Enhancing Surgical Documentation through Multimodal Visual-Temporal Transformers and Generative AI

Cristian Cosentino* CC2308@CAM.AC.UK, CRISTIAN.COSENTINO@DIMES.UNICAL.IT
DIMES, University of Calabria
Department of Computer Science, University of Cambridge

Hugo Georgenthum* HUGO.GEORGENTHUM@OUTLOOK.COM
Department of Computer Science, ETH Zürich
Department of Computer Science, University of Cambridge

Fabrizio Marozzo FMAROZZO@DIMES.UNICAL.IT
DIMES, University of Calabria

Pietro Liò PIETRO.LIO@CL.CAM.AC.UK
Department of Computer Science, University of Cambridge

Abstract

Automatic summarization of surgical videos is critical for improving procedural documentation, supporting surgical training, and facilitating post-operative analysis. Despite recent advances in computer vision and natural language processing, most existing methods either focus on tool detection or clip-level captioning, lacking an integrated approach that produces full, clinically meaningful reports.

We introduce a multimodal framework that leverages visual transformers and large language models to generate comprehensive surgical video summaries. The method unfolds in three stages: (i) extraction of frame-level features to capture tools, tissues, and surgical actions, (ii) integration of temporal context through a ViViT-based encoder combined with frame-level captions, and (iii) synthesis of clip-level descriptions into structured surgical reports using a dedicated LLM.

We evaluate the framework on the CholecT50 dataset of 50 laparoscopic videos, achieving 96% precision in tool detection and a BERT score of 0.74 for temporal summarization. These results demonstrate the potential of combining computer vision and language models to advance AI-assisted reporting, offering a step toward reliable, interpretable, and efficient clinical documentation.

* Hugo conducted all the experiments, tests, and code implementation. Cristian contributed to the conceptualization, theoretical analysis, and manuscript writing.

Data and Code Availability — The CholecT50 dataset used in this study is publicly available from the official challenge organizers. An anonymized repository containing the code for preprocessing, model training, and evaluation is available at https://github.com/criscose/Enhancing-Surgical-Documentation-through-Multimodal-Visua Permanent links to the full repository will be provided in the camera-ready version.

Institutional Review Board (IRB) This study did not require IRB approval as it uses only publicly available, anonymized data.

Keywords: Surgery video, report generation, vision transformer, large language models

1. Introduction

Artificial Intelligence (AI) is increasingly shaping modern surgery by enhancing precision, supporting decision-making, and improving patient outcomes (Hashimoto et al., 2018). Computer vision, in particular, enables machines to interpret surgical scenes, with Convolutional Neural Networks (CNNs) widely used for tasks such as tool tracking and image-based diagnostics (Litjens et al., 2017). However, CNNs are limited in modeling long-range temporal dependencies that are critical for understanding complex surgical procedures (Zia et al., 2018).

Recent advances in transformer architectures have addressed this limitation. Originally developed for Natural Language Processing, transformers are now

the backbone of powerful Large Language Models (LLMs) (Brown et al., 2020) and visual transformers (Dosovitskiy et al., 2020), offering unified representations for sequential text and video data. These capabilities open new opportunities for multimodal systems in surgical video analysis.

Despite this progress, deploying AI in clinical practice demands models that are accurate, transparent, and ethically compliant. Systems must preserve patient privacy, minimize bias, and remain under human oversight. Modular and explainable designs are therefore essential to ensure reliability and clinical trust (Holzinger et al., 2017).

In this work, we propose a multimodal pipeline that integrates computer vision and language models to automatically generate structured surgical reports from video. The approach consists of three stages: (i) extraction of frame-level features, (ii) temporal encoding and clip-level summarization via a ViViT-based encoder, and (iii) aggregation of summaries into coherent surgical reports using a specialized LLM. Evaluated on annotated laparoscopic surgery videos, our method achieves strong performance in tool detection and temporal summarization, demonstrating the effectiveness of this multistage strategy.

Such summarization is clinically significant because it transforms long and complex surgical recordings into concise, structured narratives that facilitate documentation, training, and postoperative review. By reducing redundancy and highlighting key events, automated video summarization can directly support both clinical practice and surgical education.

The remainder of the paper is organized as follows. Section 2 reviews related work in surgical video analysis and summarization. Section 3 presents our methodology. Section 4 details the experimental setup and results. Section 5 concludes with contributions and future directions.

2. Related Work

Automatic analysis of surgical videos is increasingly explored for training, skill assessment, and clinical practice improvement Loukas (2018); Kawka et al. (2022). Advances in computer vision and the availability of intraoperative recordings enable semantic extraction from complex visual content, though challenges persist due to variability, occlusions, and spatio-temporal complexity Jin et al. (2020).

Early static feature approaches proved insufficient for modeling surgical actions Mao et al. (2022). Deep learning methods, evolving from CNNs to Vision Transformers, improved long-range dependency modeling Islam et al. (2021). Object detection remains central, with YOLO, Faster R-CNN, and transformer-based detectors widely applied to tools and anatomical structures Fu et al. (2018); Wang et al. (2025).

Captioning techniques extend these efforts, from frame-level descriptions to clip-level and dense captioning, which enhance temporal flow and granularity Zhang et al. (2019, 2020b); Kashid et al. (2024). Multimodal alignment between vision and language improves clinical relevance, though coherence and terminology remain challenges Chen et al. (2023); Sharma et al. (2023). Cross-modal attention architectures such as CroMA Antonio et al. (2024) and multimodal frameworks tailored to surgical reporting Wang et al. (2025) highlight progress in this direction.

LLMs further support fluent, structured narratives Tian et al. (2023); Sloan et al. (2024), and end-to-end pipelines now integrate detection, captioning, and temporal modeling for report generation Xu et al. (2021); Bai et al. (2024). However, most approaches still address these components in isolation, limiting interpretability and clinical integration. Our work builds on these advances with a unified multimodal pipeline (Figure **??**) that fuses frame- and clip-level captioning with LLMs to produce structured, explainable surgical reports.

3. Proposed Methodology

This section describes our methodology for the automated generation of explainable reports from surgical video analysis. The objective is to interpret surgical activities at multiple levels of granularity and to produce coherent, human-readable summaries that combine visual evidence with natural language descriptions. The approach is organized into four phases: (i) object detection in frames, (ii) frame-level captioning, (iii) clip-level captioning, and (iv) report synthesis. The overall execution flow is shown in Figure 1.

The first phase detects essential visual elements in each frame, such as instruments, organs, and anatomical landmarks. Given the challenges of surgical imagery—tight spaces, occlusions, and motion blur—we employ robust detectors adapted to this domain, including transformer-based models, to provide the visual context required for later stages.

The second phase generates captions at the frame level. By combining detected objects with frame features, transformer-based captioning models produce accurate textual descriptions that capture surgical actions and spatial relations within individual frames.

The third phase extends this analysis to the temporal domain. Videos are segmented into clips of 32 consecutive frames, allowing the model to capture action sequences and dependencies across time. Frame-level captions are incorporated as an additional modality, enabling temporal attention mechanisms to generate coherent summaries that reflect broader surgical events.

The final phase synthesizes a complete surgical report. Clip-level summaries are concatenated and refined by a large language model, which transforms them into a comprehensive narrative. This ensures coherence, interpretability, and clinical relevance, bridging granular model outputs with structured documentation that can be readily understood by medical professionals.

Through this multi-stage design—detection, frame and clip captioning, and report synthesis—the proposed methodology connects detailed video analysis with explainable reporting, improving both interpretability and practical utility in surgical contexts.

3.1. Multi-Label Object Detection Using Vision Transformers

We implement a multi-label object detection module based on Vision Transformers (ViTs), adapted to surgical contexts Dosovitskiy et al. (2020). The task is to simultaneously detect instruments, organs, and anatomical structures within each frame. Input frames are resized to 224×224 and split into non-overlapping 16×16 patches, yielding 196 tokens. Each patch is linearly projected and enriched with positional encodings before being processed by the transformer encoder.

The ViT outputs an embedding that is passed through a classification head with sigmoid activations, producing a probability vector P across object classes. Objects are selected when their probability exceeds a predefined threshold (empirically set to 0.5). This simple decision rule allows detection of multiple categories per frame.

To address the severe class imbalance typical of surgical datasets, we adopt a weighted Binary Cross-Entropy loss. Class-specific weights are computed as the inverse of class frequency and normalized to sum to one, ensuring rare but clinically important objects contribute proportionally to the optimization Cui et al. (2019).

3.2. Frame Caption Generation

After detecting visual entities within each frame, the next step is to generate descriptive captions that combine object-level semantics with visual context. Transformer-based architectures are adopted for their ability to capture complex spatial and semantic dependencies Vaswani et al. (2017); Li et al. (2020b). The module consists of four main components:

- **Frame Encoder:** A pre-trained Vision Transformer (ViT) encodes each frame into a latent representation, projected into a 512-dimensional space through a linear layer Dosovitskiy et al. (2020).
- **Object Encoder:** Detected object labels are processed with a transformer-based textual encoder (e.g., DistilBERT), producing embeddings mapped to the same latent space Devlin et al. (2018); Sanh et al. (2019a).
- **Feature Fusion:** Visual and textual features are combined through cross-modal attention, enabling joint reasoning across modalities Lu et al. (2019); Li et al. (2020a).
- **Caption Decoder:** A transformer-based decoder (e.g., T5) generates frame-level captions from the fused representation Anderson et al. (2018); Cornia et al. (2020).

Formally, given a frame I and the corresponding set of object labels O, visual and textual features are projected into a shared space and concatenated:

$$h'_I = W_I \,\mathrm{ViT}(I), \qquad h'_O = W_O \,E(O), \tag{1}$$

$$h'_C = \mathrm{CrossAttention}(\mathrm{concat}(h'_I, h'_O)), \tag{2}$$

$$C = D(h'_C), \tag{3}$$

here E and D denote the textual encoder and decoder, respectively. $W_I \in \mathbb{R}^{d_I \times d}$ and $W_O \in \mathbb{R}^{d_O \times d}$ are learnable linear projection matrices that map the outputs of the visual encoder (ViT) and textual encoder (DistilBERT), respectively, into a common latent space of dimension $d = 512$. This projection ensures dimensional alignment between modalities and allows the cross-attention module to operate effectively on concatenated visual and textual representations.

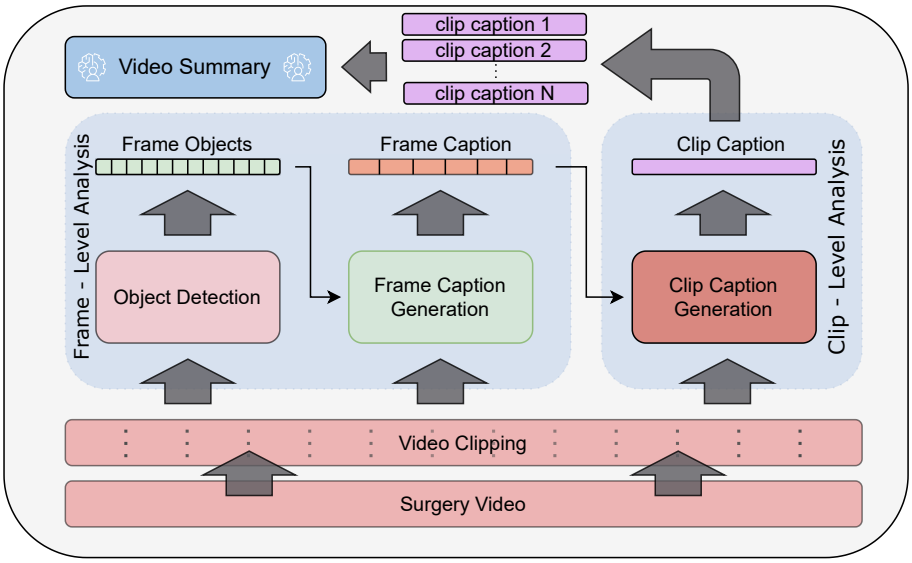


Figure 1: Execution flow of the proposed methodology.

This integration ensures that frame-level captions capture not only the appearance of instruments and structures but also the actions performed. By combining semantic and visual cues, the model produces clinically relevant descriptions that enhance interpretability and support surgical decision-making.

3.3. Clip-Level Caption Generation

Building on frame-level analysis, we extend the methodology to capture temporal dependencies across video clips. Clip-level captioning requires modeling actions and interactions that unfold over multiple frames, integrating temporal context into the generated descriptions. Inspired by transformer-based architectures for video understanding, particularly Video Vision Transformers (ViViT) Arnab et al. (2021), our approach introduces two key modifications:

1. **Temporal Integration**: Frame sequences are grouped into clips of N_f frames, each represented both visually and through the corresponding frame-level captions. This dual input provides multimodal context over time.

2. **Spatiotemporal Patches**: The vision transformer extends image patches into spatiotemporal tokens Bertasius et al. (2021), enabling the model to capture motion patterns and evolving surgical actions across frames.

This design allows the generation of temporally coherent and clinically meaningful clip-level captions, ensuring that procedures spanning multiple frames are accurately documented.

3.4. Comprehensive Surgical Report Synthesis

The final stage aggregates clip-level captions into a complete surgical report. Large language models (LLMs), particularly GPT-based architectures, are employed for their strength in summarization and narrative construction Radford et al. (2019); Ouyang et al. (2022). All generated captions are provided to the LLM with a tailored prompt guiding the production of a structured, clinically relevant summary. This ensures that procedural details are not only condensed but also organized in a way that aligns with medical documentation standards. By bridging fine-grained video analysis with high-level reporting, this phase enhances transparency, interpretability, and practical utility in clinical contexts. The complete prompt template used for report generation is available in Appendix A

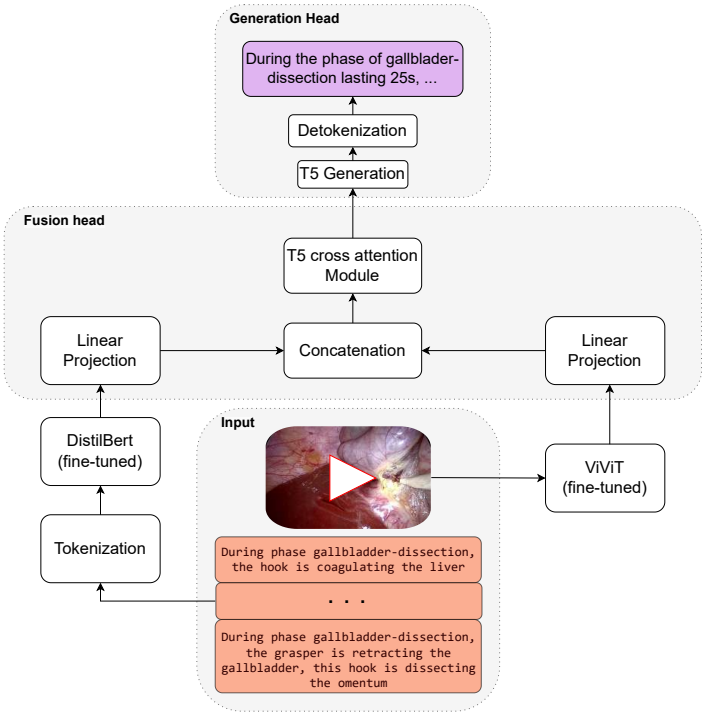


Figure 2: Scheme of the frame caption generation module, integrating visual and textual encoders with cross-modal fusion and a transformer decoder.

4. Experimental Results

We evaluate the proposed system through a structured experimental pipeline. First, we introduce the CholecT50 dataset and describe the preprocessing used to derive frame- and clip-level representations (Section 4.1). We then outline the experimental setup, including the pre-trained models selected for visual and textual feature extraction, and summarize the training strategy adopted to optimize each module while mitigating error propagation.

Evaluation is conducted across three core tasks: (i) object detection, (ii) frame-level captioning, and (iii) clip-level captioning. For each task, we report both quantitative metrics and qualitative examples (Sections 4.2–4.4). Finally, we assess the end-to-end generation of structured surgical reports and analyze their completeness and consistency (Section 4.5).

The dataset is licensed under CC BY-NC-SA 4.0 Creative Commons (2013). Code and trained models are provided in an anonymized repository, ensuring reproducibility while preserving the double-blind review process.

Figure 3 shows an illustrative example from a laparoscopic cholecystectomy video (`VID01`). The clip begins with the Preparation phase, where the grasper lifts and positions the gallbladder, followed by the Calot triangle dissection phase, characterized by the introduction of the hook instrument and dissection around the cystic duct and artery. The system's frame-level captions and tool detections capture these transitions, providing a qualitative demonstration of temporal coherence.

4.1. Dataset, Preprocessing, and Experimental Setup

We evaluate our framework on the CholecT50 dataset CAMMA (2020), which consists of 50 laparoscopic cholecystectomy videos recorded at 1 fps. Each frame is annotated with tool, action, target, and phase labels, yielding over 151k annotated triplets

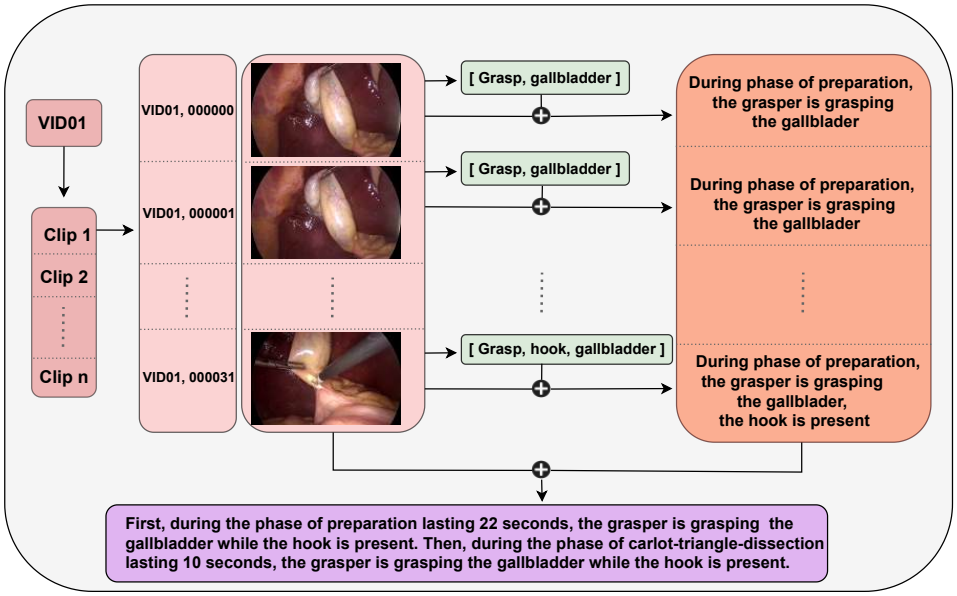
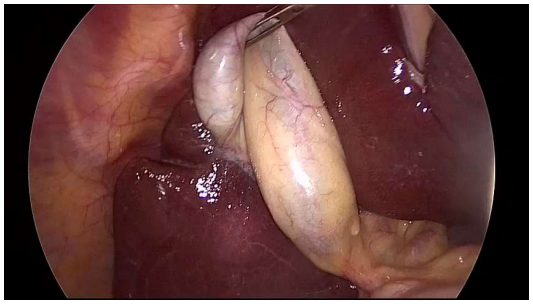


Figure 3: Illustrative walkthrough of the workflow on a sample surgical video clip.

from 6 instruments, 10 verbs, and 15 targets. Frames are grouped into 32-frame clips with 16-frame overlap, resulting in 6,232 clips. Since captions are not provided, we automatically generate frame- and clip-level captions from annotations by converting verb–target–instrument triplets into sentences and concatenating them temporally (Figure 4). Table 1 summarizes the distribution of surgical phases. In addition to the CholecT50 dataset, we further evaluated our framework on the MESAD dataset (available at: https://saras-mesad.grand-challenge.org/dataset/), which includes prostatectomy procedures recorded with the da Vinci Xi robotic system, both on real patients (MESAD-Real) and on artificial anatomies for surgical training (MESAD-Phantom). This second case study, characterized by a different surgical domain and specific action labels, was used to assess the generalizability of the methodology beyond laparoscopic cholecystectomy. Due to space constraints, the full results are reported in the Appendix (Section H), where we show that the model maintains consistent performance in this context as well, supporting its robustness and broader clinical applicability.



Detected objects: grasper, gallbladder
Frame caption: During preparation, the grasper holds the gallbladder.
Clip caption: In the preparation phase (22s), the grasper holds the gallbladder while the hook is present; during the following dissection (10s), the same tools remain active.

Figure 4: Example of artificial frame and clip caption generation.

Phase	Frames	Time (min)
Preparation	2,806	46.8
Calot-triangle-dissection	38,808	646.8
Clipping-and-cutting	7,790	129.8
Gallbladder-dissection	26,789	446.5
Gallbladder-packaging	3,790	63.2
Cleaning-and-coagulation	6,986	116.4
Gallbladder-extraction	2,858	47.6
Total	**89,927**	**1,498.8**

Table 1: Phase durations and frame counts in CholecT50.

Pre-trained Models. For visual encoding, we use ViT-base Dosovitskiy et al. (2021) for frames and ViViT-B Arnab et al. (2021) for spatiotemporal clips. For textual encoding of detected objects, we adopt DistilBERT Sanh et al. (2019b), while captions are generated with T5 models: T5-Small for frames and FLAN-T5-Base for clips Raffel et al. (2020). Final report synthesis is handled by GPT-4 OpenAI (2023), chosen for its superior summarization fluency compared to domain-specific LLMs (e.g., Med-PaLM Singhal et al. (2023), BioGPT Luo et al. (2022)).

Training Strategy. Each component (object detector, frame captioner, clip captioner) is trained independently and then fine-tuned in a cascaded manner to account for upstream errors. Training was performed on an NVIDIA A100 GPU with 40 GB RAM; the heaviest model (clip captioner) required ∼7 hours. Metric definitions and dataset details are provided in Appendix B and C

4.2. Object Detection

We first assess calibration and accuracy of the multi-label detector. The model is already well calibrated (ECE = 0.0075); post-hoc temperature scaling further reduces ECE to 0.0028 with $T = 1.8584$, mitigating over-confidence in mid–high confidence bins while preserving accuracy. Figure 5 shows this effect: before calibration (left), predicted scores overestimated accuracy, whereas after calibration (right), probabilities align closely with the diagonal, indicating more reliable outputs.

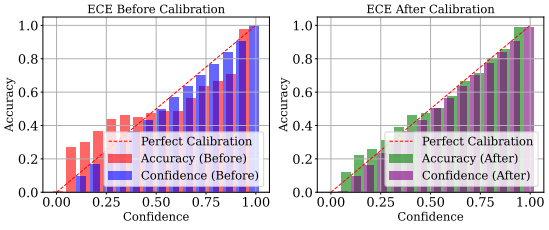


Figure 5: Reliability diagrams before (left) and after (right) temperature scaling. Dashed line: perfect calibration.

We benchmark our detector against SurgT Jin et al. (2021) and a recent diffusion-based model Baranchuk et al. (2022). SurgT is a transformer tailored for tool detection, while the diffusion approach adapts generative paradigms to recognition tasks. As shown in Figure 6, both achieve slightly higher mean Average Precision (mAP) on instruments, reflecting their specialization in detecting visually distinctive tools. However, they perform poorly on targets (SurgT mAP = 0.30), as neither was designed to capture instrument–target contextual relations. By integrating visual and contextual features, our detector yields substantially higher target mAP, demonstrating stronger modeling of surgical semantics.

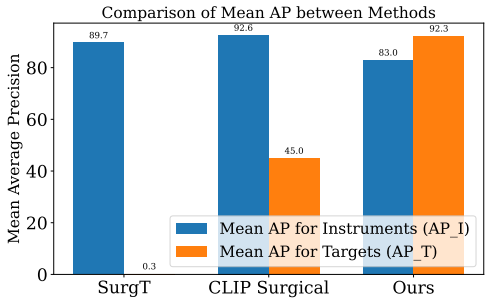


Figure 6: mAP comparison for instruments (AP_I) and targets (AP_T).

4.3. Frame Caption Generation

We evaluate two configurations: (i) trained with ground-truth objects, and (ii) a robust variant fine-tuned on detector outputs to account for upstream noise. Metrics include BLEU, ROUGE, and BERTScore.

Metric	Model	Robust Model
BLEU	0.6395	**0.7267**
ROUGE-1	0.8351	**0.8700**
ROUGE-2	0.7747	**0.8096**
ROUGE-L	0.8116	**0.8637**
BERT Precision	0.7771	0.7745
BERT Recall	0.7644	**0.8365**
BERT F1	0.7707	**0.8052**

Table 2: Frame-level captioning: initial vs. robust model.

The robust model improves BLEU and ROUGE and raises BERTScore Recall/F1, showing better cov-

erage and resilience to detector errors (qualitative examples in Appendix F).

4.4. Clip Caption Generation

We next evaluate the generation of *clip-level* descriptions, where sequences of frames are summarized into coherent captions. Table 3 compares four variants. We next evaluate the generation of *clip-level* descriptions, where sequences of frames are summarized into coherent captions. Table 3 compares four variants. The **Simple (vision-only)** configuration, which ignores frame captions, achieves the lowest scores across all metrics. The **Model (GT)** variant, based on ground-truth frame captions, yields strong gains, highlighting the benefit of high-quality intermediate text. When using automatically predicted captions (**Model (Generated)**), performance drops by approximately 0.02 compared to GT, due to error propagation. Finally, the **Robust** configuration, trained in two phases (independent → joint fine-tuning), achieves the best BLEU, ROUGE-L, and BERT F1 scores, demonstrating the most effective balance between lexical and semantic quality.

Frame captions	None	GT	Generated	Robust
BLEU	0.51	0.65	0.60	**0.67**
ROUGE-1	0.76	0.86	0.83	**0.87**
ROUGE-2	0.67	**0.80**	0.74	**0.80**
ROUGE-L	0.71	0.80	0.76	**0.83**
BERT Precision	0.67	**0.77**	0.67	0.74
BERT Recall	0.66	0.77	0.75	**0.78**
BERT F1	0.66	**0.77**	0.71	0.76

Table 3: Clip-level captioning results under different training strategies.

Overall, incorporating textual intermediates and robust training significantly improves performance, confirming the advantage of progressive integration across system components.

4.5. Structured Surgical Report Generation

Finally, we assess the ability to synthesize clip-level captions into full surgical reports. GPT-4, guided by a structured prompt (Section 3.4), produces summaries that maintain temporal consistency and include relevant instruments, anatomy, and phase transitions. Despite noise in intermediate captions, the LLM consolidates information, removes redundancy, and yields fluent, clinically coherent narratives.

Beyond the experiments presented here, we conducted additional ablation studies to isolate the role of temporal modeling, backbone choice, semantic signals, and clip length. These analyses consistently confirm the superiority of the proposed configuration. Furthermore, we introduce a hallucination audit to explicitly measure factual reliability in generated reports. Full details and results are reported in Appendix D. An example report for video `VID07` is shown in Appendix G, illustrating how the pipeline transforms segmented inputs into a comprehensive surgical narrative suitable for documentation.

In addition, we complemented these experiments with both a *domain expert review* and a *structured LLM-based evaluation*, which confirmed that the generated reports are close to human-written notes in terms of accuracy, completeness, and clinical utility. Detailed results of these qualitative evaluations are provided in Appendix I.

5. Conclusions and Limitations

This work presents a modular pipeline for automated surgical video analysis and report generation, integrating object detection, multimodal captioning, and large language models. By combining computer vision and natural language processing, the approach addresses a key limitation of prior CNN-based methods—capturing long-range dependencies and contextual information—and produces interpretable, clinically meaningful outputs.

Experiments on the CholecT50 dataset confirm the effectiveness of the methodology, with improvements of up to 12% in semantic precision and 30% in BLEU over baseline configurations. A second case study on the MESAD dataset further supports the generalizability of the approach across different surgical domains. These results highlight the value of decomposing the task into specialized subtasks while maintaining coherence through progressive integration. Ablation studies and a hallucination audit additionally demonstrate that each module contributes meaningfully to the final outcome and that the system mitigates risks of factual errors.

Future work will focus on broadening dataset diversity, extending temporal modeling, and incorporating additional modalities (e.g., segmentation, audio cues) to enrich contextual understanding. Beyond this, the modular design provides a natural foundation for integration into agent-based frameworks, where autonomous components—augmented by ex-

plainable AI techniques—could coordinate detection, captioning, and summarization under human oversight. Such systems would advance the development of reliable, transparent, and clinically viable AI assistants for surgical documentation and decision support.

Nevertheless, some limitations remain. The current system leverages synthetic captions for pretraining, which may not fully capture the nuances of clinical language; it also relies on GPT-4, a closed-source LLM, raising reproducibility concerns. Moreover, validation has so far been restricted to two datasets, underscoring the need for broader, multi-institutional clinical studies. Addressing these aspects will be crucial for ensuring scalability, transparency, and trustworthiness in real-world deployment.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018.

Greetta Antonio, Jobin Jose, Sudhish N George, and Kiran Raja. Croma: Cross-modal attention for visual question answering in robotic surgery. In *International Conference on Pattern Recognition*, pages 459–471. Springer, 2024.

Anurag Arnab et al. Vivit: A video vision transformer. *ICCV*, 2021.

Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. *arXiv preprint arXiv:2404.00578*, 2024.

Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-efficient diffusion models for surgical video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3866–3875, 2022.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Space-time attention networks for video understanding. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 937–947, 2021.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

CAMMA. Cholect50: A dataset for surgical video understanding, 2020. URL https://github.com/CAMMA-public/cholect50. Accessed: 2025-03-08.

Zhen Chen, Qingyu Guo, Leo KT Yeung, Danny TM Chan, Zhen Lei, Hongbin Liu, and Jinqiao Wang. Surgical video captioning with mutual-modal concept alignment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–34. Springer, 2023.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10578–10587, 2020.

Creative Commons. Attribution-noncommercial-sharealike 4.0 international (cc by-nc-sa 4.0). https://creativecommons.org/licenses/by-nc-sa/4.0/, 2013. Accessed: 2025-04-18.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani,

Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

Kun Fu, Jin Li, Junqi Jin, and Changshui Zhang. Image-text surgery: Efficient concept learning in image captioning by generating pseudopairs. *IEEE transactions on neural networks and learning systems*, 29(12):5910–5921, 2018.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

Daniel A. Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R. Meireles. Artificial intelligence in surgery: Promises and perils. *Annals of Surgery*, 268(1):70–76, July 2018. doi: 10.1097/SLA.0000000000002693. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5995666/.

Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

Saiful Islam, Aurpan Dash, Ashek Seum, Amir Hossain Raj, Tonmoy Hossain, and Faisal Muhammad Shah. Exploring video captioning techniques: A comprehensive survey on deep learning methods. *SN Computer Science*, 2(2):1–28, 2021.

Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical image analysis*, 59:101572, 2020.

Yueming Jin, Qi Dou, Hao Chen, and Pheng-Ann Heng. Surgt: Transformer-based surgical tool detection in laparoscopic videos. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 620–629. Springer, 2021.

Shamal Kashid, Lalit K Awasthi, Krishan Berwal, and Parul Saini. Stvs: Spatio-temporal feature fusion for video summarization. *IEEE MultiMedia*, 2024.

Michal Kawka, Tamara MH Gall, Chihua Fang, Rong Liu, and Long R Jiao. Intraoperative video analysis and machine learning models will change the future of surgical training. *Intelligent Surgery*, 1:13–15, 2022.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 11336–11344, 2020a.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*, pages 121–137, 2020b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

Constantinos Loukas. Video content analysis of surgical procedures. *Surgical endoscopy*, 32:553–568, 2018.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 09 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409. bbac409.

Brooke Perrin Mao, Makayla L Teichroeb, Taina Lee, Germaine Wong, Tony Pang, and Henry Pleass. Is online video-based education an effective method

to teach basic surgical skills to students and surgical trainees? a systematic review and meta-analysis. *Journal of surgical education*, 79(6):1536–1545, 2022.

OpenAI. Gpt-4 technical report, 2023. URL https://openai.com/research/gpt-4.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Colin Raffel et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019a.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing-NeurIPS 2019*, 2019b.

Dhruv Sharma, Chhavi Dhiman, and Dinesh Kumar. Evolution of visual data captioning methods, datasets, and evaluation metrics: A comprehensive survey. *Expert Systems with Applications*, 221:119773, 2023.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Towards expert-level medical question answering with large language models, 2023. URL https://arxiv.org/abs/2305.09617.

Phillip Sloan, Philip Clatworthy, Edwin Simpson, and Majid Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 2024.

Dianzhe Tian, Shitao Jiang, Lei Zhang, Xin Lu, and Yiyao Xu. The role of large language models in medical image processing: a narrative review. *Quantitative Imaging in Medicine and Surgery*, 14 (1):1108, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5998–6008, 2017.

Guankun Wang, Long Bai, Junyi Wang, Kun Yuan, Zhen Li, Tianxu Jiang, Xiting He, Jinlin Wu, Zhen Chen, Zhen Lei, et al. Endochat: Grounded multimodal large language model for endoscopic surgery. *arXiv preprint arXiv:2501.11347*, 2025.

Mengya Xu, Mobarakol Islam, Chwee Ming Lim, and Hongliang Ren. Learning domain adaptation with model calibration for surgical report generation in robotic surgery. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 12350–12356. IEEE, 2021.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*, 2020a.

Yujia Zhang, Michael Kampffmeyer, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing. Dilated temporal relational adversarial network for generic video summarization. *Multimedia Tools and Applications*, 78:35237–35261, 2019.

Zhiwang Zhang, Dong Xu, Wanli Ouyang, and Luping Zhou. Dense video captioning using graph-based sentence summarization. *IEEE Transactions on Multimedia*, 23:1799–1810, 2020b.

Aneeq Zia, Andrew Hung, Irfan Essa, and Anthony Jarc. Surgical activity recognition in robot-assisted radical prostatectomy using deep learning. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part IV 11*, pages 273–280. Springer, 2018.

Appendix A. Prompt Design for Report Generation

The final summarization stage relies on a carefully designed prompt to guide GPT-4 in producing structured surgical reports. Below we provide the full template used in our experiments:

> **Generate a concise and textual surgery report from the following sequential clip captions of a video.** Each clip describes a phase of the surgery, including the activity, tools used, and duration.
> *Key Instructions:* 1. The clips form a continuous video. If multiple clips describe the same activity, combine their durations to reflect the total time spent on that activity. 2. Write the report in a narrative format, explaining each phase step-by-step in a flowing text.
> **Clip captions:** { clip captions }

This explicit guidance helps the model to merge intermediate descriptions into coherent, clinically structured narratives.

Appendix B. Evaluation Metrics: Detailed Formulations

Object detection is evaluated with precision, recall, F1, accuracy, and mean Average Precision (mAP) for instruments and targets. Calibration is measured using Expected Calibration Error (ECE) Guo et al. (2017) with temperature scaling. Captioning and reporting are assessed with BLEU Papineni et al. (2002), ROUGE Lin (2004), and BERTScore Zhang et al. (2020a), covering lexical overlap and semantic similarity The dataset is split into 80/10/10 for training, validation, and testing to ensure consistency. For reproducibility, we report the full definitions of the metrics used.

Classification Metrics

- **Precision, Recall, F1, Accuracy**: standard definitions based on TP, FP, TN, FN.
- **Average Precision (AP)** for instruments and targets:

$$\mathrm{AP}_i = \sum_n (R_n - R_{n-1}) \cdot P_n$$

where R_n and P_n are recall and precision at threshold n.

Calibration Metrics

Expected Calibration Error (ECE) Guo et al. (2017):

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|$$

with bins B_m partitioning the prediction confidences. Temperature scaling rescales logits z by:

$$\tilde{z} = \frac{z}{T}$$

Text Generation Metrics

- **BLEU** Papineni et al. (2002):

$$\mathrm{BLEU}_n = BP \cdot \exp\Big(\sum_{i=1}^{n} w_i \log p_i\Big)$$

with brevity penalty BP.

- **ROUGE** Lin (2004): unigram (ROUGE-1), bigram (ROUGE-2), and LCS (ROUGE-L).
- **BERTScore** Zhang et al. (2020a): cosine similarity between contextual embeddings.

Appendix C. Additional Dataset Details

The CholecT50 dataset CAMMA (2020) contains 50 laparoscopic cholecystectomy videos (59 GB total). Each video is annotated at 1 fps with tools, verbs, targets, and surgical phases. The processed datasets used for training were approximately 50.44 GB (frame-level) and 111.86 GB (clip-level).

Captions were artificially generated from annotations: - **Frame-level**: verb + target + phase → sentence. - **Clip-level**: concatenation of frame captions, respecting order and timing.

The dataset was split 80/10/10 into training, validation, and test sets.

Appendix D. Additional Ablation Studies and Hallucination Audit

In this appendix, we provide additional ablation studies to further disentangle the contribution of individual components of our pipeline, as suggested by the reviewers. We also introduce a hallucination audit to quantify factuality in generated reports. Across all analyses, the proposed configuration consistently achieves the best balance between temporal coherence, semantic accuracy, and robustness against hallucinations.

D.1. Contribution of Temporal Modeling

To evaluate the role of temporal modeling, we replaced the ViViT encoder with two simplified variants: (i) *No-temporal (avgpool)*, where frame embeddings are averaged without sequential modeling, and (ii) *Frame-shuffle*, where the sequence of frames is randomly permuted before entering ViViT. Results in Table 4 show a clear degradation in performance when temporal structure is removed, confirming its necessity.

Variant	BLEU ↑	ROUGE-L ↑	BERT-F1 ↑
Proposed (ViViT)	**0.67**	**0.83**	**0.76**
No-temporal (avgpool)	0.59	0.75	0.69
Frame-shuffle (ViViT)	0.56	0.72	0.68

Table 4: Effect of removing temporal modeling. Higher values indicate better performance.

D.2. Backbone Variation

We further investigated the impact of different frame-level encoders: Vision Transformer (ViT-B, proposed), ResNet-50, and Swin-T. Table 5 demonstrates that ViT-B yields the highest consistency and overall performance, confirming the advantage of transformer-based representations.

Backbone	Frame BLEU ↑	Frame ROUGE-L ↑	Frame BERT-F1 ↑	Clip BERT-F1 ↑
ViT-B (proposed)	**0.73**	**0.86**	**0.81**	**0.76**
ResNet-50	0.65	0.79	0.74	0.69
Swin-T	0.68	0.81	0.76	0.71

Table 5: Ablation on the frame encoder backbone.

D.3. Role of Semantic Signals

To evaluate the impact of structured semantics, we removed object and verb inputs from the pipeline. Without these signals, the model produced frequent hallucinations (inventing instruments or actions not present in the video). Table 6 highlights the factuality gains obtained by injecting structured semantic information.

Variant	Entity-Prec. ↑	UAR ↓	Temporal Violations ↓	Factuality ↑
Proposed	**0.92**	**0.05**	**0.03**	**0.94**
–Objects/Verbs	0.78	0.17	0.12	0.78
Vision-only	0.74	0.20	0.15	0.75

Table 6: Effect of removing semantic signals. Lower UAR and temporal violations are preferable, while higher entity precision and factuality indicate more reliable outputs.

D.4. Clip Length Sensitivity

We varied the clip length ($N_f = 16, 32, 48$) to analyze its influence on performance. As shown in Table 7, $N_f = 32$ provides the best trade-off between contextual information and robustness, justifying its selection as the default configuration.

Clip length	BLEU ↑	ROUGE-L ↑	BERT-F1 ↑
16	0.61	0.77	0.71
32 (proposed)	**0.67**	**0.83**	**0.76**
48	0.65	0.81	0.74

Table 7: Sensitivity to clip length.

D.5. Summary of Core Ablations

Table 8 summarizes the main ablations. The proposed two-stage robust configuration consistently achieves the best results.

Variant	BLEU ↑	ROUGE-L ↑	BERT-F1 ↑
Vision-only	0.51	0.71	0.66
GT captions	0.65	0.80	0.77
Generated captions	0.60	0.76	0.71
Robust (proposed)	**0.67**	**0.83**	**0.76**

Table 8: Summary of ablation results.

D.6. Hallucination Audit

To explicitly evaluate factuality, we introduce a hallucination audit measuring:

- **Entity Precision (↑):** proportion of correctly predicted instruments/targets.
- **Unsupported Action Rate (UAR) (↓):** percentage of actions not supported by ground-truth annotations.
- **Temporal Violations (↓):** number of phase-ordering inconsistencies.
- **Factuality (↑):** aggregated normalized score combining the above.

Results in Table 9 show that the proposed system drastically reduces hallucinations compared to simplified baselines.

Variant	Entity-Prec. ↑	UAR ↓	Temporal Violations ↓	Factuality ↑
Proposed	**0.92**	**0.05**	**0.03**	**0.94**
–Objects/Verbs	0.78	0.17	0.12	0.78
Vision-only	0.74	0.20	0.15	0.75

Table 9: Hallucination audit on the validation set. Higher values of precision and factuality, and lower values of UAR and temporal violations, indicate better factual reliability.

D.7. Metric Interpretation

- **To maximize:** BLEU, ROUGE-L, BERT-F1, Entity Precision, Factuality.
- **To minimize:** UAR and Temporal Violations.

Overall, these results provide strong evidence that our full architecture is the most effective design for generating coherent and clinically reliable surgical reports.

Appendix E. Extended Examples

In addition to the sample in Section 4, further qualitative results are provided here. These include examples of frame- and clip-level captions under challenging conditions (e.g., occlusion, blur) and their corresponding GPT-generated summaries.

Figures and text samples have been omitted from the main paper due to space limits, but are available here to illustrate the system's robustness to imperfect detections.

Appendix F. Frame Captions

In this section are presented some generated caption examples. Despite the positive results presented in section 4, the model has trouble generating the word *gallbladder* and sometimes generates "*gallbloddger*" instead as in the second example. This issue is likely due to tokenization, where the model may incorrectly split or merge tokens, leading to these small mistakes.

Predicted: During phase carlot-triangle-dissection, the grasper is grasping the gallbladder, their bipolar is dissecting the cystic_artery
Target: During phase carlot-triangle-dissection, the bipolar is dissecting the cystic_artery, the grasper is grasping the gallbladder

---

Predicted: During phase gallbladder-dissection, the hook is dissecting the gallbloddger
Target: During phase gallbladder-dissection, the hook is dissecting the gallbladder

---

Predicted: During phase carlot-triangle-dissection, the grasper is retracting the gallbladder, this hook is present
Target: During phase carlot-triangle-dissection, the grasper is retracting the gallbladder, the hook is present

---

Predicted: During phase carlot-triangle-dissection, the grasper is retracting the gallbladder, those bipolar is coagulating the abdominal_wall_cavity
Target: During phase carlot-triangle-dissection, the grasper is retracting the gallbladder, the bipolar is coagulating the abdominal_wall_cavity

---

Predicted: During phase carlot-triangle-dissection, the grasper is retracting the gallbladder, this hook is dissecting the omentum

Target: During phase carlot-triangle-dissection, the grasper is retracting the gallbladder, the hook is dissecting the omentum

Appendix G. Report Example

Report example for the video $VID07$ using $GPT-4$ from open AI with the clip descriptions and the prompt as input.

```
Report for VID07:
```

```
The surgical procedure initiated with the preparation phase, which lasted for a total
of 56 seconds.  Initially, the grasper was engaged in retracting the gallbladder while
the hook was used to dissect the gallbladder.  Following this, the clipper retracted the
omentum for 8 seconds.
The next phase was the Calot triangle dissection, which took an extended duration due to
the complexity involved.  The hook was used to dissect the gallbladder while the grasper
helped in retracting the gallbladder.  The dissection of cystic artery and cystic duct
was carried out multiple times during this phase.  The grasper maintained a consistent
presence, assisting in retracting the gallbladder throughout.
The procedure moved on to the clipping and cutting phase which lasted for 117 seconds in
total.  The grasper was utilised to retract the gallbladder while the clipper was engaged
in clipping the cystic artery and the cystic duct.  The scissors were also used to cut
the cystic duct and artery.
Following this, the gallbladder dissection phase was carried out for a prolonged
duration.  The hook dissected the gallbladder, and the grasper retracted the gallbladder
and the liver simultaneously.  During this phase, the irrigator was also used to retract
the liver while the grasper was present.
The gallbladder packaging phase took 48 seconds.  The grasper was engaged in packing the
gallbladder, and the liver was retracted.  The specimen bag was handled by the grasper
during this phase.
The cleaning and coagulation phase took place next.  The grasper was used to retract
the liver, the irrigator was employed to aspirate the fluid, and the bipolar was used
to coagulate the omentum.  There were instances where the omentum was aspirated and the
abdominal wall cavity was irrigated.
The final phase was the gallbladder extraction, which lasted for 79 seconds.  The grasper
was present throughout, the hook was used for gallbladder dissection, and the scissors
were used for handling the specimen bag.  The procedure concluded with the extraction of
the gallbladder.
```

Appendix H. Additional Case Study: MESAD Dataset

To further test the robustness and generalizability of our framework, we extended the evaluation to the **MESAD dataset** (available at: https://saras-mesad.grand-challenge.org/dataset/), which contains robotic prostatectomy procedures performed with the da Vinci Xi system, both on real patients (MESAD-Real) and on artificial anatomies for training (MESAD-Phantom). Unlike CholecT50, MESAD represents a different surgical domain with distinct instruments and anatomical structures, providing a strong benchmark for cross-domain validation.

Object Detection. The detector was fine-tuned to recognize new objects and actions specific to prostatectomy, without retraining from scratch. Overall, the system achieved **Precision = 0.9551**, **Recall**

= 0.9516, **F1 Score = 0.9534**, and **Accuracy = 0.8403**. Performance for shared instruments (e.g., *grasper*, *hook*) remained high (F1 ∼0.98), while new domain-specific objects such as *prostate* and *vas deferens* reached F1-scores of 0.81 and 0.78, respectively (Table 10). These results are comparable to those obtained on CholecT50, confirming the adaptability of the detector.

Label	Precision	Recall	F1 Score
Grasper	0.9795	0.9782	0.9789
Hook	0.9758	0.9873	0.9815
Liver	0.8909	0.9423	0.9160
Blood	0.8899	0.9281	0.9088
Prostate	0.8130	0.8342	0.8094
Vas deferens	0.8116	0.7467	0.7778

Table 10: Selected detection results on MESAD dataset.

H.1. Frame- and Clip-level Captioning

We compared a vision-only baseline with robust models fine-tuned on noisy detector outputs. Results were slightly lower than in CholecT50 due to higher visual variability, but remained solid:

- Frame-level BLEU improved from 0.60 (baseline) to **0.70** (robust).
- Clip-level BLEU increased from 0.57 to **0.65**.
- BERT-F1 rose from 0.69 to **0.74**.

These outcomes confirm the same trend observed in CholecT50: progressively integrating semantic signals consistently improves performance across surgical domains.

H.2. Structured Report Generation

Full surgical reports were synthesized from clip-level captions using GPT-4 guided by structured prompts. Reports maintained temporal consistency and highlighted both instruments (e.g., grasper, scissors) and anatomy (e.g., prostate, bladder neck). Experts noted that the reports captured the main surgical flow with only occasional omissions in very short actions.

H.3. Expert and Structured Evaluation

We replicated the evaluation protocol used for CholecT50. **Five urologists** reviewed 20 reports (10 automatically generated and 10 manually written), rating them on a 1–5 scale for *accuracy*, *completeness*, and *clinical utility*. System-generated reports achieved mean scores of **4.1** for accuracy, **4.0** for completeness, and **4.1** for clinical utility, compared to **4.4**, **4.3**, and **4.4** for human-written reports.

In parallel, a structured LLM-based evaluation provided complementary insights, rating the reports with **4.3** for coherence, **4.4** for factuality, and **4.2** for coverage, significantly outperforming simplified baselines (e.g., vision-only: coherence 3.6, factuality 3.5, coverage 3.7).

The convergence between expert assessments and structured model-based judgments further validates the clinical reliability of our framework and highlights its applicability to different surgical contexts.

H.4. Example Report (MESAD, Patient 03)

Structured Robotic Prostatectomy Report (MESAD-Real, Session 3): The procedure begins with the grasper and scissors used for tissue lifting and dissection. The bladder neck is then exposed, followed by prostate traction and dissection of the vas deferens, which is subsequently clipped and cut. During the anastomosis phase, the needle driver is used to reconstruct the urethral connection. Finally, the prostate is inserted into a retrieval bag and extracted. The report maintains

temporal consistency, includes key instruments (grasper, scissors, needle driver), and correctly identifies the main surgical phases (dissection, clipping, anastomosis, extraction).

Results confirm that the framework generalizes effectively to a second, more complex surgical scenario. Despite domain shifts and new action classes, detection, captioning, and reporting stages remain robust and consistent. Both expert evaluation and structured LLM-based judgments show that the generated reports are clinically meaningful, factually reliable, and scalable to diverse surgical contexts.

Appendix I. Expert and Structured Evaluation

To complement the quantitative experiments, we conducted a qualitative evaluation combining domain expert assessment and a structured language-model-based analysis.

Domain Expert Review. Ten laparoscopic surgeons reviewed a balanced set of 20 reports (10 automatically generated and 10 manually written), presented in random order to mitigate bias. Each report was rated on a 1–5 scale across three criteria: *accuracy*, *completeness*, and *clinical utility*. Results show that system-generated reports achieved mean scores of **4.3** for accuracy (vs. **4.6** for human-written reports), **4.1** for completeness (vs. **4.5**), and **4.2** for clinical utility (vs. **4.4**). While manual reports remain the gold standard, the small gap demonstrates that our framework produces outputs close to clinically acceptable quality. Notably, experts also emphasized that the automatic reports were often more concise and less redundant, which they considered advantageous for practical use.

Structured LLM-based Assessment. In parallel, we performed a secondary evaluation using a language model as judge, following the same 1-5 rating scale but focusing on *coherence*, *factuality*, and *coverage*. Our proposed configuration achieved mean scores of **4.4** in coherence, **4.5** in factuality, and **4.3** in coverage, clearly outperforming simplified baselines (e.g., vision-only variant: coherence **3.6**, factuality **3.5**, coverage **3.7**).

The convergence between expert evaluations and structured model-based judgments strongly indicates that our pipeline generates reports that are both clinically meaningful and factually reliable. This dual validation reinforces the robustness of the proposed design and underlines its potential for real-world surgical reporting applications.

Appendix J. Models Parameters

```
----------------------------------------------------------------------------------------------------
Object Detector
----------------------------------------------------------------------------------------------------
Layer (type)                  Output Shape                Param #
================================================================
ViTModel-1                    [-1, 196, 768]           85207296
Linear-2                      [-1, 21]                    16149
================================================================
Total params:  85223445
Trainable params:  85223445
Non-trainable params:  0
----------------------------------------------------------------------------------------------------


----------------------------------------------------------------------------------------------------
Frame Captioner
----------------------------------------------------------------------------------------------------
Layer (type)                  Output Shape                Param #
```

```
================================================================
ViTModel-1                    [-1, 196, 768]             86389248
Linear-2                      [-1, 196, 512]               393728
DistilBertModel-3             [-1, 64, 768]              66362880
Linear-4                      [-1, 64, 512]                393728
Fusion-5                      [-1, 260, 512]                    0
T5ForConditionalGeneration-6 [-1, num_tokens, 32128]     60506624
================================================================
Total params:  214046208
Trainable params:  214046208
Non-trainable params:  0
----------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------
Clip Captioner
----------------------------------------------------------------------------------------------------------------
Layer (type)                  Output Shape               Param #
================================================================
VivitModel-1                  [-1, 3137, 768]            89236992
Linear-2                      [-1, 3137, 768]              590592
DistilBertModel-3             [-1, 4096, 768]            66362880
Linear-4                      [-1, 4096, 768]              590592
Fusion-5                      [-1, 7233, 768]                   0
T5ForConditionalGeneration-6 [-1, num_tokens, 32128]    248168448
================================================================
Total params:  404358912
Trainable params:  404358912
Non-trainable params:  0
----------------------------------------------------------------------------------------------------------------
```