The detailed steps of this post-processing workflow are summarised in Algorithm 1, which outlines the procedure for identifying anchoring frames, creating microclips, and mapping timeline labels from LS-generated JSON files. The complete codebase is available in our GitHub repository `EVR` (https://github.com/bilalcodehub/evr), which includes Python scripts such as `split.py`, `map.py`, and `microclip.py` to support each step. Designed for flexibility, the `EVR` library can be applied to a wide range of surgical video datasets, ensuring systematic processing and transformation of raw video and annotations into a structured format suitable for machine learning applications.

Algorithm 1 Post-Processing for Generating ESV Dataset.

Data: ESV videos path, `offset` (microclip length)
Result: Extracted frames, timestamped labels, and corresponding microclips
begin
  `/* Step 1: Keyframe Extraction */`
  foreach *video in dataset* **do**
    timestamps ← extract keyframe timestamps;
    keyframes ← extract keyframes from video using timestamps;
  end
  `/* Step 2: Load Annotations */`
  annotation data ← load JSON annotations;
  `/* Step 3: Label Mapping for Keyframes */`
  foreach *keyframe in extracted keyframes* **do**
    video_name, timestamp ← parse filename;
    labels ← find labels from annotation data using video_name and timestamp;
    save keyframe filename, labels to results;
  end
  `/* Step 4: Save Results */`
  write results to `CSV`;
  `/* Step 5: Create Microclips */`
  foreach *entry in results* **do**
    video_name, timestamp ← parse filename;
    start_time ← calculate start time (max(0, timestamp - `offset`));
    microclip_file ← create microclip using `FFmpeg`;
    save microclip to designated folder;
  end
end

Data Records

We are releasing the TEMSET-24K[39] dataset to the surgical data science community to advance timeline segmentation capabilities and video-based analytics with a focus on improving surgical performance. All videos are from patients who have consented for de-identified videos to be made open-source for service evaluation, non-commercial education and research purposes. Users must agree not to attempt use for any other purpose. The fully deidentified video dataset is hosted and can be accessed at https://zenodo.org/records/14016844 after completing a data sharing agreement.

The dataset is packaged as a zipped `temset` folder (~20*GB*), which includes several subfolders: `videos`, `microclips`, and accompanying metadata files.

The `videos` folder contains subdirectories for each surgical case, pseudonymised as `TEMS-001`, `TEMS-002`, etc. Within each folder, an `originals` subfolder holds the high-resolution, multipart ESV recordings. These files were compressed to ~ 10% of their original size using a bespoke pipeline to facilitate storage and handling. We stored compressed videos in the `videos` folder for easy access and manipulation. Each surgery folder also includes a corresponding expert-annotated JSON file (e.g., `TEMS-001.json`) exported from LS, containing timeline segmentation labels based on our proposed dense TEMS taxonomy. These annotations are aligned with the videos and serve as the foundation for generating machine learning-ready labels.

The `microclips` folder contains the curated 30-second retrospective clips for each identified anchoring frame. Each microclip is saved as an `.mp4` file and named using a consistent format that combines the surgical ID, video part, and timestamp. These clips are not random snapshots but temporally coherent video segments prepared for training timeline segmentation models.

The file `timeline_labels.csv`, located in the root `temset` folder, contains the processed labels for all microclips. Each row corresponds to a microclip and includes a unique filename (comprising the surgical ID, video name, and timestamp) for reproducibility and traceability to the original ESV files. Labels are provided both as dot-separated triplets (`timeline_label`) and as separate columns: `timeline_phase_label`, `timeline_task_label`, and `timeline_action_label` to support both single- and multi-target formulation.

An additional column, `time_to_finish`, records the estimated remaining surgical time for each microclip. This is computed by subtracting the microclip's anchoring frame timestamp from the total duration of the surgery, enabling research into temporal prediction tasks such as surgical progress estimation.

Figure 4 illustrates the end-to-end methodology adopted to create the TEMSET-24K dataset, from surgical video acquisition to annotation and post-processing for timeline segmentation.

Technical Validation

Annotation Assessment. To ensure the consistency of labelling in the dataset, we designed an annotation process involving a team of colorectal cancer surgery specialists, all accredited with fellowship status with the Royal College of Surgeons (RCS, UK). The process began with one surgeon annotating one full video in a shared setting to demonstrate the annotation procedure for the multipart ESV files. Following this, another surgeon logged into the LS server using their credentials and navigated to the project they intended to annotate, accessing the individual video clips for annotation. The LS user interface provided a comma-separated list of phases, tasks, and actions for annotating the timeline of each video clip. Annotations were initially performed by one surgeon and subsequently validated by at least two other surgeons for cross-checking purposes. In cases of conflicting boundaries between the start and end of the labelling triplets, discussions were held to finalize the annotations that was agreed by all surgeons. We employed multifaceted strategies involving our proposed dense taxonomy, collaboratively annotating one full surgery in shared settings, and holding iterative discussions to resolve conflicts to achieve consistent annotations of the complex workflow scenes based on all surgeons' inputs. The final annotations consisted of labels made up of five phases, 12 tasks, and 21 actions as defined by the proposed taxonomy. These annotations were then programmatically exported from LS in JSON format, along with the corresponding ESV files.

Deep Learning Model Training. *Data Pre-Processing.* To improve the field of view, irrelevant areas were cropped from ESV images comprising black regions. The input image was first converted to grayscale, and a binary threshold was used to isolate the circular surgical region from the background. This step enhanced the visibility of the surgical scene. Subsequently, the largest contour was identified within the thresholded image and computed its minimum enclosing bounding box. A mask corresponding to this circular region was created and applied to the original image to extract the surgical area while ignoring the background. The bounding box of the surgical region was cropped and this cropped image was resized to its original size using bilinear interpolation. This method ensures that only the relevant surgical view is retained and standardised, facilitating improved visualisation and analysis of the surgical scene.

Problem Formulation. A key objective of this study was to learn an unknown function $\mathbf{F}$ that maps high-dimensional TEMS endoscopic surgical videos $\mathbf{X} \in \mathbb{R}^{T\times H\times W\times 3}$ to a multitarget label triplet $\mathbf{Y} \in \{\text{Phase, Task, Action}\}$, where, T, H, and W denote the sequence length (no. of frames in the video), height, and width of the frames, respectively. To achieve this, this study proposes a Spatiotemporal Adaptive LSTM Network (STALNet) that learns the desired mapping. As shown in Fig. 5, STALNet integrates a TimeDistributed video encoder $\mathbf{E}^T$, followed by an adaptive long-short term memory network (LSTM) module having attention as the last layer $\mathbf{M}_{\text{AA-LSTM}}$ to capture spatial and temporal dependencies in the ESV data. Let ϕ be the feature extraction function using the backbone. The output of the encoder is given by:

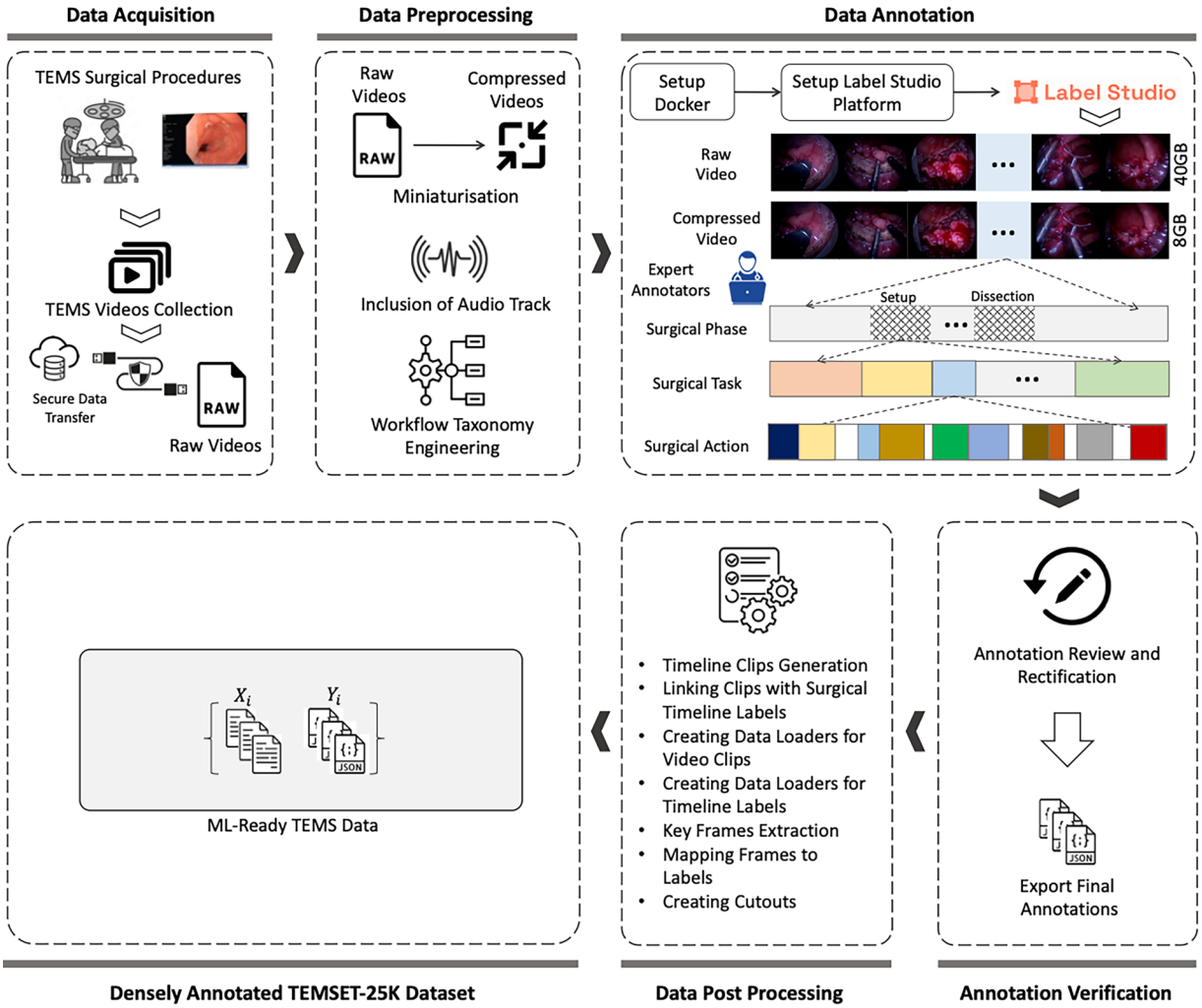

Fig. 4 Methodology adopted for annotating TEMS surgical videos for surgical timeline segmentation includes six major steps. (**1**) TEMS surgical data acquisition (**2**) Data Preprocessing, (**3**) Data Annotation, (**4**) Annotation Verification, (**5**) Data Post Processing, and (**6**) Surgical data preparation for training timeline segmentation models.

$$\mathbf{F} = \mathbf{E}^T(\phi(\mathbf{X})); \mathbf{X} \in \mathbb{R}^{B\times T\times C\times H\times W}, \tag{1}$$

where, B is the batch size, T is the sequence length, C is the number of channels, and H and W are the height and width of the frames, respectively. We experimented with various encoders, including `ConvNeXt` (convnext_small_in22k)[40], `SWIN V2` (swinv2_base_window12_192-22k)[41], and `ViT` (vit_small_patch16_224)[42,43]. These encoders were chosen for their proven ability to capture detailed spatial features across different scales, which is crucial for accurately interpreting surgical video frames. The extracted features are fed into an `Adaptive LSTM` module. This module consists of multiple LSTM layers, where the number of LSTMs depends on the input sequence length T. Each LSTM processes the sequence of features and produces hidden states. Let $\mathbf{h}_t$ represent the hidden state at time step t. The hidden states are computed as:

$$\mathbf{H}_t = \mathbf{M}_{\mathrm{AA-LSTM}}(\mathbf{F}_t, \mathbf{h}_{t-1}),$$

where $\mathbf{H}_t \in \mathbb{R}^{B\times D}$. Multiple LSTM layers were applied to capture temporal dependencies across the sequence. Incorporating LSTMs into the proposed solution in an adaptive manner significantly improved the model's capacity for surgical scene understanding, as this approach leverages and preserves the temporal coherence in the videos, improving the stability and accuracy of the timeline predictions. The final hidden states from each LSTM layer are collected as $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \ldots, \mathbf{H}_T] \in \mathbb{R}^{T\times B\times D}$ and their information across the sequence is aggregated using an attention mechanism. The attention weights are computed by applying a linear layer to the hidden states:

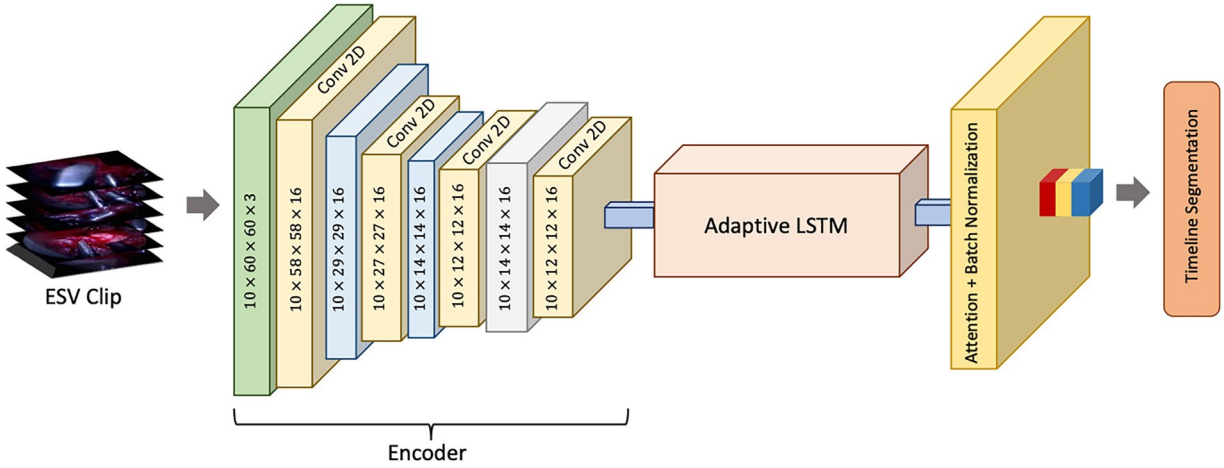

Fig. 5 Proposed SpatioTemporal Adaptive LSTM Network (STALNet) for Surgical Timeline Segmentation. This network diagram shows the process by which ESV clips are analysed by encoders in order to apply reliable timeline segments.

$$\mathbf{A}_t = \text{softmax}\,(\mathbf{W}_a\mathbf{H}_t),$$

where $\mathbf{W}_a \in \mathbb{R}^{D\times 1}$ is the attention weight matrix. The attention-weighted output is computed as a weighted sum of the hidden states:

$$\mathbf{O} = \sum_{t=1}^{T}\mathbf{A}_t\mathbf{H}_t \in \mathbb{R}^{B\times D}.$$

The final output is obtained by passing the attention-weighted output through a fully connected layer followed by batch normalisation:

$$\mathbf{Y} = \text{BatchNorm}\,(\mathbf{W}_h\mathbf{O}),$$

where $\mathbf{W}_h \in \mathbb{R}^{D\times(P+T+A)}$, with P, T, and A representing the number of phases, tasks, and actions, respectively. A technique was employed here for mean ensembling to create more robust learners for each model, followed by heuristic-based prediction correction to address sporadic predictions.

The model is trained using a custom loss function that combines the losses for phase, task, and action predictions. The total loss is given by:

$$\mathcal{L} = \alpha\mathcal{L}_p + \beta\mathcal{L}_t + \gamma\mathcal{L}_a,$$

where $\mathcal{L}_p$, $\mathcal{L}_t$, and $\mathcal{L}_a$ are the individual losses for phase, task, and action predictions, and α, β, and γ are their respective weights. Each of these losses is computed using the `CrossEntropyLossFlat` function applied to each of the output triplets.

DL Model Implementation. The model described in this paper was implemented using the `fastai`[44] library. A server with 4 NVIDIA LS40 GPUs was used for training and validation. To enhance model convergence, the default `ReLU` activation function was replaced with the `Mish` activation function, which demonstrated superior performance in our experiments. Additionally, we substituted the default `Adam` optimiser with `ranger`, a combination of `RectifiedAdam` and the `Lookahead` optimisation technique, providing more stable and efficient training dynamics. To further optimise the training process, the `to_fp16()` method was employed to reduce the precision of floating-point operations, thereby enabling half-precision training and improving computational efficiency. The `lr_find` method was utilised to determine the optimal learning rate for the model, implementing a learning rate slicing technique. This approach assigned higher learning rates to the layers closer to the model head and lower learning rates to the initial layers, facilitating more effective training. For benchmarking, we initially evaluated several network architectures, including a basic image classifier, to establish a trivial baseline. This simple approach, however, produced significant sporadic predictions due to the absence of sequence modelling, highlighting the necessity for a more sophisticated model.

Model Validation. The model described in this paper was validated against the human annotator ground truth using the server with NVIDIA LS40 GPUs. We compared the proposed `STALNet` architecture with various encoder backbones, including `ConvNeXt`, `SWIN V2`, and `ViT`. The output results were analysed against the baseline to look at comparative performance metrics and how they captured the spatiotemporal dependencies that are crucial for the surgical timeline segmentation task.

Sr#	Model Descripton	ConvNeXt		ViT		SWIN V2	
		Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
1	Baseline Vision Classifier	80.36%	72.99%	75.23%	60.87%	78.74%	66.70%
2	STALNet (Ours)	**91.69%**	82.78%	83.02%	68.29%	91.42%	**86.02%**

Table 1. Comparison of Surgical Timeline Segmentation Models.

Phase Name	ConvNeXt		ViT		SWIN V2	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
[01] Setup	0.99 ± 0.09	0.97 ± 0.02	0.98 ± 0.13	0.94 ± 0.05	0.99 ± 0.10	0.97 ± 0.03
[02] Dissection	0.99 ± 0.10	0.99 ± 0.00	0.97 ± 0.17	0.97 ± 0.00	0.99 ± 0.11	0.99 ± 0.00
[03] Specimen Removal	1.00 ± 0.03	0.97 ± 0.03	1.00 ± 0.04	0.95 ± 0.05	1.00 ± 0.02	0.99 ± 0.01
[04] Closure	0.99 ± 0.09	0.99 ± 0.00	0.98 ± 0.14	0.98 ± 0.01	0.99 ± 0.08	0.99 ± 0.00
[05] Scope Removal	1.00 ± 0.03	1.00 ± 0.00	1.00 ± 0.05	0.99 ± 0.01	1.00 ± 0.03	1.00 ± 0.00

Table 2. Performance of the STALNet model on Surgical Phases across different encoders.

Statistical Analysis. For our model evaluation, we utilised standard metrics including accuracy, F1 score, and ROC (Receiver Operating Characteristic) curves. To illustrate model variability, standard deviation is reported for accuracy and F1 scores. The following equations define these metrics:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \times 100\%, \\ \text{Precision} &= \frac{TP}{TP + FP}, \\ \text{Recall} &= \frac{TP}{TP + FN}, \\ \text{F1 Score} &= 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \end{aligned} \tag{2}$$

We computed these statistics at two levels: 1) Overall Model Performance: We reported the overall accuracy and F1 score on the entire validation set. 2) Class-Specific Performance: These metrics were computed for each taxonomy triplet class (phase, task, and action) to identify which classes the model struggles with the most. Additionally, ROC curves were used to visually investigate model performance. True positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) were derived from the predictions, which were then used to compute precision and recall, leading to the construction of ROC curves plotted using Scikit-learn. To enhance our analysis, we implemented custom visualisations showing video clips, target labels, and model predictions. We employed color coding (red for incorrect and green for correct predictions) for easy interpretation. All data and model results were visualised and analysed using Matplotlib, NumPy, and Scikit-learn.

Model Performance Evaluation. Table 1 presents the accuracy and F1 scores for each model across the three encoder architectures. The baseline image classification learner, which predicts timeline labels based solely on individual images, achieved an F1 score of 72.99% with the ConvNeXt encoder, 66.7% with the SWIN V2 encoder, and 60.87% with the ViT encoder. These results indicate the fundamental capability of deep learning models for surgical timeline segmentation but also highlight the limitations of relying solely on spatial information. In contrast, our proposed `STALNet` demonstrated significant performance improvements over the baseline model. On average, `STALNet` achieved an F1 score of 82.78% and an accuracy of 91.69%, reflecting an average performance gain of 9.79% in F1 score and 11.38% in accuracy compared to the baseline model. These improvements underscore the importance of incorporating spatiotemporal information for surgical timeline segmentation. Furthermore, the performance varied between different model encoders used in the time-distributed layer for feature extraction. Among the evaluated encoders, the ConvNeXt encoder achieved the highest accuracy with 91.69%, slightly better than the SWIN V2 encoder at 91.41%. However, the highest performing F1 score, which is a significant metric for evaluating timeline segmentation, was achieved by the SWIN V2 encoder at 86.02%, which is approximately 3.24% higher than the ConvNeXt encoder's F1 score of 82.78%. This demonstrates that while ConvNeXt offers marginally better accuracy, SWIN V2 excels in terms of F1 score, highlighting its superior performance in capturing relevant features for timeline segmentation. Despite the higher F1 score of SWIN V2, it required substantial computation during both training and deployment phases. On the other hand, ConvNeXt not only delivered competitive performance but also offered a more computationally efficient solution, making it a practical choice for real-world applications. Overall, the `STALNet` model, particularly with the ConvNeXt encoder, demonstrated superior performance in segmenting surgical timelines. This highlights the efficacy of integrating spatiotemporal features and selecting robust encoder architectures to balance performance and computational efficiency.

Task Name	ConvNeXt		ViT		SWIN V2	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
[01] Scope Setup	0.99 ± 0.10	0.96 ± 0.04	0.99 ± 0.11	0.94 ± 0.05	0.99 ± 0.09	0.96 ± 0.04
[02] Instrument Setup	1.00 ± 0.02	0.94 ± 0.06	1.00 ± 0.03	0.81 ± 0.19	1.00 ± 0.02	0.92 ± 0.08
[03] Site Setup	1.00 ± 0.06	0.83 ± 0.17	1.00 ± 0.07	0.84 ± 0.16	1.00 ± 0.07	0.82 ± 0.18
[04] Pressure Setup	0.99 ± 0.07	0.93 ± 0.07	0.99 ± 0.11	0.85 ± 0.15	0.99 ± 0.08	0.93 ± 0.07
[05] Landmarking	0.99 ± 0.07	0.98 ± 0.02	0.98 ± 0.13	0.93 ± 0.06	0.99 ± 0.09	0.97 ± 0.03
[06] Mucosal Dissection	0.98 ± 0.14	0.95 ± 0.04	0.94 ± 0.23	0.87 ± 0.10	0.98 ± 0.15	0.95 ± 0.04
[07] Submucosal Dissection	0.98 ± 0.14	0.96 ± 0.03	0.96 ± 0.20	0.91 ± 0.06	0.98 ± 0.14	0.96 ± 0.03
[08] Circular Muscle Dissection	0.99 ± 0.11	0.96 ± 0.04	0.96 ± 0.19	0.86 ± 0.12	0.98 ± 0.12	0.95 ± 0.04
[09] Longitudinal Muscle Dissection	0.99 ± 0.11	0.97 ± 0.02	0.97 ± 0.17	0.93 ± 0.06	0.99 ± 0.11	0.97 ± 0.02
[10] Specimen Removal	1.00 ± 0.03	0.98 ± 0.02	1.00 ± 0.04	0.96 ± 0.04	1.00 ± 0.02	0.99 ± 0.01
[11] Suturing	0.99 ± 0.09	0.99 ± 0.00	0.98 ± 0.14	0.98 ± 0.01	0.99 ± 0.09	0.99 ± 0.00
[12] Scope removal	1.00 ± 0.03	1.00 ± 0.00	1.00 ± 0.05	0.99 ± 0.01	1.00 ± 0.03	1.00 ± 0.00

Table 3. Performance of the STALNet model on Surgical Tasks across different encoders.

Action Name	ConvNeXt		ViT		SWIN V2	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
[01] Aspiration	0.97 ± 0.18	0.85 ± 0.13	0.93 ± 0.26	0.69 ± 0.27	0.97 ± 0.18	0.85 ± 0.13
[02] Bleeding	0.99 ± 0.07	0.82 ± 0.17	0.99 ± 0.10	0.67 ± 0.33	0.99 ± 0.08	0.80 ± 0.19
[03] Clipping Suture	1.00 ± 0.07	0.88 ± 0.12	0.99 ± 0.10	0.71 ± 0.28	1.00 ± 0.06	0.91 ± 0.09
[04] Debris Wash	1.00 ± 0.01	0.50 ± 0.50	1.00 ± 0.01	0.50 ± 0.50	1.00 ± 0.01	0.50 ± 0.50
[05] Deflate Rectum	0.99 ± 0.07	0.89 ± 0.10	0.99 ± 0.10	0.77 ± 0.23	0.99 ± 0.08	0.89 ± 0.11
[06] Dissection	0.93 ± 0.26	0.87 ± 0.08	0.87 ± 0.34	0.76 ± 0.16	0.93 ± 0.26	0.86 ± 0.10
[07] Fluid Wash	1.00 ± 0.05	0.76 ± 0.24	1.00 ± 0.06	0.78 ± 0.22	1.00 ± 0.06	0.73 ± 0.27
[08] Haemostatis	1.00 ± 0.03	0.85 ± 0.15	1.00 ± 0.03	0.79 ± 0.21	1.00 ± 0.03	0.85 ± 0.15
[09] Inflate Rectum	1.00 ± 0.07	0.83 ± 0.17	0.99 ± 0.08	0.70 ± 0.29	1.00 ± 0.06	0.87 ± 0.13
[10] Instrument Positioning	0.91 ± 0.29	0.82 ± 0.12	0.81 ± 0.39	0.67 ± 0.21	0.89 ± 0.32	0.80 ± 0.14
[11] Marking	1.00 ± 0.07	0.91 ± 0.08	0.99 ± 0.10	0.80 ± 0.20	0.99 ± 0.07	0.90 ± 0.09
[12] No Action	0.97 ± 0.18	0.87 ± 0.11	0.93 ± 0.25	0.77 ± 0.20	0.96 ± 0.20	0.85 ± 0.13
[13] Out of Body	0.99 ± 0.09	0.98 ± 0.02	0.99 ± 0.11	0.96 ± 0.03	0.99 ± 0.09	0.97 ± 0.02
[14] Retraction	0.95 ± 0.22	0.78 ± 0.19	0.92 ± 0.26	0.66 ± 0.30	0.94 ± 0.24	0.77 ± 0.20
[15] Scope Insertion	1.00 ± 0.07	0.95 ± 0.05	0.99 ± 0.07	0.94 ± 0.06	1.00 ± 0.06	0.96 ± 0.04
[16] Scope Positioning	0.96 ± 0.19	0.87 ± 0.11	0.92 ± 0.27	0.73 ± 0.23	0.96 ± 0.19	0.87 ± 0.11
[17] Scope Removal	1.00 ± 0.03	1.00 ± 0.00	1.00 ± 0.05	0.99 ± 0.01	1.00 ± 0.03	1.00 ± 0.00
[18] Smoke	1.00 ± 0.07	0.87 ± 0.13	0.99 ± 0.10	0.71 ± 0.28	0.99 ± 0.08	0.82 ± 0.17
[19] Specimen Removal	1.00 ± 0.03	0.97 ± 0.03	1.00 ± 0.04	0.95 ± 0.05	1.00 ± 0.02	0.99 ± 0.01
[20] Stitching	0.98 ± 0.15	0.93 ± 0.06	0.95 ± 0.21	0.85 ± 0.12	0.97 ± 0.16	0.92 ± 0.06
[21] Washout	0.99 ± 0.11	0.91 ± 0.08	0.97 ± 0.17	0.79 ± 0.20	0.99 ± 0.12	0.90 ± 0.09

Table 4. Performance of the STALNet model on Surgical Actions across different encoders.

The `STALNet` model was also evaluated for its performance on each of the taxonomy triplets (phase, task, action) as shown in Tables 2, 3, and 4, respectively. The evaluation of phase segmentation reveals that the model performs exceptionally well across all phases, with only minor fluctuations in performance using different encoders. The ROC curves show its efficacy across these triplet behaviours (see Fig. 6). For example, the "Dissection" phase achieved an F1 score of 99.0% with no variance and an accuracy of 99.0% with a variance of 11.0% with the SWIN V2 encoder. Similarly, the "Setup" phase showed high performance with an F1 score of 98.0% and an accuracy of 99.0%, both exhibiting low variances (1% and 9%, respectively with ConvNeXt and SWIN V2 encoders). Even the "Closure" phase, despite being one of the more challenging phases due to its fewer instances, maintained an F1 score and accuracy of 100% for both with variances of 0% and 5%, respectively with the SWIN V2 encoder. These results indicate that the model effectively captures and segments the different phases consistently across three distinct encoders. In task segmentation, the model showed strong and consistent performance across most tasks. For instance, tasks such as "Longitudinal Muscle Dissection" and "Suturing" achieved high F1 scores of 99% for each, with accuracies of 100% and 99%, and low variances (1% and 0%, and 7% and 8%, respectively) with the ConvNeXt encoder. This consistency reflects the model's robust ability to segment tasks accurately. Conversely, the "Site" task, which had a significantly lower F1 score of 67% with high variance 33% with the ConvNeXt encoder. This indicates that the model struggles more with tasks that are less frequently represented in the dataset. For action segmentation, the model demonstrated high performance on

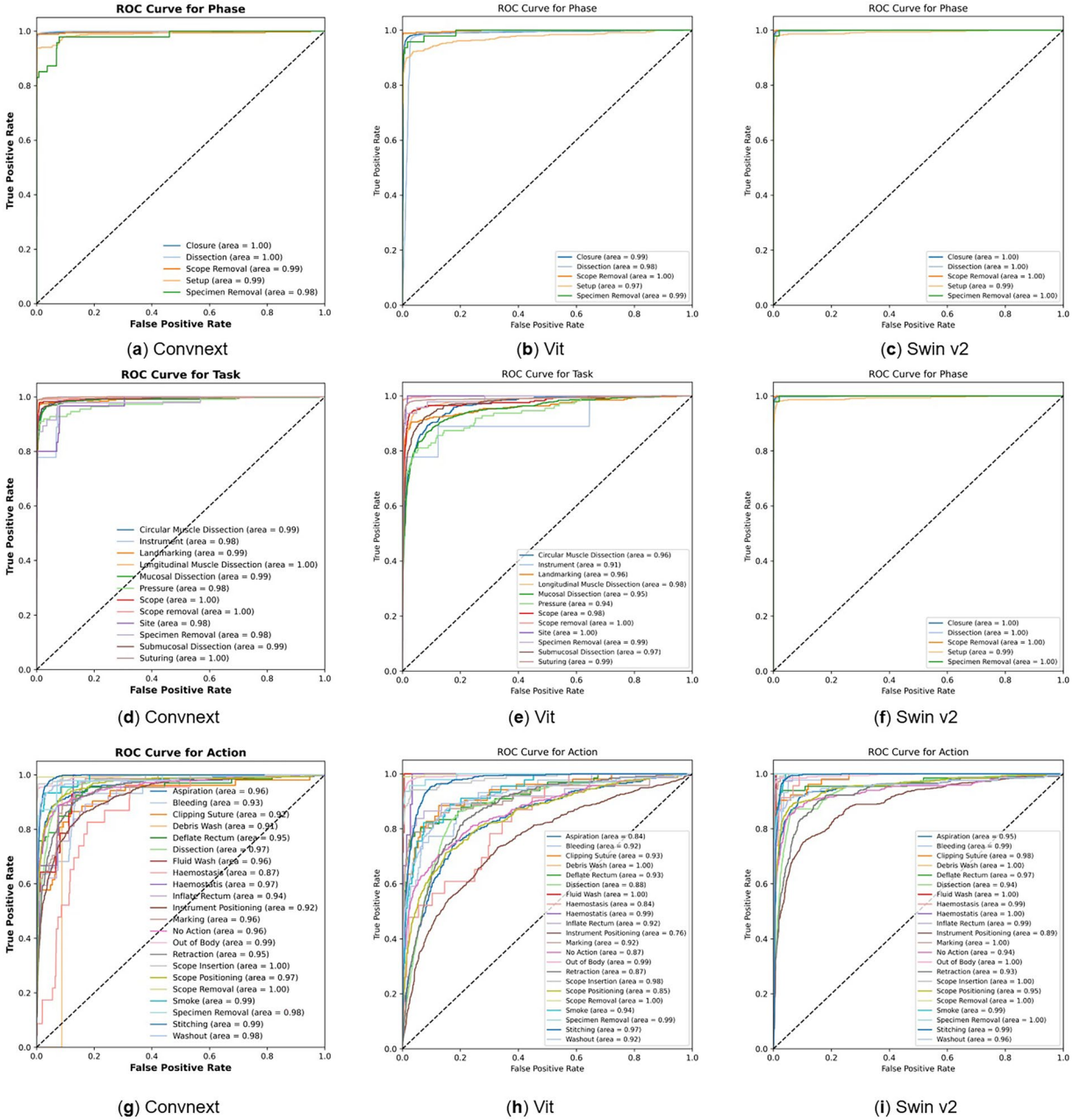

Fig. 6 STALNet Performance Review using ROC Curves for Taxonomy Triplets. The top row of ROC curves shows the performance of ConvNeXt, ViT and SWIN V2 encoders on labelling high level TEMS surgical "Phases". The next two rows show the performance of STALNet encoders on labelling TEMS surgical "Tasks" (intermediate level) and "Actions" (the fine level).

frequently occurring actions such as "Scope Insertion" and "Stitching" achieving F1 scores of 99% and 95%, and accuracies of 100% and 98%, respectively with the ConvNeXt encoder. The variances for "Scope Insertion" were 1% for the F1 score and 3% for accuracy, while "Stitching" had variances of 4% and 15%, indicating stable and reliable performance. However, actions like "Debris Wash" and "Haemostatis," which had lower F1 scores of 50% for each, also exhibited higher variances 50% for each of the above actions with the ConvNeXt encoder. These findings suggest that the model's performance is consistent for well-represented actions, but struggles with less frequent actions.

In summary, our technical validation is deliberately structured to demonstrate the effectiveness of STALNet's multi-target modelling strategy, which offers superior performance and semantic consistency compared to flat single-label approaches. In early experiments, we trained STALNet as a single-label classifier across all 84 triplet combinations. This unitarget formulation consistently plateaued at ~ 72% accuracy and struggled to model the underlying dependencies between triplet components. While it did not produce invalid triplets—since each output class was predefined—it lacked interpretability and failed to generalise well to complex surgical workflows.

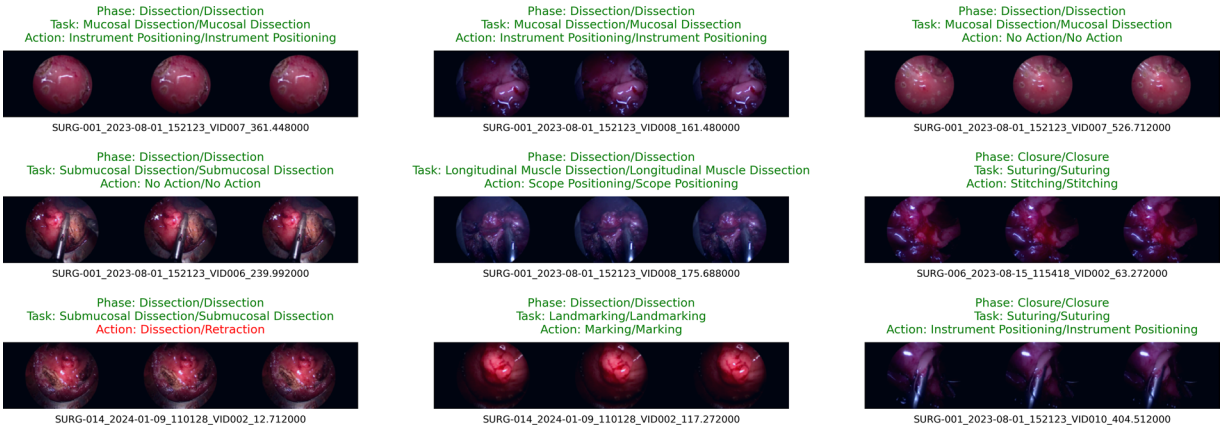

Fig. 7 STALNet: Batch of results for visual inspection. This figure illustrates the output of the STALNet model compared to human annotations—the ground truth (GT). Each tile displays the first, middle, and last frames of a video clip, along with predictions and GT for each taxonomy triplet (Phase, Task, Action) at the top. Green font indicates agreement with the GT, while red font indicates disagreement. In this example, there is widespread agreement except for one microclip where the model predicted the action "retraction" instead of "dissection" as labeled by the human annotators.

We also explored a multi-head architecture without tailored loss weighting. This improved expressiveness but still resulted in clinically implausible combinations, as the model lacked guided supervision to respect the hierarchical structure between phases, tasks, and actions. Our final multi-target approach, with three prediction heads and tailored loss functions for each triplet component, enabled the model to learn semantic relationships across components. This design achieved up to 91.7% accuracy and 86.0% F1 score on individual elements (see Tables 1 to 4), while effectively avoiding unrealistic triplet outputs by learning their internal structure. Although the results are shown in separate tables for interpretability, they originate from a single, unified model trained jointly with a triplet-aware loss.

The results confirm that the `STALNet` model with the ConvNeXt encoder performs well and consistently across phases, tasks, and actions with sufficient training data, as evidenced by low variance in well-represented classes. However, as the number of classes increases—from five phases to 11 tasks to 21 actions—the modelling task becomes more challenging, leading to higher variance and lower performance for less frequent classes. This trend underscores the complexity of handling a larger number of classes and highlights the need to address class imbalance. Techniques such as weighted dataloaders and customised loss functions can mitigate these issues, improving the model's robustness and performance across all categories.

The results also illustrate the model's superior capabilities in capturing the nuances of surgical workflows. The ROC curves highlights that the Swin V2 encoder outperforms other encoders in terms of accuracy and F1 score. The model's output is visually depicted in an infographic in Fig. 7. This shows the input video clips with predicted and actual taxonomy triplet labels from a batch. This visualisation clearly demonstrates the trends discussed in the performance tables and ROC curves, providing a comprehensive understanding of the model's efficacy in real-world scenarios.

The focus of this study was to provide a high-fidelity resource that enables the development of AI models for accurate surgical video indexing, such as our proposed STALNet architecture. While the objective is not to directly evaluate models for upstream tasks like surgical skill assessment—which require deeper reasoning and semantic understanding—this foundational work is essential for enabling scalable retrospective video analysis and supporting future clinical applications. To support this, the structured phase-task-action triplet taxonomy was co-designed with a panel of expert colorectal surgeons, aiming not only to capture workflow granularity but also to embed clinically meaningful signals that could potentially serve as proxies for surgical competence. For example, metrics derived from such factors as the frequency of intraoperative adverse events (e.g., bleeding), the length of inactive periods ("no action"), or the volatility of phase transitions—could, in future studies, be investigated as indicators of procedural fluency or surgeon expertise. These hypotheses are particularly relevant for distinguishing between experienced and novice operators, as variability in temporal workflow progression may indeed reflect differences in training or technical confidence.

Usage Notes

The dataset described in this study is available and is designed to facilitate the training and evaluation of machine learning models for surgical timeline segmentation based on the proposed taxonomy. Users can use several software packages to analyse and process the dataset, with Python being particularly useful for data handling, preprocessing, and model training. Key libraries include *FFmpeg* and *av* for video processing and frame extraction, timm for accessing various pre-trained models, *PyTorch* for deep learning model implementation and training, *fastai* for simplifying the training process and integrating with *PyTorch, nbdev* for creating reproducible and literate programs, and *Matplotlib* for visualising data and model performance. It is recommended to normalise the video frames to standardise the input data. Microclips can be generated based on custom logic

using tools like FFmpeg. Additionally, *Matplotlib* and *pandas* can be used to analyse data distribution and class imbalance in the dataset.

When integrating or comparing this dataset with others, it is essential to ensure consistent preprocessing steps to maintain uniformity. Utilising common evaluation metrics can help effectively compare performance across different datasets. Considering the temporal nature of surgical workflow data when combining it with static datasets is crucial to preserve contextual information. This project has been formally registered, and patients have given consent for their fully anonymised data to be shared openly to support surgical quality improvement, education and research. All consent forms have been checked by the core clinical team. All data has been fully anonymised in line with UK GDPR and NHS information governance standards, ensuring that individuals cannot be identified. Surgeons and data scientists can access the dataset at the following link(after signing the data sharing agreement): https://zenodo.org/records/14016844. By following these guidelines and using the tools and recommendations provided, researchers can effectively leverage this dataset to advance the field of surgical timeline segmentation and related applications. For additional resources, code, and tools, please refer to the Code Availability section.

In this study, the proposed timeline segmentation model has been employed to index a large number of trans-anal endoscopic microsurgery procedures, potentially creating an intuitive front-end platform for surgeons, educators and surgical training committees to analyze surgical videos effectively in a time-effective manner. The methodology described here is generalizable and can be used in any form of surgery where video recording is performed. It is possible to create search capabilities of the ESV searching platform, which leverages the timeline segmentation models to efficiently analyze multipart surgical videos of a single video or across a large library of videos. Surgeons will have the ability to search within a single video or across their entire "personal" surgical video bank using the timeline labels generated by the model. Taking this to the next level, service evaluation, training or NHS governance committees would be able to do this at scale to ensure quality of surgical procedures are maintained at scale and across surgical domains. This personalised searching capability is crucial for improving surgical technique and demonstrating effectiveness in various governance tasks, such as service evaluation or appraisal.

Currently, trainees or surgeons do not routinely submit VBAs for appraisal. This is polyfactorial but may be due to video file size, difficulty scrolling through large videos to find "key steps" and lack of a reliable standardised process to index videos. The model demonstrated here has the potential for clinicians to use STALNet to index their videos so that they can quickly locate video clips of specific intra-operative surgical events from large video datasets. Once the model has been validated in a clinical trial, a future project may focus on video library analysis to identify key behaviors that can be modified to improve future surgical performance. The timeline also enables comparisons between surgeons based on their surgical behavior, task efficiency, end-to-end operative progression, and intraoperative risk management. Any future clinical feedback system should allow users to reliably and securely filter and sort videos by type, speciality, and hospital, providing a powerful tool for detailed clinical data science analysis. This sets the foundation for continuous improvements in surgical practice in a large cohort of surgeons. The possibilities offered by this system are vast, empowering clinicians to conduct comprehensive reviews of intra-operative tasks to improve surgical outcomes for patients.

Code availability

The necessary scripts used in the generation and processing of the dataset for this study is available in the GitHub repository at https://github.com/bilalcodehub/evr. This repository contains all the necessary scripts and tools for working with the dataset. Included are data preprocessing scripts for normalising video frames, generating microclips using *FFmpeg*, and handling data distribution and class imbalance with *pandas* and *Matplotlib*. Additionally, the repository provides model training and evaluation scripts for implementing and training deep learning models using *PyTorch* and *fastai*, with configurations for integrating pre-trained models from timm. There are also tools for evaluating model performance using metrics like accuracy, F1 score, and ROC curves, as well as for visualising data and model results with *Matplotlib*.To ensure reproducibility, nbdev scripts for creating reproducible and literate programs are included.

The repository also provides detailed documentation on the versions of software used and instructions on how to set up and run the scripts. Specific variables and parameters used to generate, test, and process the current dataset are provided within the scripts, ensuring the study can be replicated accurately. We aim to facilitate the reuse of our dataset and the replication of our study, allowing other researchers to build upon our work in the field of surgical timeline segmentation. For further assistance, please refer to the documentation in the GitHub repository or contact the corresponding authors.

Received: 12 November 2024; Accepted: 18 July 2025;
Published online: 14 August 2025

References

1. Weiser, T. G. *et al.* Estimate of the global volume of surgery in 2012: an assessment supporting improved health outcomes. *The Lancet* **385**, S11 (2015).
2. Nepogodiev, D. *et al.* Global burden of postoperative death. *The Lancet* **393**, 401 (2019).
3. England, N. Secure data environment. www.digital.nhs.uk/services/secure-data-environment-service (2021).
4. NHS, N. Z. Delivering a net zero nhs. https://www.england.nhs.uk/greenernhs/a-net-zero-nhs/ (2020).
5. Maier-Hein, L. *et al.* Surgical data science for next-generation interventions. *Nature Biomedical Engineering* **1**, 691–696 (2017).
6. Reiley, C. E., Lin, H. C., Yuh, D. D. & Hager, G. D. Review of methods for objective surgical skill evaluation. *Surgical endoscopy* **25**, 356–366 (2011).
7. Goodman, E. D. *et al.* A real-time spatiotemporal ai model analyzes skill in open surgical videos. *arXiv preprint arXiv:2112.07219* (2021).

8. Padoy, N. Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy & Allied Technologies* **28**, 82–90 (2019).
9. Huaulmé, A. *et al*. Offline identification of surgical deviations in laparoscopic rectopexy. *Artificial Intelligence in Medicine* **104**, 101837 (2020).
10. Kadkhodamohammadi, A. *et al*. Towards video-based surgical workflow understanding in open orthopaedic surgery. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* **9**, 286–293 (2021).
11. Holden, M. S. *et al*. Feasibility of real-time workflow segmentation for tracked needle interventions. *IEEE Transactions on Biomedical Engineering* **61**, 1720–1728 (2014).
12. Padoy, N. *et al*. Statistical modeling and recognition of surgical workflow. *Medical image analysis* **16**, 632–641 (2012).
13. Lin, H. C. *et al*. Automatic detection and segmentation of robot-assisted surgical motions. In *International conference on medical image computing and computer-assisted intervention*, 802–810 (Springer, 2005).
14. Dergachyova, O., Bouget, D., Huaulmé, A., Morandi, X. & Jannin, P. Automatic data-driven real-time segmentation and recognition of surgical workflow. *International journal of computer assisted radiology and surgery* **11**, 1081–1089 (2016).
15. Jin, Y. *et al*. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging* **37**, 1114–1126 (2017).
16. Blum, T., Feußner, H. & Navab, N. Modeling and segmentation of surgical workflow from laparoscopic video. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010, Proceedings, Part III 13*, 400–407 (Springer, 2010).
17. Twinanda, A. P. *et al*. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**, 86–97 (2016).
18. Ramesh, S. *et al*. Multi-task temporal convolutional networks for joint recognition of surgical phases and steps in gastric bypass procedures. *International journal of computer assisted radiology and surgery* **16**, 1111–1119 (2021).
19. Gao, X., Jin, Y., Long, Y., Dou, Q. & Heng, P.-A. Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24*, 593–603 (Springer, 2021).
20. Funke, I., Rivoir, D.,, Krell, S. & Speidel, S. TUNeS: A Temporal U-Net With Self-Attention for Video-Based Surgical Phase Recognition, *in IEEE Transactions on Biomedical Engineering*, **72**(7), pp. 2105-2119 (2025).
21. Huber, M. Video-based content analysis. In Campbell, A. G., Hong, L., Meinel, F. & Zallio, M. (eds.) *Handbook of Research on Multimodal Human Computer Interaction and Pervasive Services*. https://www.taylorfrancis.com/chapters/edit/10.4324/9780429316647-5/video-based-content-analysis-matthias-huber (CRC Press, 2020).
22. Liu, L. & Özsu, M. T. *Encyclopedia of database systems*, vol. 6 (Springer New York, 2009).
23. Feldman, L. S. *et al*. SAGES Video-Based Assessment (VBA) program: a vision for life-long learning for surgeons. *Surgical endoscopy* **34**, 3285–3288 (2020).
24. Vercauteren, T., Unberath, M., Padoy, N. & Navab, N. CAI4CAI: the rise of contextual artificial intelligence in computer-assisted interventions. *Proceedings of the IEEE* **108**, 198–214 (2019).
25. Nwoye, C. I. *et al*. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23*, 364–374 (Springer, 2020).
26. Mascagni, P. *et al*. Computer vision in surgery: from potential to clinical value. *npj Digital Medicine* **5**, 163 (2022).
27. Lewandrowski, K.-U. *et al*. Regional variations in acceptance, and utilization of minimally invasive spinal surgery techniques among spine surgeons: results of a global survey. *Journal of Spine Surgery* **6**, S260 (2020).
28. Richards, M. K. *et al*. A national review of the frequency of minimally invasive surgery among general surgery residents: assessment of ACGME case logs during 2 decades of general surgery resident training. *JAMA surgery* **150**, 169–172 (2015).
29. Paysan, D., Haug, L., Bajka, M., Oelhafen, M. & Buhmann, J. M. Self-supervised representation learning for surgical activity recognition. *International Journal of Computer Assisted Radiology and Surgery* **16**, 2037–2044 (2021).
30. Valderrama, N. *et al*. Towards holistic surgical scene understanding. In *International conference on medical image computing and computer-assisted intervention*, 442–452 (Springer, 2022).
31. Ayobi, N. *et al*. Pixel-wise recognition for holistic surgical scene understanding. *arXiv preprint arXiv:2401.11174* (2024).
32. Bawa, V. S. *et al*. The saras endoscopic surgeon action detection (esad) dataset: Challenges and methods. *arXiv preprint arXiv:2104.03178* (2021).
33. Nwoye, C. I. *et al*. CholecTriplet2022: Show me a tool and tell me the triplet—an endoscopic vision challenge for surgical action triplet detection. *Medical Image Analysis* **89**, 102888 (2023).
34. Murali, A. *et al*. The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *arXiv preprint arXiv:2312.12429* (2023).
35. Nwoye, C. I., Zaid, F., Lavanchy, J. & Padoy, N. Cholectrack20: A multi-perspective tracking dataset for surgical tools. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, Wed June 11th-Sun June 15th, 2025* (2025).
36. Ye, Z. *et al*. A comprehensive video dataset for surgical laparoscopic action analysis. *Scientific Data* **12**, 1–10 (2025).
37. Bach, S. P. *et al*. Radical surgery versus organ preservation via short-course radiotherapy followed by transanal endoscopic microsurgery for early-stage rectal cancer (trec): a randomised, open-label feasibility study. *The Lancet Gastroenterology & Hepatology* **6**, 92–105 (2021).
38. Gurevych, I., De Castilho, R. E. & Biemann, C. Webanno: A flexible, web-based and visually supported system for distributed annotations. *51st Annual Meeting …* (2013).
39. Ashraf, S. & Muhammad, B. Densely annotated transanal endoscopic microsurgery dataset for timeline segmentation https://doi.org/10.5281/zenodo.14016844 (2024).
40. Liu, Z. *et al*. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).
41. Liu, Z. *et al*. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019 (2022).
42. Steiner, A. *et al*. How to train your vit? data, augmentation, and regularization in vision transformers. arXiv657 preprint arXiv:2106.10270 (2021).
43. Caron, M. *et al*. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660 (2021).
44. Howard, J. & Gugger, S. Fastai: a layered api for deep learning. *Information* **11**, 108 (2020).

Acknowledgements

The authors express their deepest gratitude to the patients who consented to have their procedures recorded and made open source for educational, service improvement, and research purposes. The authors thank the University Hospitals Birmingham (UHB) CARMS and data team for guidance. We also extend our thanks to the theatre and anaesthetic staff at Good Hope Hospital, (UHB, UK), whose dedication over the past 20 years has established the institution as a world-class centre for TEMS. In addition, we are grateful to all UHB staff for

fostering an environment that embraces and supports service improvement and innovation. Shazad Ashraf has a Birmingham Health Partnership fellowship that is supported by Metchley Park Medical Society, a charity dedicated to supporting clinical research, education and innovation in Birmingham. Andrew Beggs is funded by an MRC Senior Clinical Fellowship Award (ref MR/X006433/1). The authors also acknowledge the support of Birmingham City University for enabling protected time for Muhammad Bilal to undertake this research and advance collaborative efforts.

Author contributions

M.B. and S.A. conceived the project idea, S.A., D.B., S.K., N.L., Mo.A., S.B., and A.B. were clinical domain experts and assisted in labelling and validation, M.B., Ma.A., A.H., K.S., I.Q., P.C., Z.K., A.Q., A.B., J.Q. and S.A. conducted the experiment design and analysis, M.B., Ma.A., A.H., H.V., M.C., A.Q., J.Q. and S.A. analysed the results. M.B., Ma.A. and S.A. drafted the first version of the manuscript. All authors reviewed and helped edit the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.B. or S.A.

Reprints and permissions information is available at www.nature.com/reprints.

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