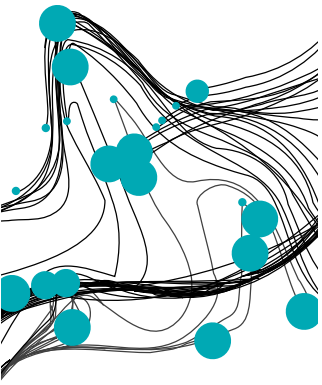
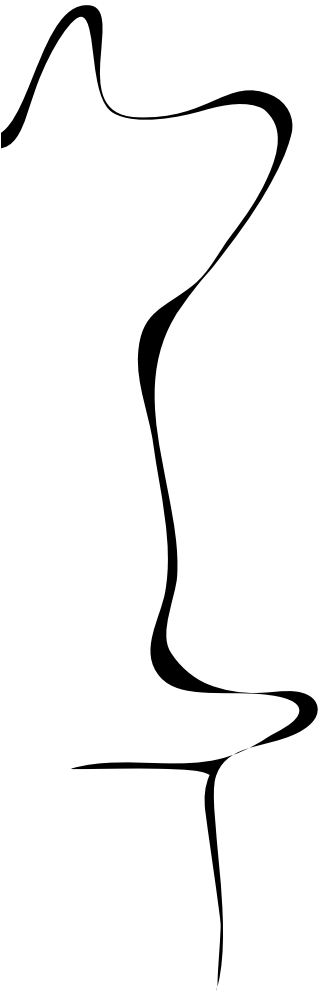


.26389

DMB

DATA MANAGEMENT AND BIOMETRICS



SURGICAL VIDEO TRIPLET RECOGNITION WITH MULTIMODAL LEARNING

Yueming Wu

MASTER'S ASSIGNMENT

Committee:
dr. E. Talavera Martínez MSc
dr. D.V. Le Viet Duc

January, 2025

2025DMB0001
Data Management and Biometrics
EEMathCS
University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

UNIVERSITY OF TWENTE. | DIGITAL SOCIETY INSTITUTE

Abstract

Surgical action triplet recognition is crucial for understanding surgical workflows. This work presents a novel multimodal approach that leverages the complementary strengths of RGB features and segmentation information to improve triplet recognition accuracy. Our key innovation lies in the integration of the Segment Anything Model (SAM) with a CAM-guided prompting mechanism, coupled with a gated cross-attention architecture for effective modality fusion. The system not only achieves improved triplet recognition performance but also demonstrates capability in weakly supervised instrument and anatomy segmentation. Through extensive experimentation on the CholecT45 dataset, we show that our fusion approach with selective information flow outperforms traditional concatenation-based methods. We also provide insights into the limitations of certain modalities, such as optical flow in low frame rate scenarios, and the challenges of using generic vision-language models in medical contexts. Our approach offers practical benefits for surgical workflow analysis while reducing the annotation burden through its dual-use nature.

1. Introduction

The field of Artificial Intelligence (AI) has experienced remarkable evolution in recent years, particularly in deep learning architectures and training strategies. Starting from AlexNet [22], which revolutionized computer vision with its groundbreaking performance on ImageNet [12], deep learning models have grown increasingly sophisticated. The development progressed through deeper architectures like ResNet [19], which solved the vanishing gradient problem through residual connections, to the emergence of Vision Transformers (ViT) [13], which adapted the self-attention mechanism for visual tasks. These architectural advances have been complemented by innovative learning mechanisms, such as attention mechanisms [43], contrastive learning strategies [10], and self-supervised learning approaches [9], enabling models to learn more robust and generalizable feature representations. Parallel to these developments, multi-modal learning has emerged as a powerful paradigm, enabling models to process and integrate information from different data sources. Notable advances include approaches in visual-text models [35], and cross-modal transformers [20]. These developments have significantly enhanced models' ability to learn comprehensive feature representations by leveraging complementary information across different modalities.

This rapid progress in AI has catalyzed transformative advances in the medical field. Beginning with foundational work like U-Net [37] for medical image segmentation, AI applications have expanded to encompass increasingly complex tasks, from predictive medicine [1] and clinical decision-making [6, 34] to sophisticated diagnostics. Among these applications, surgical video analysis has emerged as a particularly promising domain, offering the potential to enhance surgical safety, improve training, and optimize procedural workflows. Early work in this field demonstrated the feasibility of using deep learning for surgical scene understanding, as shown by Funke et al. [14] who applied temporal convolutional networks for real-time detection of surgical tools, and Twinanda et al. [42] who developed frameworks for recognizing surgical phases in laparoscopic cholecystectomy videos. The practical impact of these advances is significant, as highlighted by [5], where the integration of AI in surgical decision support has enhanced clinical outcomes through real-time analysis and feedback during procedures. Beyond intraoperative support, these systems have proven invaluable for surgical education, as noted in [17], where annotated recordings not only facilitate training but also enable objective skill assessment.

The analysis of surgical videos provides comprehensive insights essential for understanding surgical actions. Several datasets have been developed to facilitate research in this area by providing structured and annotated data for various tasks such as skill assessment, workflow recognition, and action detection. For instance, the RMIT [40] dataset captures minimally invasive surgical procedures with annotations aimed at understanding surgical tool usage and hand motions. The JIGSAWS [15] dataset, on the other hand, focuses on fine-grained skill evaluation, offering kinematic and video data from simulated surgical tasks performed. Similarly, the HeiCo [31] provides detailed recordings of colorectal surgeries, emphasizing workflow analysis and decision-making processes. These datasets offer unique opportunities for research in various areas, such as safety analysis, workflow optimization, and skill assessment. For instance, Baghdadi et al. [4] leveraged structured data for performance assessment in pelvic lymph node dissection, and other researchers, such as those in [44], utilized neural networks to assess surgical skills and recognize tasks automatically.

Among these datasets, the CholecT45 [32] dataset stands out as a comprehensive resource specifically designed to break down complex surgical procedures into meaningful components. It introduces a novel framework for representing surgical activities as triplets of {instrument, verb, target}, capturing the interactions between surgical tools, actions, and anatomical targets. For example, a laparoscopic cholecystectomy can be represented with triplets like {Grasper, Hold, Gallbladder} or {Scissors, Cut, Cystic Duct}. These triplets provide a structured approach to understanding surgical workflows.

However, the association of these triplet components—linking the correct instrument, verb, and target in a cohesive manner—still presents challenges. Several underlying difficulties contribute to this issue. For instance, multiple surgical activities often occur within a single frame, making it challenging to distinguish between overlapping actions. Instruments may also be obscured or overlap with one another, complicating the detection process. Moreover, relying solely on information from RGB image is insufficient for determining the specific action being performed, as the spatial relationship alone does not capture the full dynamics of the interaction.

In this study, we aim to address the following research question:

- **Main Research Question:** How can we leverage multimodal information to improve the recognition of surgical action triplets?
- **Sub-Research Questions:**
    1. How can video data be best represented for surgical action triplet recognition?
    2. How can different data modalities be effectively fused to enhance recognition performance?

This study makes several contributions to the field of surgical action understanding:

1. We introduce a novel architecture that effectively combines RGB and segmentation information for surgical action triplet recognition. Our approach leverages the universal segmentation capabilities of SAM, enhanced by CAM-guided prompting, to obtain robust spatial representations that transfer well to the medical domain despite the domain gap.
2. We develop a gated cross-attention mechanism that enables selective fusion of different modalities, maintaining RGB as the primary information source while dynamically incorporating complementary segmentation features. This mechanism has proven crucial for handling the varying reliability of different modalities during surgical procedures.

2. Related Work

In this section, we discuss about the action triplet task and reviewing the current methods employed for its recognition. We then introduce multi-modal learning, with a particular emphasis on various forms of visual modality and the fusion techniques used to integrate them.

2.1. Surgical Triplet Recognition

Surgical action triplets offer a method for understanding the detailed interactions and activities within surgical procedures. By analyzing video data captured from surgical devices, these triplets with the format $\{$instrument, verb, target$\}$ provide a detailed breakdown of the instruments used, the actions performed, and the targets affected, helping to achieve a more refined understanding of the surgical workflow.

Several approaches have been proposed to address the challenge of surgical action triplet detection. The MCIT-IG [39] method introduces a two-stage pipeline that utilizes both image data and Region of Interest(ROI) information for feature extraction, followed by a Graph Neural Network(GNN) for the classification task. Rendezvous [33], on the other hand, proposes an instrument-centric network, addressing the complexity of multiple organs being present simultaneously but only one target being acted upon by an instrument. Since the verb is determined by the instrument's action, the method employs instrument class activation maps to guide the detection of the target and verb components of the triplet, conditioned on the instrument's visual cues. Building on this, RIT [38] enhances the network by focusing on improving verb detection, leveraging both current and past frames to refine the verb component of the triplet. By fusing verb features from multiple frames through a weighted sum, this approach achieves better temporal understanding of actions. Similarly, another method [11] introduces a disentanglement framework that uses class activation maps for guidance. By breaking down the task into smaller steps, it effectively addresses challenges such as the simultaneous appearance of multiple tools or irrelevant surgical activities. It enhance results through soft label generation using self-distillation. Additionally, another recent approach [26] proposes a generative framework for surgical triplet recognition, employing a diffusion model. This model integrates joint space learning and association guidance to improve the accuracy of triplet detection and recognition in surgical videos.

Despite the progress made by these methods, several limitations remain. One common issue is the reliance on a single modality, such as image data or class activation maps, without incorporating additional sources of information like optical flow or depth. By not combining these modalities, these approaches miss out on valuable context and motion cues that could improve the detection and recognition of action triplets. Another challenge lies in the use of class activation maps, which primarily provide instrument location information. While this is useful for detecting the presence and position of surgical tools, it does not capture the association between the verb and the target. This disconnect makes it difficult to fully model the interactions necessary for accurate triplet detection. Temporal information is also crucial for action recognition, as surgical actions unfold over time. Some methods attempt to incorporate temporal data by using sliding windows or weighted features from previous frames. However, the use of sliding windows

requires the tuning of a hyperparameter to control the window size, which can limit the model's ability to generalize across different scenarios. This approach may fail to capture long-range dependencies or adapt to the variability in the duration of surgical actions, leading to suboptimal results.

2.2. Multimodal Learning

Humans naturally perceive and understand the world through multiple sensory modalities. For instance, we comprehend spoken language not only through the sound of speech but also by observing the lip movements of the speaker. Similarly, object recognition involves multiple senses: we recognize objects through vision, the sound they produce, and sometimes their texture when touched. This integration of different sensory inputs significantly enhances the accuracy and reliability of human perception and decision-making.

In multi-modal learning, a similar principle applies, where combining various types of data can improve the performance of machine learning models. There are different types of modalities: some are raw, such as images, audio signals, and videos, which directly capture the sensory input. Other modalities are more abstract and derived from raw data. These include semantic segmentation maps, depth information, optical flow, and even language transcriptions extracted from audio. These higher-level representations provide richer, complementary information that can aid tasks like object detection, action recognition, and scene understanding. The challenge in multi-modal learning lies in not only processing raw sensory data but also integrating abstracted, task-specific modalities to create more robust and intelligent systems.

In the last few years, deep learning has made substantial progress in computer vision, primarily using single-modality data such as images. Early breakthroughs, such as Convolutional Neural Networks (CNNs) [19, 22, 41], revolutionized image classification by introducing architectures capable of learning hierarchical features from images. Later, the introduction of Vision Transformers (ViTs) [13] extended this progress, applying self-attention mechanisms to vision tasks. In parallel, models like ViViT [3] adapted transformer-based architectures to video data, further advancing the field by handling the temporal dimension of video sequences. These methods, however, largely focused on a single modality, limiting the potential to leverage richer data sources. In natural language processing (NLP), similar developments have occurred, with models such as Recurrent Neural Networks (RNNs) and transformers achieving remarkable success in tasks like machine translation, sentiment analysis, and question answering. Yet, these models also predominantly relied on a single modality data.

Recently, the field has seen a growing shift toward multi-modality, where various types of data, such as images, text, audio, and more, are integrated to enhance performance across a wide range of tasks. In computer vision, this typically involves modalities like images, audio, text, depth data, and optical flow. One notable success in this area is CLIP [35], which combines image and text modalities. By learning to associate images with textual descriptions, CLIP [35] has demonstrated the ability to perform zero-shot classification tasks, significantly broadening the capabilities of vision models. The recent rise of multi-modal large language models (MLLMs) has further pushed the boundaries of what multi-modality can achieve. Frameworks like LLaVA [27] represent pioneering efforts in integrating diverse modalities, such as text and vision, into unified models. These advances highlight the growing importance of multi-modality in modern computer vision, where utilizing data from multiple sources is increasingly becoming essential for achieving state-of-the-art performance.

However, despite its power and potential, multi-modal learning also comes with significant challenges. One major challenge is how to effectively represent and extract meaningful features from each modality. Each modality has different characteristics: images are spatial, audio signals are temporal, and text is symbolic. Designing models that can capture the essence of each type of data while keeping the representations consistent is non-trivial. Another challenge is how to fuse different modalities and align them in a meaningful way. It can be difficult to combine image data with text or audio, as these modalities may not naturally correlate in a simple manner. Learning robust cross-modal relationships and ensuring that the fusion of modalities enhances the overall performance, rather than introducing noise or confusion, is an ongoing area of research. Addressing these challenges is crucial for the continued success and advancement of multi-modal learning.

2.3. Visual Modality

In Computer Vision, there are multiple modalities that capture different aspects of visual information. Images, for example, exhibit a complex spatial structure that goes beyond their simple 3D pixel representation in the form of RGB values. While an RGB image provides information about color and intensity, it does not explicitly reveal deeper aspects such as depth or object segmentation. These additional layers of information, like depth maps or segmentation masks, can provide critical insights into the geometric and structural properties of a scene. Similarly, videos add an extra layer of complexity, as they consist of sequential image frames that encode temporal information. Analyzing video requires consideration of the dependencies and relationships between frames over time, making it necessary to capture both spatial and temporal features. This is where modalities like optical flow, which captures the motion be-

tween consecutive frames, become essential. In this section, we will explore different visual modalities—such as image, depth, and optical flow—and discuss their respective feature extraction methods.

2.3.1 Image Modality

Images are typically represented as a grid of pixels. This 3D pixel representation encodes a vast amount of visual information, such as color, brightness, and contrast. However, it doesn't directly convey higher-level properties such as object boundaries, depth, or semantic content. This makes feature extraction techniques critical for understanding the underlying structure of images.

CNNs are the foundational models used to process and analyze image data. They are built upon convolutional layers, which apply small filters over the input image to capture local patterns. These filters slide across the image, detecting simple structures such as edges, textures, and corners in the early layers. As the network deepens, CNNs progressively learn more complex and abstract features by stacking multiple convolutional layers, each one building on the representations learned by the previous layer. This hierarchical feature learning enables CNNs to focus on local spatial hierarchies, starting from low-level features and eventually capturing high-level features such as object parts and entire objects. ResNet [19] introduced the concept of residual connections, allowing networks to be substantially deeper without suffering from vanishing gradient problems by learning residual mappings. EfficientNet [41] improved CNN performance by scaling the network architecture efficiently across depth, width, and resolution with a simple compound scaling method. NFNet [8] builds on EfficientNet [41] but removes the need for batch normalization, enabling faster training with adaptive gradient clipping. ConvNeXt [29] modernized CNNs by adopting architectural innovations from the Transformer family, such as large kernel sizes and LayerNorm, further improving their competitive performance in vision tasks while maintaining the core efficiency of convolutional networks.

Meanwhile, ViTs are based on the Transformer architecture. Unlike CNNs, which focus on local spatial patterns, ViT [13] divide the input image into fixed-size patches and treat each patch as a sequence of tokens. These tokens are then passed through Transformer layers that use self-attention mechanisms to capture relationships between the patches. This approach allows ViT to focus on global feature extraction by modeling long-range dependencies between all patches, rather than being limited to local regions as CNNs are. This makes ViT particularly effective for tasks that require a comprehensive understanding of the entire scene, as they are not biased toward local features like CNNs. Swin Transformer [28] introduces hierarchical feature learning by using non-overlapping windows to partition images and applying self-attention within each window. This hierarchical design helps capture both local and global features efficiently. CLIP [35] leverages both image and text data by training a vision encoder and a text encoder jointly, using contrastive learning to align visual representations with natural language descriptions. DINO [9] is a self-supervised learning method that uses a student-teacher framework to train vision models without labels.

Both CNNs and ViTs are widely used as visual encoders for computer vision tasks. While CNNs focus primarily on local, neighboring dependencies by leveraging small, localized receptive fields, ViTs are designed to capture long-range interactions across the entire image through self-attention mechanisms. As it is stated in [29], one of the key advantages of ViTs is their scaling power. When provided with larger models and datasets, ViTs have the potential to outperform traditional CNN models like ResNets by a substantial margin. This scalability allows ViTs to excel in tasks such as image classification, particularly when dealing with extensive data and powerful computational resources. However, one of the main challenges of ViTs is their global attention mechanism, which has a quadratic complexity in relation to the input size. This makes them efficient for tasks like ImageNet classification, where the input image resolution is relatively small.

Ultimately, the choice between different frameworks depends on the specific requirements of the task. For instance, LLaVA [27] leverages the CLIP [35] encoder as a vision expert to capture rich visual information, benefiting from CLIP's [35] ability to align visual and textual representations. However, models like Flamingo [2] argue that NFNet [8] can outperform CLIP [35] in certain tasks because NFNet [8] is better at capturing fine-grained spatial information. In another approach, SPHINX [25] utilizes mixed embeddings from both CNNs and ViTs for visual encoding, combining the strengths of CNN's local feature extraction with ViT's global attention to create a more robust visual encoder. Moreover, it states that supervised learning models like ConvNeXt [29] and ViT [13] can impose explicit semantic information through category labels. In contrast, self-supervised models like DINO [9] explore implicit signals through pretext tasks, such as learning to match different augmented views of the same image. This forces the model to learn more generalized, context-independent representations of visual data without the need for labeled datasets.

2.3.2 Depth Modality

Depth estimation is a critical aspect of visual modality because it provides a sense of three-dimensional structure from two-dimensional images. While traditional RGB im-

ages offer information about color and intensity, they lack the ability to capture the spatial relationships and distances between objects in a scene. Depth estimation addresses this limitation by predicting the distance of objects from the camera, which is essential for understanding the geometry of a scene.

Depth estimation becomes particularly important for tasks like action recognition, where understanding the motion and spatial relationships of objects or people within a scene is crucial. In these scenarios, depth information helps distinguish between overlapping objects and provides insights into the relative movements of individuals or objects. By knowing the distance and position of key elements in a video, models can more accurately recognize and interpret complex actions and gestures, especially in dynamic environments. Without depth information, many subtle cues in movement could be lost, leading to less accurate predictions.

One prominent model is MiDaS [36]. MiDaS [36] is trained on a diverse set of datasets and uses a ResNet backbone to extract rich feature representations. Later MiDaSv3 [7] extend it to different backbones, such as Swin [28]. It has proven highly effective at providing detailed and accurate depth maps.

2.3.3 Segmentation

Segmentation mirrors how humans perceive and understand the world. When we look at a scene, our brain intuitively segments it into various objects which allows us to comprehend and interact. As an abstract modality, segmentation is particularly important because it provides a clear delineation of objects within an image. Unlike the raw RGB representation, where object boundaries are often implicit and can overlap with other visual elements, segmentation offers explicit information about the shape and location of objects. This is especially useful for aligning different modalities, such as combining image data with depth, optical flow, or text annotations. By providing a localized and detailed understanding of object boundaries, segmentation enhances the ability of multi-modal systems to align and integrate diverse forms of data.

Several notable models have been developed for segmentation. One early technique used for understanding regions of interest within an image is the Class Activation Map (CAM) [47]. CAM [47] highlights the areas in an image that a neural network focuses on when making predictions about its class. By visualizing the regions that contribute most to a specific classification, CAM [47] provide insight into how a model sees objects within an image. R-CNN [16] perform segmentation by first generating region proposals and then using CNNs to classify and refine these regions. FPN [24] enhances the capability by creating feature maps at multiple scales. It builds a pyramid of features from different layers of the network and merges them in a top-down fashion. This multiscale feature representation allows FPN [24] to capture both fine details and high-level information. One of the most advanced and successful segmentation models is the Segment Anything Model (SAM) [21]. The architecture consists of a encoder to generate high-quality embeddings, a prompt encoder, and a mask decoder. It combines this with a dynamic prompt-based system, where prompts in the form of points, boxes, or masks are provided to guide the segmentation process.

2.3.4 Optical Flow

Optical flow captures the velocity and direction of motion by analyzing the changes in pixel intensities over time. Optical flow provides valuable information about how objects move through a scene, making it an essential tool for understanding temporal dynamics in video sequences.

Optical flow is particularly important for tasks such as action recognition in video analysis. Video analysis requires temporal coherence to capture how actions unfold over time. Optical flow enables models to incorporate this temporal information, making it a necessary modality for tasks that rely on understanding motion. In tasks like surgical action recognition, precise hand movements and tool manipulations are key to understanding the surgeon's actions. The shifts in motion can provide crucial information about what is happening at any given moment.

2.4. Multi-Modality Fusion

Multi-modality fusion aims to combine information from different modalities to create a unified representation that captures the cross-modal interactions between individual elements.

Traditionally, there have been two main approaches to modality fusion: early fusion and late fusion. In early fusion, features from different modalities are combined at the input level, before being processed by the model, while in late fusion, the model processes each modality independently, and the results are combined only at the decision-making stage. However, late fusion has a notable drawback, as it tends to treat each modality as an isolated branch, ignoring the relationships and interactions between modalities throughout the learning process. This limitation can lead to suboptimal performance because cross-modal dependencies are not fully exploited.

Moreover, as highlighted in [46], a key challenge in late fusion networks is unimodal bias. This occurs when the network over-depends on one modality, effectively sidelining others during joint training. In such cases, the model may fail to fully utilize the richness of multi-modal data, relying too heavily on the most dominant or easily learnable modal-

ity. In contrast, early fusion networks, by integrating the modalities from the beginning, tend to encourage the model to make use of all input data. This can help mitigate unimodal bias by promoting the extraction of complementary information from each modality, leading to a more balanced and effective use of multi-modal inputs.

For early fusion, there are some simple but effective techniques. One such method is Sequence Append, which involves directly appending the visual tokens from different backbones into a longer sequence. Another common technique is Channel Concatenation, which concatenates visual tokens along the channel dimension without increasing the sequence length. By keeping the sequence length fixed, this method ensures that the fused representation remains computationally efficient, while still combining the information from different modalities. Both of these strategies typically involve some degree of interpolation and flattening of features, as the representations from different modalities need to be aligned and normalized before fusion. Additionally, advanced techniques have been developed. LLaVa-HR [30] injects high-resolution features into low-resolution vision encoders using a mixture of resolution adapter. This technique enables high-resolution data to influence low-resolution vision encoders, enriching the joint representation with finer details while maintaining computational efficiency.

However, these methods may not fully capture the relationships between different modalities. By merely appending or concatenating features, these approaches assume that the combined information is sufficient for the model to learn cross-modal dependencies, but they do not explicitly model interactions between modalities. This can lead to suboptimal performance.

To address these limitations, more sophisticated fusion schemes have been developed. One such approach is Flamingo [2], which adopts a more structured fusion method. Flamingo [2] leverages the Perceiver resampler, which resamples the inputs from different modalities to a reduced dimension, unifying them into a more manageable representation. Once the dimension is reduced, a cross-attention module is employed to explicitly explore the relationships between the different modalities. This cross-attention mechanism allows the model to dynamically attend to relevant information across modalities, the unified representation is fed into the classification network for task-specific predictions. BLIP-2 [23] uses a Q-Former to fuse different modalities. The Q-Former acts as a query-based fusion module, where a set of learnable queries interact with the visual and textual features to extract the most relevant information from both modalities. By fusing the information in this structured manner, BLIP [23] is able to capture more fine-grained relationships between language and vision.

3. Methodology

3.1. Problem Formulation

Given a surgical video frame I_t at time step t, our goal is to predict the surgical action triplet $y_t = <i_t, v_t, g_t>$, where $i_t \in \mathcal{I}$ denotes the instrument class from instrument space. $\mathcal{I}$ with $|\mathcal{I}| = 6$ classes $v_t \in \mathcal{V}$ represents the verb class from verb space. $\mathcal{V}$ with $|\mathcal{V}| = 10$ classes $g_t \in \mathcal{G}$ indicates the target class from target space $\mathcal{G}$ with $|\mathcal{G}| = 15$ classes.

The task can be formulated as learning a mapping function f_θ parameterized by θ:

$$f_\theta : I_t \rightarrow P(y_t)$$

where $P(y_t)$ represents the probability distribution over the combined triplet space $\mathcal{T} = \mathcal{I} \times \mathcal{V} \times \mathcal{G}$ with $|\mathcal{T}| = 100$ valid triplet combinations.

For a video sequence of length T, we denote the input sequence as $\mathbf{I} = \{I_1, ..., I_T\}$ and the corresponding ground truth triplet sequence as $\mathbf{Y} = \{y_1, ..., y_T\}$. Each frame I_t is additionally associated with various modalities. $I_t^{rgb} \in \mathbb{R}^{H \times W \times 3}$ as RGB modality, $I_t^{seg} \in \{0, 1\}^{H \times W}$ as segmentation modality, $I_t^{flow} \in \mathbb{R}^{H \times W \times 2}$ modality with Optical flow. H and W denote the height and width of the input frame respectively.

The model's objective is to maximize the prediction accuracy across the triplet space while effectively leveraging the complementary information from different modalities.

3.2. Multimodal Feature Learning Framework

To address the challenge of surgical triplet recognition, we propose a two-stage network, as illustrated in Figure 1. The first stage focuses on feature extraction, where we explore three different modalities to enhance the representation of surgical activities. These modalities include both raw image-based information and more abstract features, segmentation and optical flow, designed to provide a comprehensive understanding of the surgical scene. The second stage is modality fusion.

3.2.1 Modality-specific Feature Extraction

Our framework extracts complementary features from three distinct modalities: RGB frames, segmentation masks, and optical flow. Each modality stream is designed to capture specific aspects of surgical actions.

RGB. For the RGB, we employ a Swin Transformer [28] backbone, which has demonstrated superior performance in capturing hierarchical visual features. Given an input frame I_t^{rgb}, the RGB encoder f_{rgb} produces features:

$$h_t^{rgb} = f_{rgb}(I_t^{rgb}) \in \mathbb{R}^{N_{rgb} \times D_{rgb}}$$

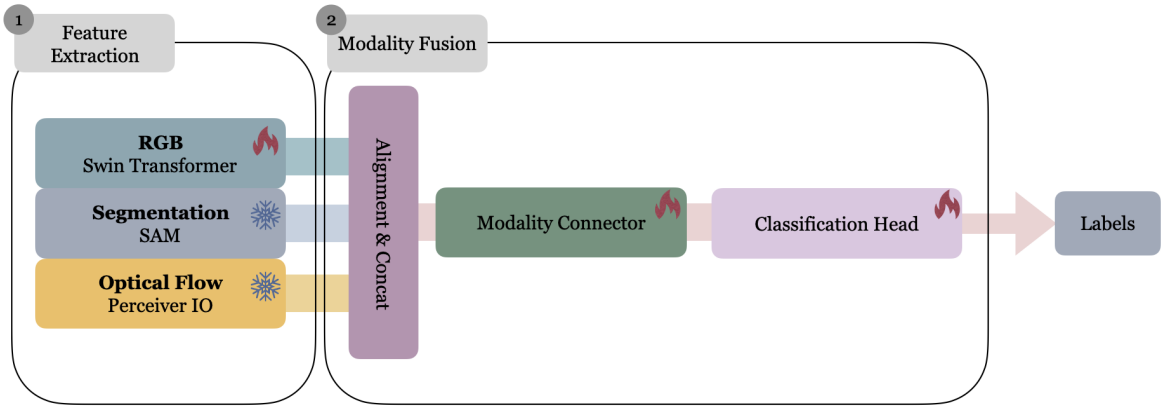


Figure 1. General Framework. The framework consists of two stages: (1) Feature extraction, where features from different modalities are extracted using frozen feature extractors for segmentation and Optical Flow. The RGB encoder is fine-tuned. (2) Fusion, where the extracted features are fused using a modality fusion network. In this stage, the network is unfrozen and fine-tuned to generate the final representation.

where N_{rgb} represents the number of spatial tokens and D_{rgb} is the feature dimension. The Swin Transformer [28] shifted window attention mechanism is particularly effective at capturing both local surgical tool details and global scene context.

Segmentation. We explore three distinct approaches for segmentation feature extraction, each addressing different aspects of the segmentation challenge:

Our initial approach utilizes SAM [21] with random point prompts to generate segmentation masks:

$$M_t^{inst}, M_t^{targ} = f_{sam}(I_t, p_t^{random})$$

where p_t^{random} represents randomly sampled prompt points. While this approach produces high-quality boundary segmentations, it lacks semantic consistency across frames - the same instrument or target may receive different segment labels in consecutive frames, potentially confusing the network.

To maintain semantic consistency while leveraging SAM's [21] powerful representation, we directly utilize SAM's [21] image encoder features:

$$h_t^{seg} = f_{downsample}(f_{sam}^{enc}(I_t)) \in \mathbb{R}^{N_{seg} \times D_{seg}}$$

This approach preserves high-level semantic information while avoiding the semantic inconsistency of mask labels, though it may lose some explicit spatial information.

Our third approach combines class activation mapping (CAM) with SAM [21] to achieve both semantic alignment and precise segmentation. First, we generate class-specific attention maps using ResNet34:

$$A_t^{inst}, A_t^{targ} = f_{cam}(I_t)$$

Convert attention maps to bounding boxes through thresholding:

$$b_t^{inst}, b_t^{targ} = f_{thresh}(A_t^{inst}, A_t^{targ})$$

Use boxes as SAM [21] prompts to generate semantically aligned masks:

$$M_t^{inst}, M_t^{targ} = f_{sam}(I_t, b_t^{inst}, b_t^{targ}) \in \mathbb{R}^{2 \times H \times W}$$

This approach maintains semantic consistency across frames while providing explicit spatial segmentation, as the CAM guidance ensures consistent identification of instruments and targets. The resulting binary masks encode both spatial and semantic information, providing a strong foundation for action recognition.

Optical Flow Stream For motion feature extraction, we adopt the Perceiver IO architecture to process optical flow fields between consecutive frames. Given frame pairs (I_t, I_{t-1}), the flow computation and feature extraction are:

$$F_t = f_{flow}(I_t, I_{t-1})$$

$$h_t^{flow} = f_{perc}(F_t) \in \mathbb{R}^{N_{flow} \times D_{flow}}$$

where F_t represents the computed flow field and h_t^{flow} are the extracted motion features.

Each modality stream outputs features with potentially different spatial dimensions (N_{rgb}, N_{seg}, N_{flow}) and potentially different feature dimensions (D_{rgb}, D_{seg}, D_{flow}). These heterogeneous features are subsequently aligned and fused in the second stage of our framework.

3.2.2 Feature Dimension Alignment

Different encoder architectures produce features with varying sequence lengths and formats. To enable effective fusion, we align these features to a common sequence length N while maintaining their respective feature dimensions. We employ three alignment strategies based on the feature format.

Spatial Feature Alignment For features in spatial format (H×W×C), such as segmentation masks, we employ convolutional layers to adjust the spatial dimensions: $h_{aligned} = f_{conv}(h_{spatial})$ where the convolutional operations are designed to output the desired sequence length N through appropriate stride and kernel size configurations.

Linear Projection. For tokenized features, the simplest alignment approach uses linear projection:

$$h_{aligned} = f_{linear}(h_{tokens})$$

This approach is computationally efficient but maintains a fixed mapping between input and output tokens.

Perceiver Resampler. For tokenized features, we also employ a Perceiver-based resampler that uses learnable latent queries Q to attend to the input sequence:

$$h_{aligned} = \text{MultiHeadAttention}(Q, h_{tokens}, h_{tokens})$$

This self-attention mechanism allows the model to learn dynamic, content-aware feature resampling, particularly effective for variable-length sequences from different modalities.

Each modality's features maintain their original feature dimensions (D_{rgb}, D_{seg}, D_{flow}) after alignment, as these dimensional differences are handled in the subsequent fusion stage. The critical goal is achieving consistent sequence length N across modalities to enable effective cross-modal attention mechanisms.

3.3. Modality Fusion Strategies

After aligning features across modalities, we explore different architectures to effectively fuse RGB features with complementary information from other modalities. Given that RGB features inherently contain the most comprehensive information—capturing texture, color, and fine-grained spatial details crucial for action recognition—we use RGB as the main branch. Auxiliary modalities, such as segmentation and optical flow, serve as complementary information providers to enhance RGB features rather than as primary sources.

Initially, we concatenate features from different modalities along the channel dimension. This unified representation serves as the input to our fusion strategies, which focus on selective information combination through attention mechanisms. Below, we describe our two proposed fusion methods:

Gated Attention Fusion

The first approach utilizes a gated attention mechanism that processes mask features to generate attention weights for RGB features:

$$h_{fused} = h_{rgb} \odot \sigma(W(h_{rgb} \odot h_{aligned}))$$

where σ denotes the sigmoid function, and W is a learned gating mechanism that provides additional feature selection. The element-wise multiplication ($\odot$) allows the model to selectively focus on relevant spatial regions while maintaining feature coherence.

Gated Cross-Attention Fusion Our second approach employs a more sophisticated gated cross-attention mechanism:

$$Q = h_{rgb}, K = V = h_{aligned}$$

$$A = \text{MultiHead}(Q, K, V)$$

$$g = \sigma(\gamma)$$

$$h_{fused} = h_{rgb} + g \cdot A$$

where γ is a learnable parameter, and the cross-attention allows for more complex interactions between modalities. We use RGB features as queries, allowing them to actively seek relevant complementary information from auxiliary modalities. As RGB features encapsulate the most detailed information about the surgical scene, making them the primary source for guiding attention. The gate g scales the attended features before adding them to the RGB features. The residual connection preserves the integrity of the RGB features, while the gating mechanism controls how much auxiliary information is integrated.

The final fused features h_{final} are then passed to a classification head for triplet prediction:

$$y_t = f_{cls}(h_{fused}) \in \mathbb{R}^{|\mathcal{T}|}$$

where $|\mathcal{T}|$ represents the number of valid triplet combinations.

To address the significant class imbalance in surgical triplet recognition, we employ focal loss as our training objective:

$$\mathcal{L}_{focal} = -\alpha(1 - p_t)^{\gamma} \log(p_t)$$

where p_t is the model's estimated probability for the target class, α is a balancing factor, γ is the focusing parameter that adjusts the rate at which easy examples are down-weighted. This approach is particularly effective for our problem as it allows the model to focus on learning the rare but valid surgical action triplets without being overwhelmed by the dominant combinations that naturally occur more frequently in surgical procedures.

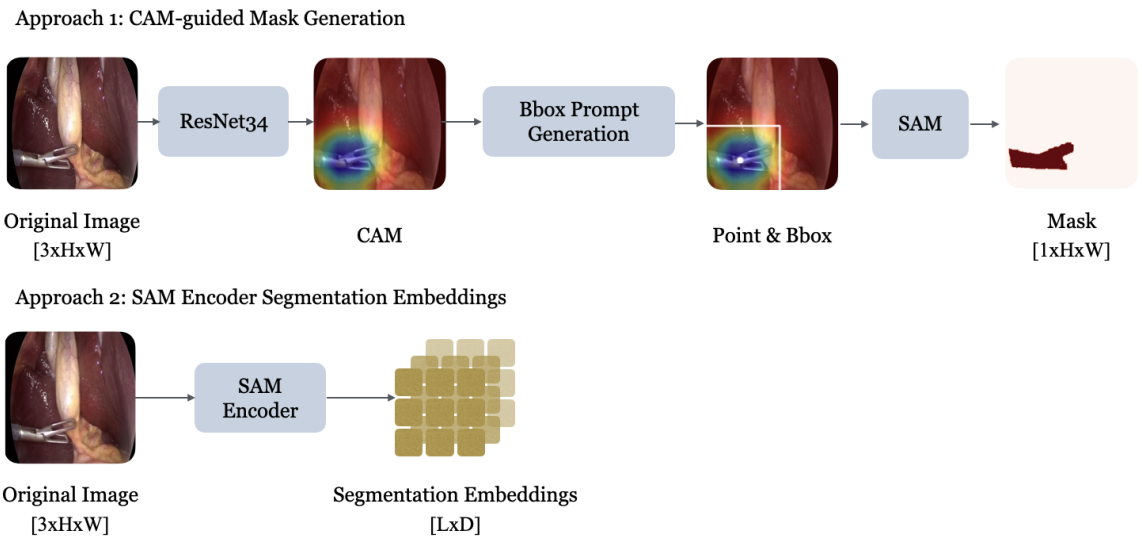


Figure 2. Two alternative approaches for segmentation feature extraction, where H and W denote the height and width of the input image respectively. Top: CAM-guided SAM [21] pipeline, where ResNet-34 processes the input image ([3×H×W]) to generate class activation maps (CAM), which guide bounding box and point prompt generation for SAM [21] to produce semantically-aligned masks ([1×H×W]). Bottom: Direct feature extraction using SAM [21] encoder, transforming the input image ([3×H×W]) into segmentation embeddings ([L×D]), where L is the sequence length determined by the number of image patches and D is the feature dimension of each embedding vector, providing high-level semantic representations without explicit mask generation.

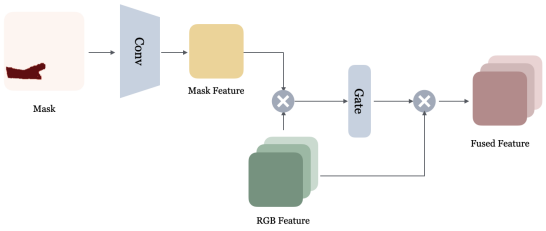


Figure 3. Detailed structure of gated attention fusion module for mask embedding. We apply convolution block to the generated mask, and apply it to the original RGB feature. We use a sigmoid gate to control the contribution. Lastly, we multiply it again to get the combined feature.

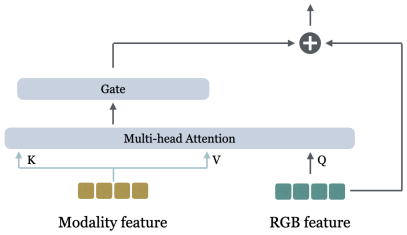


Figure 4. Detailed structure of the gated cross-attention fusion module. RGB features serve as queries (Q) while modality features serve as keys (K) and values (V) in the multi-head attention mechanism. The attention output is regulated by a learnable gating mechanism before being combined with the original RGB features through a residual connection.

3.4. Temporal Smoothing

To mitigate frame-to-frame prediction noise and enhance temporal consistency in surgical triplet recognition, we incorporate a smoothing strategy that leverages a sliding window of recent frames. Let $p_t \in \mathbb{R}^{|\mathcal{T}|}$ denote the prediction probability vector for the current frame t, and let N be the size of the sliding window. We first compute the mean prediction over the sliding window:

$$\bar{p}_t = \frac{1}{|W_t|} \sum_{i \in W_t} p_i, \tag{1}$$

where W_t is the set of indices corresponding to the frames in the current window (with $|W_t| \leq N$ when fewer than N frames are available).

We then compute the smoothed prediction $\hat{p}_t$ as a combination of the current frame's prediction and the windowed average:

$$\hat{p}_t = \alpha\, p_t + (1 - \alpha)\, \bar{p}_t, \tag{2}$$

with $\alpha \in [0, 1]$ acting as a smoothing factor that controls the trade-off between the current prediction and the historical average. This approach leverages recent information to reduce transient fluctuations while still adapting quickly to

new observations.

4. Experiments and Results

4.1. Implementation Details

We conducted all experiments using NVIDIA A10 and A16 GPUs, with our framework implemented in PyTorch. For training, we employed the Adam optimizer with an initial learning rate of 2e-4 and weight decay of 1e-6. A cosine annealing scheduler with warm restarts was adopted, where T_0 was set to epochs + 1 and T_mult to 1, with the learning rate decreasing to a minimum of 2e-5. The model was trained for 2 epochs with a batch size of 16 and gradient accumulation steps of 2 to simulate larger batch sizes while managing memory constraints. To address the inherent class imbalance in surgical action triplet detection, we implemented Focal Loss with $\alpha = 0.8$ and $\gamma = 2$. Our data augmentation strategy included vertical and horizontal flips, contrast adjustments, and 90-degree rotations. For optimization stability, we apply gradient clipping with a maximum norm of 1.0 and employ layer-wise learning rate decay for the transformer layers. The learning rate for the attention modules is set to 1.5 times.

4.2. Dataset Analysis and Insights

4.2.1 Dataset Overview

The CholecT45 [32] dataset comprises 45 videos of cholecystectomy procedures, with frames extracted at 1 frame per second (fps). The dataset contains 90,489 annotated frames with 127,385 triplet instances, where each triplet follows the structure of $\{$instrument, verb, target$\}$. The annotation space consists of 100 unique triplet classes, derived from 6 instrument classes, 10 verb classes, and 15 target classes.

4.2.2 Class Distribution Analysis

Our analysis reveals distinct distributional patterns across individual components and their combinations, as it is shown in Fig 6. At the component level, while certain classes dominate (e.g., graspers and hooks among instruments, retract and dissect among verbs), the distribution reflects the fundamental requirements of cholecystectomy procedures.

However, as it is shown in Fig 5, the triplet distribution exhibits more pronounced imbalance, following a natural long-tailed pattern that reflects the inherent structure of surgical procedures. The most frequent combination is $\langle$grasper, retract, gallbladder$\rangle$, while rare combinations like $\langle$bipolar, grasp, cystic-plate$\rangle$ appear fewer than 10 times. This imbalance is not merely a dataset artifact but rather represents genuine surgical workflow constraints – certain instruments are designed for specific actions on particular targets, while some combinations rarely occur in standard procedures. This natural but extreme imbalance poses a unique challenge for model supervision.

4.2.3 Temporal Characteristics

Analysis of action duration distributions reveals complex temporal dynamics. Table 1 presents statistics for the primary surgical actions. Retraction actions exhibit the highest variance. Dissection actions show similarly broad distribution, while actions like irrigation maintain consistently shorter durations. We observe that action durations follow heavy-tailed distributions, with significant positive skew indicated by standard deviation. This temporal heterogeneity poses challenges for models attempting to capture action dynamics.

4.2.4 Sequential Dependencies

Analysis of inter-triplet transitions reveals highly structured temporal patterns in surgical workflows. As it is shown in Table 2, we quantify these dependencies by computing transition probabilities between consecutive frames. Certain surgical steps show deterministic transitions, particularly in critical phases. For example, after scissors-cut-adhesion operations, the workflow invariably transitions to grasper-retract-omentum, indicating a standardized procedural sequence. The $\langle$grasper, retract, gallbladder$\rangle$ action serves as a central "hub" state, being the dominant subsequent action.This reflects its role as a stabilizing action between other surgical steps.

In summary, these analyses highlight the technical challenge that the dataset triplet class is extremely imbalance. Moreover, its multi-scale temporal dynamics, requiring models to capture both brief and extended action sequences. The identified characteristics inform our architectural decisions, particularly in addressing class imbalance through targeted data sampling strategies.

4.3. Evaluation Metrics

We adopt mean Average Precision (mAP) as our primary evaluation metric, computing it at four different levels to comprehensively assess model performance. For instrument detection (mAP_I), for verb (mAP_V), and target (mAP_T) detection, we evaluate the model's ability to correctly identify surgical actions and anatomical targets respectively. The triplet detection metric (mAP_{IVT}) provides the most important evaluation, requiring all three components (instrument, verb, target) to match the ground truth simultaneously.

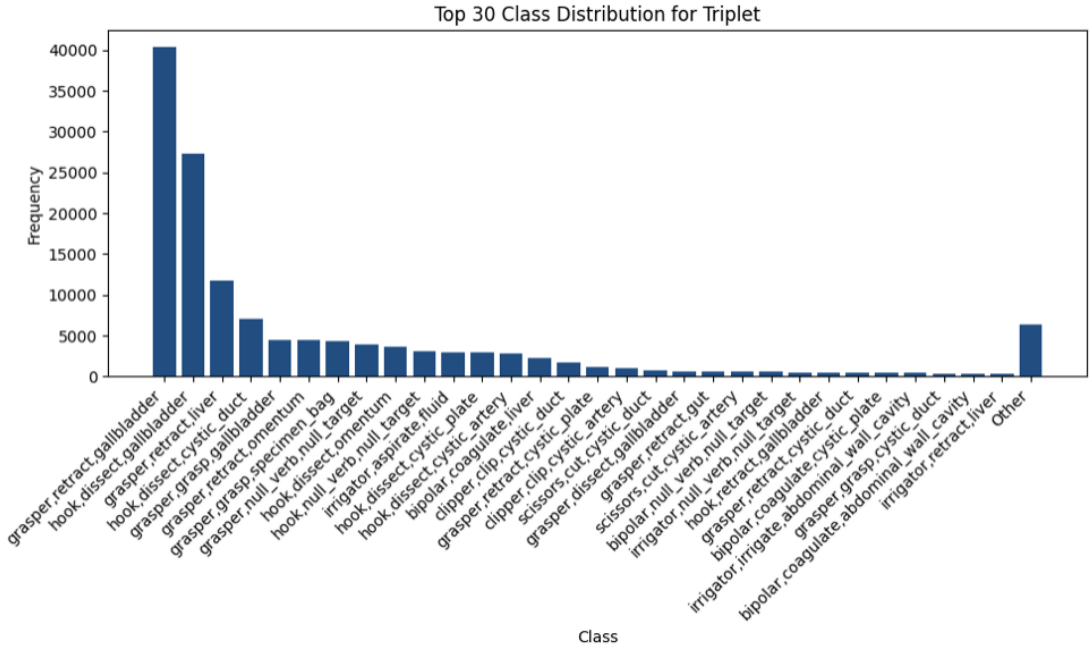


Figure 5. Distribution of surgical action triplets in CholecT45 [32] dataset. Top 30 most frequent triplet combinations are shown individually, with remaining 70 combinations aggregated as "others" due to their low occurrence.

Actions	Mean duration(s)	Median duration(s)	Min duration(s)	Max duration(s)	Standard deviation(s)
retract	39.99	12.00	1.00	1073.00	84.04
null	8.94	6.00	1.00	231.00	13.83
dissect	55.62	38.00	1.00	756.00	61.33
grasp	21.32	10.00	1.00	256.00	30.06
aspirate	10.40	6.00	1.00	108.00	11.93
coagulate	18.37	12.00	1.00	98.00	16.85
irrigate	3.62	3.00	1.00	14.00	2.58
clip	20.63	17.00	2.00	110.00	14.63
cut	21.86	17.00	3.00	151.00	19.64
pack	12.25	10.50	4.00	24.00	6.35

Table 1. Temporal characteristics of surgical actions in CholecT45 [32] dataset.

Source Action	Target Action	$n_{dominant}$	n_{total}	r_d
⟨scissors, cut, adhesion⟩	⟨grasper, retract, omentum⟩	96	96	1.000
⟨bipolar, coagulate, cystic_pedicle⟩	⟨grasper, retract, gallbladder⟩	35	35	1.000
⟨scissors, cut, cystic_plate⟩	⟨grasper, retract, gallbladder⟩	17	17	1.000
⟨scissors, dissect, cystic_plate⟩	⟨grasper, retract, gallbladder⟩	12	12	1.000
⟨bipolar, dissect, cystic_artery⟩	⟨grasper, retract, gallbladder⟩	89	90	0.989
⟨grasper, dissect, gallbladder⟩	⟨grasper, retract, gallbladder⟩	418	423	0.988
⟨hook, dissect, cystic_artery⟩	⟨grasper, retract, gallbladder⟩	2069	2272	0.911
⟨hook, dissect, cystic_plate⟩	⟨grasper, retract, gallbladder⟩	1818	2029	0.896

Table 2. Dominant surgical action transitions with high determinism ($r_d > 0.85$). $n_{dominant}$ represents the count of the most frequent transition, n_{total} is the total transitions from the source action, and r_d is the dominance ratio.

4.4. Results

4.4.1 Ablation Study

Table 3 presents our comprehensive ablation studies on different modalities, dimension alignment methods, and fusion strategies. Several notable and sometimes counterintuitive findings emerge from these experiments.

First, contrary to our initial hypothesis, the RGB-only baseline demonstrates superior performance in instrument with 90.0% and verb with 65.7% detection compared to all multimodal variants. This suggests that the base model effectively captures spatial features crucial for instrument

Modalities	Align	Fusion	mAP_I	mAP_V	mAP_T	mAP_{IVT}
OF	-	-	71.2	32.4	15.8	9.2
SEG	-	-	80.3	45.6	34.5	23.0
RGB	-	-	**90.0**	**65.7**	47.9	32.6
RGB+SEG$_f$	LP	GCA	84.0	60.0	45.8	30.0
RGB+SEG$_f$	PR	GCA	85.7	62.8	**51.7**	**35.7**
RGB+SEG$_f$	PR	LP	81.0	58.4	49.7	34.6
RGB+SEG$_m$	CF	GA	84.2	61.2	50.1	34.2
RGB+SEG$_m$	CF	LP	80.8	52.4	45.3	30.9
RGB+OF	LP	GCA	84.6	59.4	33.5	28.5
RGB+OF	PR	GCA	79.8	54.1	33.2	26.0
RGB+OF	LP	LP	83.7	57.8	32.6	26.2
RGB+SEG$_f$+OF	LP	GCA	81.4	58.8	33.5	27.7
RGB+SEG$_f$+OF	LP	LP	79.4	54.3	32.9	25.8

Table 3. Ablation studies on modalities and fusion methods. Modalities: RGB images (RGB), segmentation features (SEG$_f$), segmentation masks (SEG$_m$), optical flow (OF). Dimension alignment: Linear projection (LP), Perceiver resampler (PR), Convolutional features (CF). Fusion: Gated attention (GA), Gated cross-attention (GCA).

recognition and action understanding from RGB inputs alone. For models using only the Optical Flow and only the Segmentation modalities, they achieve relatively low performance which highlights that it lacks the semantic richness required for effective instrument and action recognition. These observations suggest that neither modality in isolation is sufficient for robust triplet detection, underscoring the necessity of multimodal fusion to combine complementary cues for improved performance.

The integration of segmentation features (SEG$_f$) with RGB shows mixed results. While it leads to a decrease in instrument and verb detection, it significantly improves target detection and overall triplet recognition ($mAP_T = 51.7\%$, $mAP_{IVT} = 35.7\%$). This improvement in triplet detection suggests that segmentation features provide valuable contextual information for understanding instrument-target interactions, despite the slight compromise in individual component detection.

We also found that the integration of segmentation features with RGB shows superior result compared to CAM guided SAM [21] mask generation. The reason might be the target mask is less accurate. The generate bounding boxes usually cant cover all areas, lead to some noise to the mask generation. Such imperfect mask might introduce noise. However, the triplet recognition outscores the RGB baseline, indicating it indeed provide useful location information.

Surprisingly, optical flow (OF) features consistently degrade performance across all metrics when compared to both the baseline and SEG variants. The RGB+OF combination with LP achieves only $mAP_{IVT} = 28.5\%$, substantially lower than the baseline's 32.6%. This unexpected degradation suggests that motion features, as captured by our optical flow implementation, may introduce noise rather than beneficial temporal information for surgical action recognition.

Regarding dimension alignment methods, Linear Projection (LP) generally outperforms Perceiver Resampler (PR) when comparing similar modality combinations. For instance, in the RGB+OF configuration, LP achieves better results across all metrics ($mAP_I = 84.6\%$ vs 79.8%, $mAP_V = 59.4\%$ vs 54.1%). This indicates that the additional complexity of PR might not be justified for our specific task.

The combination of all modalities (RGB+SEG$_f$+OF) does not yield better results than simpler configurations, achieving only $mAP_{IVT} = 27.7\%$. This suggests that the potential benefits of multimodal fusion might be offset by the challenges of effectively combining features from disparate sources, particularly when incorporating optical flow information.

4.4.2 Comparison with State-of-the-Art

Table 4 presents a comparison of our approach with existing state-of-the-art methods for surgical action triplet detection. While our model achieves competitive performance ($mAP_{IVT} = 35.7\%$), it falls short of the current best result from TERL ($mAP_{IVT} = 39.0\%$). This performance gap reveals several important insights about the task.

Moreover, the introduction of temporal smoothing offers additional benefits. When temporal smoothing is applied, our method exhibits a noticeable improvement in temporal consistency, with video-level triplet mAP increasing from 51.7% to 54.0% and an overall mAP_{IVT} rise from 35.7% to 36.4%. These gains indicate that smoothing over a sliding window of recent frames helps reduce transient fluctuations in the predictions, thereby better capturing the evolving dy-

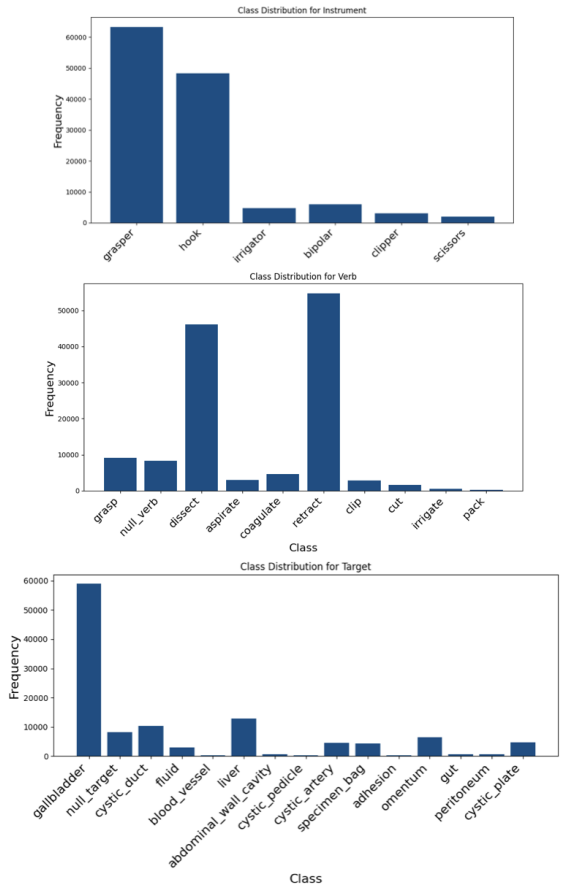


Figure 6. Component-wise distribution analysis of surgical actions in CholecT45 [32] dataset. Left: Distribution of 6 instrument classes. Middle: Distribution of 10 verb classes. Right: Distribution of 15 target classes.

namics inherent in surgical procedures.

A notable trend emerges: the majority of successful approaches leverage Class Activation Maps as a form of semantic guidance, highlighting the crucial role of spatial-semantic information in surgical action understanding. Our approach explores an alternative path using explicit segmentation features, achieving competitive performance ($mAP_{IVT} = 35.7\%$) while validating the importance of semantic guidance.

The superior performance of TERL ($mAP_{IVT} = 39.0\%$) can be attributed to its tail-enhanced representation learning strategy, which specifically addresses the long-tailed distribution nature of surgical action triplets. TERL leverages contrastive learning with a global memory bank to capture discriminative features specifically for tail classes, while also maintaining semantic relationships through component class prototypes. Our approach, while effective at modeling cross-modal relationships, does not explicitly handle class imbalance, which appears to be crucial for this

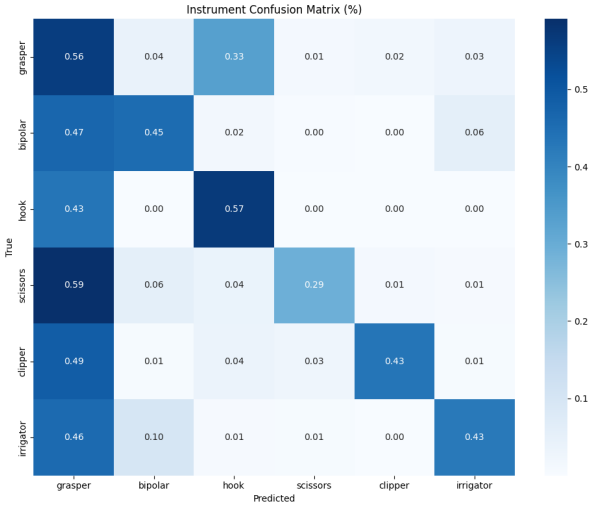


Figure 7. Confusion Matrix for Instrument Classification.

Figure 8. Confusion Matrix for Verb Classification.

task. This is evident in the performance gap in verb detection ($mAP_V = 71.3\%$ vs 62.8%), where rare actions are more prevalent.

Building upon the observation of class imbalance effects, the confusion matrices, as it is shown in Figure7 provide deeper insights into the model's behavior. In the instrument detection case, while grasper achieves 56% accuracy, it significantly contributes to false positives across other classes, with 47% of bipolar forceps and 43% of hooks being misclassified as graspers. This systematic misclassification pattern stems from the grasper's dominance in the training data, causing the model to develop a strong prior bias towards grasper features. The target detection matrix further reinforces this phenomenon, where dominant classes like gallbladder (82.3% accuracy) and cystic plate

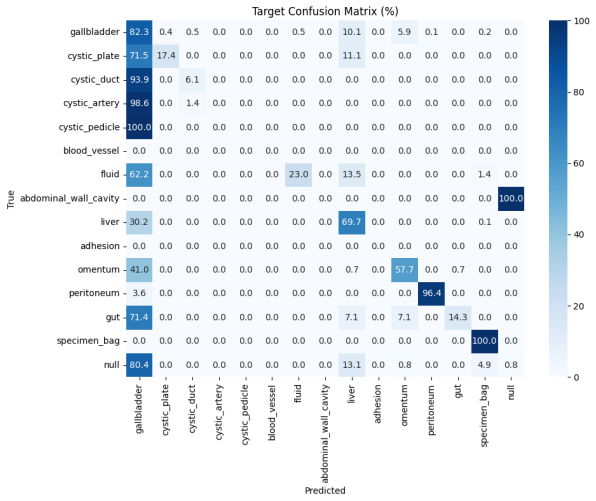


Figure 9. Confusion Matrix for Target Classification.

(71.5% accuracy) show high true positive rates but also contribute to false positives in related anatomical structures. For instance, liver samples are misclassified as gallbladder in 30.2% of cases, suggesting the model struggles to distinguish fine-grained features between anatomically adjacent structures when trained on imbalanced data. The verb detection matrix reveals similar challenges, with 'retract' actions being over-predicted across multiple true classes. This manifests in high false positive rates, where 41-49% of 'dissect', 'coagulate', and 'clip' actions are misclassified as 'retract'. Interestingly, despite these significant misclassifications, the model maintains relatively high average precision scores across classes. This apparent contradiction can be attributed to the model's ability to assign higher confidence scores to true positive cases while struggling with decision boundaries between classes due to imbalanced training examples. The high AP scores suggest that the model learns meaningful feature representations but fails to establish appropriate decision thresholds for balanced classification performance.

SDSwin's relatively good performance ($mAP_{IVT} = 36.1\%$) with only RGB input and knowledge distillation suggests that our multimodal approach might benefit from similar self-distillation techniques to refine the feature representations.

Notably, our model shows competitive performance in target detection ($mAP_T = 51.7\%$) compared to TERL ($mAP_T = 54.0\%$), suggesting that our segmentation-based approach effectively captures spatial relationships between instruments and anatomical structures. However, the lower instrument detection performance ($mAP_I = 85.7\%$ vs 91.5%) indicates that our current fusion strategy might be suboptimal for preserving fine-grained instrument features.

5. Discussion

5.1. Segmentation Superiority Analysis

Our experiments revealed that the integration of segmentation information provides crucial advantages for surgical action triplet detection. The superiority of segmentation-based features can be attributed to two key components: the SAM [21] encoder features and the approximate semantic masks.

The SAM [21] encoder features provide rich visual representations that are particularly well-suited for surgical environments. The SAM [21] encoder, through its vision transformer architecture, captures both fine-grained details and global context. This hierarchical understanding is essential for surgical scenes where instruments interact with anatomical structures at various scales. Even without specific prompts, the SAM [21] encoder generates features that are inherently attuned to object boundaries and salient regions, making them particularly effective at capturing instrument-tissue interactions.

The approximate semantic masks, generated through our two-stage pipeline combining SAM [21] with CAM-guided bbox generation, provide complementary benefits. The masks effectively highlight regions of interest, helping the model focus on relevant instrument-tissue interactions while suppressing background noise.

5.2. Effective Fusion Strategies

The gated cross-attention mechanism emerged as the most effective approach for fusing RGB and segmentation features. The cross-attention mechanism enables each RGB feature to dynamically attend to relevant mask features, creating a flexible and context-aware fusion process.

The introduction of the learnable gating parameter proved particularly important. It controls the influence of modality features on RGB features, preventing feature corruption when mask predictions are unreliable. Therefore, it adaptively balancing the contribution of each modality.

This analysis suggests that the success of our approach stems from the synergistic combination of rich visual features from the SAM [21] encoder, structured spatial information from semantic masks, and an adaptive fusion mechanism that can optimally leverage both sources of information.

5.3. Key Architectural Decisions

Our architectural choices were heavily influenced by recent advances in vision models and multimodal fusion strategies, though with several important adaptations for the surgical domain.

The conventional wisdom of using frozen pre-trained encoders such as Swin Transformer, DINOv2, ConvNext

Model	Year	Backbone	Modality	mAP_I	mAP_V	mAP_T	mAP_{IVT}
Tripnet [32]	2020	ResNet-18	RGB & CAM	89.9	59.9	37.4	24.4
Attention Triplet [33]	2022	ResNet-18	RGB & CAM	89.1	61.2	40.3	27.2
Rendezvous [38]	2023	ResNet-18	RGB & CAM	89.3	62.0	40.0	29.4
SDSwin [45]	2023	Swin	RGB	-	-	-	36.1
TERL [18]	2024	Swin	RGB & CAM	91.5	71.3	54.0	**39.0**
ours (w/o TS)	2025	Swin	RGB & SEG	85.7	62.8	51.7	35.7
ours (w/ TS)	2025	Swin	RGB & SEG	85.6	63.9	54.0	36.4

Table 4. Performance comparison of models from recent years. "w/o TS" and "w/ TS" denote our method without and with temporal smoothing (TS), respectively.

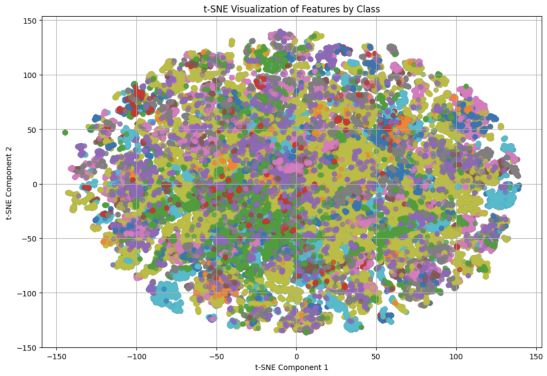


Figure 10. T-SNE feature extracted based on SAM. It shows a relatively uniform distribution without clear clustering.

proved challenging in our surgical context. While this approach has been successful in many recent works by only training the fusion components, we found the domain gap between general vision datasets and surgical scenes to be too significant. This necessitated careful fine-tuning of these backbones to adapt to the medical domain's unique characteristics.

Modern vision architectures often face an inherent trade-off between capturing semantic information and preserving local structural details. Our analysis reveals an interesting dynamic between the RGB backbone and SAM features that effectively balances this trade-off. As it is shown in Figure 10, SAM extracts generalizable features rather than overly specialized ones. The overlap between different colors suggests SAM captures shared visual patterns across different surgical actions. The Swin Transformer backbone, operating on RGB inputs, demonstrates superior capability in extracting semantic information crucial for action recognition. This is evidenced by its strong performance in distinguishing different surgical actions when used alone. However, like many semantic-focused architectures, it may not fully capture fine-grained spatial relationships and boundary information.

An initial concern was that SAM's class-agnostic nature might introduce confusion - the same instrument could have different representations across frames. However, this potential inconsistency is effectively managed through two mechanisms:

- Hierarchical Information Processing: The RGB backbone maintains primary control over semantic understanding, while SAM features serve as refinement signals rather than primary classification cues.
- Attention-Based Feature Selection: Our gated cross-attention mechanism allows the RGB features to selectively query relevant structural information from SAM features. This selective attention means that even if SAM represents the same instrument differently across frames, the RGB features can attend to the relevant aspects of these representations based on the current context.

For example, when the RGB stream identifies a grasping action, it can attend to SAM's boundary information to refine its understanding of the instrument-tissue interaction, regardless of SAM's specific representation of the instrument. This dynamic ensures that SAM's variable representations enhance rather than confuse the model's predictions.

Some sophisticated modules that have shown promise in other domains, such as the Perceiver resampler, underperformed in our context. We hypothesize this is due to it increased model complexity. The surgical domain's inherent structure possibly being better served by simpler, more direct architectural choices.

5.4. Practical Applicability

Our system demonstrates significant practical value beyond its primary function of triplet detection through its ability to perform weakly supervised segmentation. As shown in Figure 11, the integration of SAM with CAM-guided prompting enables the generation of meaningful segmentation masks using only triplet-level supervision, particularly excelling in instrument segmentation. This success can be attributed to the visual distinctiveness of surgical instruments and the strong correlation between instru-

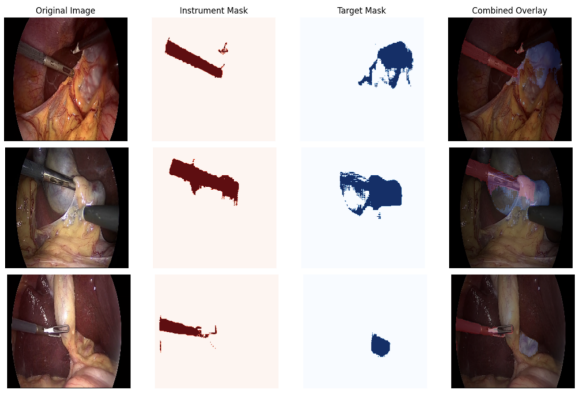


Figure 11. Overview of the generated mask.

ment locations and action labels, which leads to more precise CAM predictions. From a practical implementation standpoint, this dual-use capability offers significant advantages by providing both action recognition and spatial information without requiring expensive pixel-level annotations, making it particularly valuable for surgical workflow analysis and training applications.

5.5. Answers to Research Questions

Our investigation into multimodal surgical action triplet recognition has yielded answers to our initial research questions.

Regarding our first question: **How can video data be best represented for surgical action triplet recognition?** Our research revealed that a combination of RGB features and segmentation information provides the most effective representation for surgical action triplet recognition. The Segment Anything Model (SAM) [21] proved particularly valuable, as its universal segmentation capabilities transfer well to the medical domain. While we explored various modalities including optical flow, the temporal sparsity of surgical videos (1 FPS) limited the utility of motion-based features.

For our second question: **How can different data modalities be effectively fused to enhance recognition performance?**, we discovered that fusion strategies with RGB as the primary modality consistently outperform naive fusion approaches. The gated cross-attention mechanism emerged as the optimal solution, allowing the model to dynamically weight the contribution of different modalities. This approach proved superior to alternatives such as simple concatenation or self-attention across all modalities. The gating mechanism played a crucial role, enabling the model to selectively incorporate segmentation information while maintaining the primacy of RGB features. Our experiments demonstrated that removing this gating component led to degraded performance, highlighting its importance in effective multimodal fusion.

Answering our main research question: **How can we leverage multimodal information to improve the recognition of surgical action triplets?**, we found that successful multimodal integration requires both careful selection of complementary modalities and appropriate fusion architecture design. Through extensive experimentation, we identified that combining RGB features with SAM-generated segmentation masks provides the most effective representation for our task, while optical flow proved less beneficial due to temporal constraints. Our work demonstrates the value of leveraging pre-trained foundation models, as the frozen SAM model effectively transfers its universal segmentation capabilities to the surgical domain without requiring domain-specific fine-tuning. This finding suggests a promising direction for leveraging powerful modern models to enhance domain-specific tasks. For modality fusion, we found that a gated cross-attention mechanism achieves optimal performance by allowing dynamic weighting of modalities while maintaining RGB as the primary information source. This architecture enables the model to selectively incorporate complementary segmentation information while preventing potential degradation from less relevant features. Looking forward, exploring higher frame rates or incorporating additional modalities such as textual information could provide new avenues for capturing temporal dynamics and semantic relationships in surgical action recognition.

6. Conclusion

This work investigated the challenge of surgical action triplet recognition through the lens of multimodal learning. Our research demonstrates that effective combination of RGB and segmentation information, enabled by the universal capabilities of SAM [21] and our proposed gated cross-attention mechanism, can significantly improve triplet recognition accuracy while providing additional segmentation capabilities.

The success of our approach provide several key findings. First, the combination of SAM's [21] robust feature extraction and CAM-guided prompting provides effective spatial understanding that reduce the domain gap typically challenging medical applications. Second, our gated fusion mechanism proves essential for selective information flow, allowing the model to dynamically weight different modalities' contributions. Third, the system's ability to generate segmentation masks from triplet-level supervision demonstrates the potential for reducing annotation requirements in practical applications.

Future lines of work. However, our work also reveals important limitations and areas for future research. The ineffectiveness of optical flow features at low frame rates suggests the need for better temporal modeling approaches.

The challenges faced with generic vision-language models highlight the ongoing need for medical-domain-specific solutions. A significant limitation we encountered is the inherent class imbalance in surgical action datasets, where certain triplet combinations appear far more frequently than others. This imbalance poses challenges for model training and can lead to biased predictions favoring common actions while struggling with rare but potentially critical surgical events.

To address these limitations, future work could explore several directions. The integration of temporal modeling at higher frame rates could improve action understanding. The development of medical-specific vision-language models could enhance semantic understanding. To tackle the dataset imbalance, future research could investigate techniques such as hierarchical classification approaches that first identify the action category before fine-grained triplet recognition, dynamic sampling strategies that prioritize rare triplets during training, or contrastive learning methods that help learn more discriminative features for rare classes.

In conclusion, our work provides practical insights for developing deployable systems in clinical settings. The dual-use nature of our approach, combined with its reduced annotation requirements, offers a promising direction for building more intelligent surgical assistance systems.

References

[1] Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. Exploring the impact of artificial intelligence: Prediction versus judgment. *Information Economics and Policy*, 47:1–6, 2019. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4, 6

[3] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021. 3

[4] Amir Baghdadi, Ahmed A Hussein, Youssef Ahmed, Lora A Cavuoto, and Khurshid A Guru. A computer vision technique for automated assessment of surgical performance using surgeons' console-feed videos. *International journal of computer assisted radiology and surgery*, 14:697–707, 2019. 1

[5] Jeremy A Balch, Benjamin Shickel, Azra Bihorac, Gilbert R Upchurch Jr, and Tyler J Loftus. Integration of ai in surgical decision support: improving clinical judgment. *Global Surgical Education-Journal of the Association for Surgical Education*, 3(1):56, 2024. 1

[6] Casey C Bennett and Kris Hauser. Artificial intelligence framework for simulating clinical decision-making: A markov decision process approach. *Artificial intelligence in medicine*, 57(1):9–19, 2013. 1

[7] Reiner Birkl, Diana Wofk, and Matthias Müller. Midas v3. 1–a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 5

[8] Andy Brock, Soham De, Samuel L Smith, and Karen Simonyan. High-performance large-scale image recognition without normalization. In *International conference on machine learning*, pages 1059–1071. PMLR, 2021. 4

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1, 4

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[11] Yiliang Chen, Shengfeng He, Yueming Jin, and Jing Qin. Surgical activity triplet recognition via triplet disentanglement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 451–461. Springer, 2023. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 3, 4

[14] Isabel Funke, Sebastian Bodenstedt, Florian Oehme, Felix von Bechtolsheim, Jürgen Weitz, and Stefanie Speidel. Using 3d convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition in video. *CoRR*, abs/1907.11454, 2019. 1

[15] Yixin Gao, S Swaroop Vedula, Carol E Reiley, Narges Ahmidi, Balakrishnan Varadarajan, Henry C Lin, Lingling Tao, Luca Zappella, Benjamın Béjar, David D Yuh, et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *MICCAI workshop: M2cai*, volume 3, page 3, 2014. 1

[16] R Girshick, J Donahue, T Darrell, and J Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. arxiv e-prints. *arXiv preprint arXiv:1311.2524*, 396, 2013. 5

[17] Jason L Green, Visakha Suresh, Peter Bittar, Leila Ledbetter, Suhail K Mithani, and Alexander Allori. The utilization of video technology in surgical education: a systematic review. *journal of surgical research*, 235:171–180, 2019. 1

[18] Shuangchun Gui and Zhenkun Wang. Tail-Enhanced Representation Learning for Surgical Triplet Recognition . In *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume LNCS 15011. Springer Nature Switzerland, October 2024. 15

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 4

[20] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 1

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5, 7, 9, 12, 14, 16, 19

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1, 3

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 6, 19

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 5

[25] Ziyi Lin, Chris Liu, Renrui Zhang, Peng Gao, Longtian Qiu, Han Xiao, Han Qiu, Chen Lin, Wenqi Shao, Keqin Chen, Jiaming Han, Siyuan Huang, Yichi Zhang, Xuming He, Hongsheng Li, and Yu Qiao. Sphinx: The joint mixing of weights, tasks, and visual embeddings for multi-modal large language models, 2023. 4

[26] Daochang Liu, Xintao Hu, Mubarak Shah, and Chang Xu. Surgical triplet recognition via diffusion model. *arXiv preprint arXiv:2406.13210*, 2024. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 3, 4

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 5, 6, 7

[29] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[30] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 6

[31] Lena Maier-Hein, Martin Wagner, Tobias Ross, Annika Reinke, Sebastian Bodenstedt, Peter M. Full, Hellena Hempe, Diana Mindroc-Filimon, Patrick Scholz, Thuy Nuong Tran, Pierangela Bruno, Anna Kisilenko, Benjamin Müller, Tornike Davitashvili, Manuela Capek, Minu Tizabi, Matthias Eisenmann, Tim J. Adler, Janek Gröhl, Melanie Schellenberg, Silvia Seidlitz, T. Y. Emmy Lai, Bünyamin Pekdemir, Veith Roethlingshoefer, Fabian Both, Sebastian Bittel, Marc Mengler, Lars Mündermann, Martin Apitz, Annette Kopp-Schneider, Stefanie Speidel, Hannes G. Kenngott, and Beat P. Müller-Stich. Heidelberg colorectal data set for surgical data science in the sensor operating room, 2021. 1

[32] Chinedu Innocent Nwoye, Cristians Gonzalez, Tong Yu, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. 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*, pages 364–374. Springer, 2020. 1, 10, 11, 13, 15

[33] Chinedu Innocent Nwoye, Tong Yu, Cristians Gonzalez, Barbara Seeliger, Pietro Mascagni, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis*, 78:102433, 2022. 2, 15

[34] Trishan Panch, Peter Szolovits, and Rifat Atun. Artificial intelligence, machine learning and health systems. *Journal of global health*, 8(2), 2018. 1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 19

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

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. 1

[38] Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. Rendezvous in time: an attention-based temporal fusion approach for surgical triplet recognition. *International Journal of Computer Assisted Radiology and Surgery*, 18(6):1053–1059, 2023. 2, 15

[39] Saurav Sharma, Chinedu Innocent Nwoye, Didier Mutter, and Nicolas Padoy. Surgical action triplet detection by mixed supervised learning of instrument-tissue interactions. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 505–514. Springer, 2023. 2

[40] Raphael Sznitman, Karim Ali, Rogério Richa, Russell H Taylor, Gregory D Hager, and Pascal Fua. Data-driven visual tracking in retinal microsurgery. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012: 15th International Conference, Nice, France, October 1-5, 2012, Proceedings, Part II 15*, pages 568–575. Springer, 2012. 1

[41] Mingxing Tan. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019. 3, 4

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

[43] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 1

[44] Ziheng Wang and Ann Majewicz Fey. SATR-DL: improving surgical skill assessment and task recognition in robot-assisted surgery with deep neural networks. *CoRR*, abs/1806.05798, 2018. 1

[45] Amine Yamlahi, Thuy Nuong Tran, Patrick Godau, Melanie Schellenberg, Dominik Michael, Finn-Henri Smidt, Jan-Hinrich Noelke, Tim Adler, Minu Dietlinde Tizabi, Chinedu Nwoye, Nicolas Padoy, and Lena Maier-Hein. Self-distillation for surgical action recognition, 2023. 15

[46] Yedi Zhang, Peter E Latham, and Andrew Saxe. A theory of unimodal bias in multimodal learning. *arXiv preprint arXiv:2312.00935*, 2023. 5

[47] B Zhou, A Khosla, A Lapedriza, A Oliva, and A Torralba. Learning deep features for discriminative localization. arxiv. *arXiv preprint arXiv:1512.04150*, 2015. 5

7. Appendix

7.1. Unsuccessful Attempts and Learnings

Our research journey included several approaches that, while promising in theory, failed to deliver expected results. Analyzing these failures provides valuable insights for future research directions.

7.1.1 Optical Flow Limitations

The attempt to incorporate optical flow as an additional modality faced significant challenges. The primary limitation stemmed from the dataset's 1 FPS sampling rate. This low temporal resolution resulted in large motion gaps between consecutive frames. It lead to the loss of subtle motion patterns crucial for action understanding. Therefore, the unreliable flow estimations that failed to capture meaningful motion information. The poor quality of optical flow features ultimately contributed little to the model's understanding of surgical actions

7.1.2 Failed Fusion Strategies

Several alternative fusion approaches were explored but proved less effective. We first explored naive modality concatenation, which involved concatenating all modality features followed by self-attention. This approach showed poor performance due to the confusion in prioritizing different modalities and the loss of RGB features' primacy as the main information source. The results reinforced the importance of maintaining RGB as the primary modality with other modalities serving supporting roles. Our experiments demonstrated the crucial role of the gating mechanism in effective fusion. By employing a sigmoid gate before the residual connection, the model learns to dynamically adjust the contribution of each modality at the feature level. This selective mechanism allows the model to emphasize relevant segmentation features when they provide complementary information (such as clear instrument boundaries) while suppressing them when RGB features alone are sufficient or when segmentation information might be misleading. The improvement in performance with gated fusion suggests that different surgical actions benefit from varying degrees of segmentation information, making this dynamic weighting capability essential for optimal recognition.

7.1.3 Alternative Semantic Alignment Approaches

We explored using vision-language models for semantic alignment of SAM [21] masks, which proved unsuccessful. The approach attempted to match masked images with predicted labels using vision-language models such as CLIP [35] and BLIP [23]. However, the lack of medical domain knowledge in CLIP/BLIP training data lead to poor performance in distinguishing similar anatomical features. The approach struggled particularly with target organ identification, while different anatomical structures produced very similar feature representations. Thus, the generic visual features from vision-language models proved insufficient for medical-specific distinctions

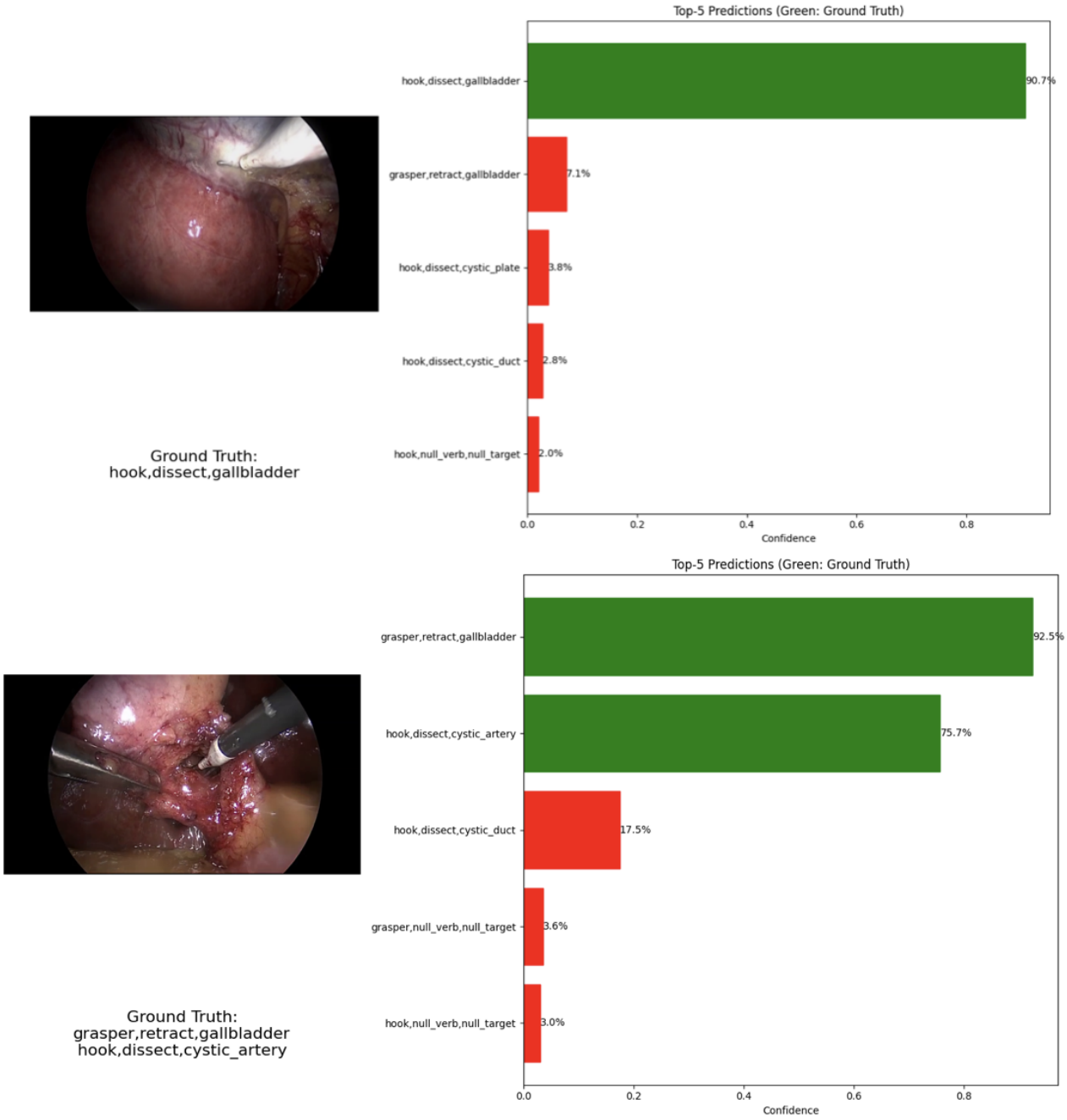


Figure 12. Correct Recognition Result.In this surgical frame, the model correctly identified all ground truth labels (shown in green bars), demonstrating accurate recognition of surgical workflow.

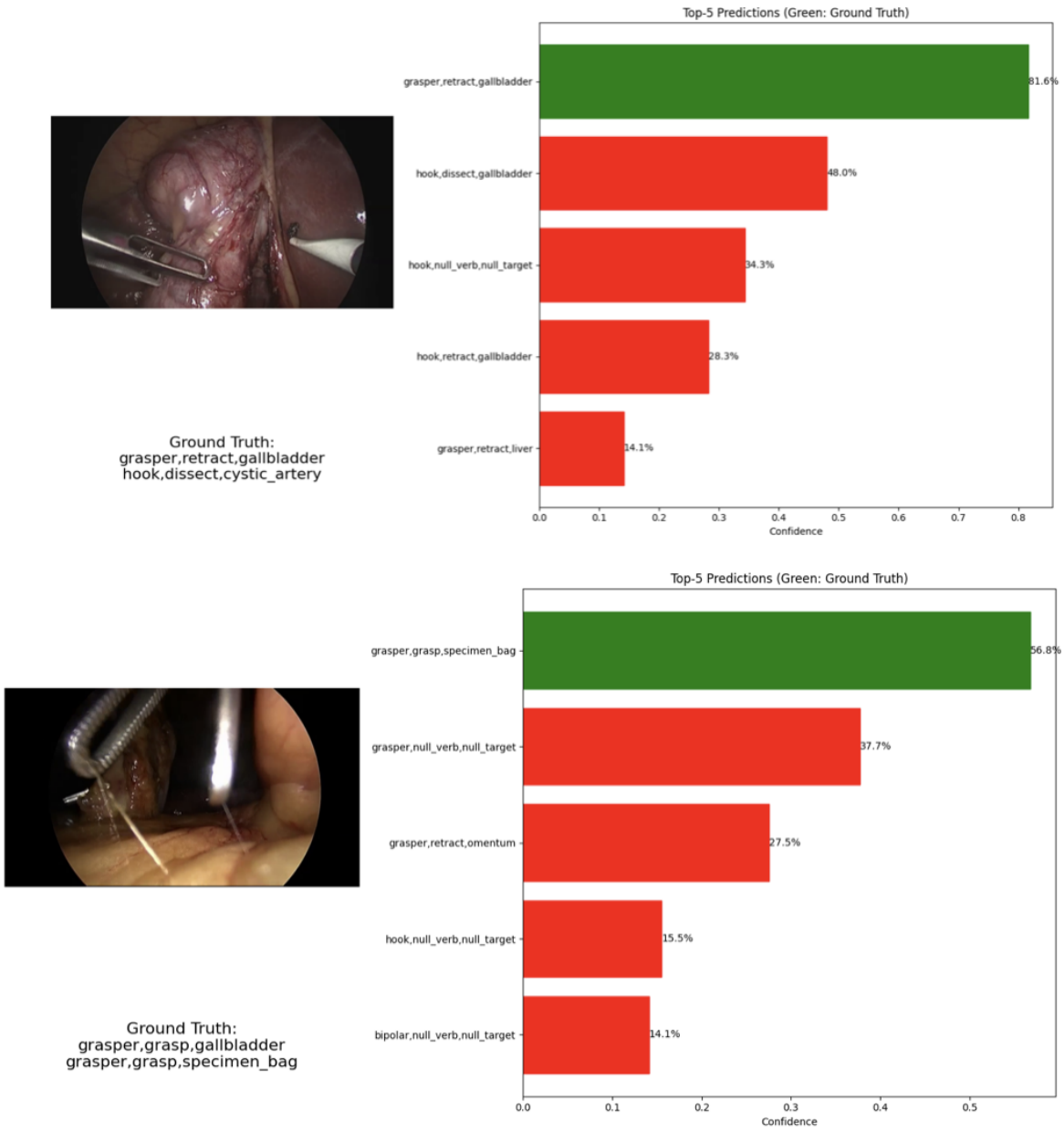


Figure 13. Partial Match Recognition Result. The model captured some but not all ground truth labels (green bars = correct, red = incorrect). This indicates partial understanding of the surgical phase.

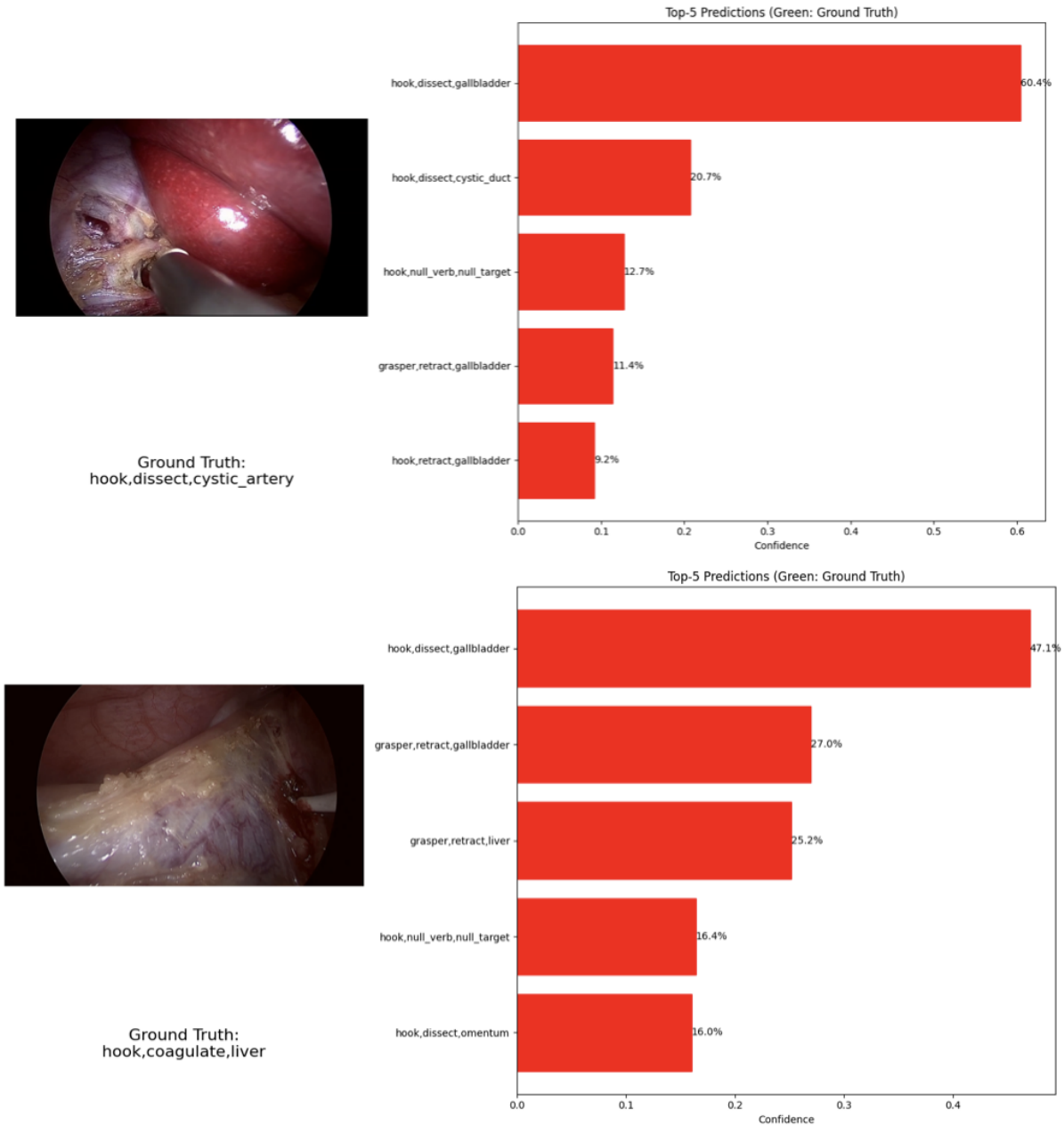


Figure 14. Complete Misclassification Recognition Result. Despite high confidence scores, the model failed to identify any ground truth labels (all bars in red), suggesting a challenging frame.