Rethinking CLIP-Style Fusion for Surgical Video Analysis in Low-Data Scenarios

FAHIM FARDIN

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on **"Rethinking CLIP-Style Fusion for Surgical Video Analysis in Low-Data Scenarios"**, submitted by **Fahim Fardin (ID: 221-35-1045)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S M Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name	:	Fahim Fardin
Date of Birth	:	25 November 2002
Title	:	Rethinking CLIP-Style Fusion for Surgical Video Analysis in Low-Data Scenarios
Academic Session	:	2022-2025

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997) *
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done) *
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

FAHIM

(Student's Signature)

221-35-1045

Student ID

Date: 27-12-2025

(Supervisor's Signature)

Mr. Musabbir Hasan Sammak

Name of Supervisor

Date: 27-12-2025

NOTE: * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City,
Ashulia.Dhaka, Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.

Author's Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours
faithfully,

_____
(Supervisor's Signature)

Date:

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian, *Daffodil International University* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.

(Supervisor's Signature)

Full Name : Mr. Musabbir Hasan Sammak

Position : Lecturer (Senior Scale)

Date : 27-12-2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Rethinking CLIP-Style Fusion for Surgical Video Analysis in Low-Data Scenarios

(Student's Signature)

Full Name : Fahim Fardin
ID Number : 221-35-1045
Date : 27-12-2025

Rethinking CLIP-Style Fusion for Surgical Video Analysis in Low-Data Scenarios

Fahim Fardin

Thesis submitted in fulfillment of the requirements for
the award of the degree of
Bachelor of Science

Department of Software Engineering (Major in Data Science)

DAFFODIL INTERNATIONAL UNIVERSITY

November 2025

ACKNOWLEDGEMENTS

All praise and gratitude are due to Allah (SWT), the Most Gracious, the Most Merciful. Without his infinite grace, strength, and guidance, this journey would never have been possible. Without his divine will, this work would not have been possible.

I want to express my deepest gratitude to my esteemed supervisor, Mr. Musabbir Hasan Sammak. Your unwavering support, insightful advice, and deep knowledge have contributed immensely to the completion of this thesis. Your guidance has not only shaped this research but has also been a source of inspiration for me, for which I am sincerely grateful.

I am also deeply grateful to my family, whose unconditional love, patience, and encouragement have always been a source of inspiration for me. I am also grateful to my friends, whose cooperation and support have helped me move this work forward

DEDICATION

This thesis is dedicated to my family and mentors for their unwavering support, guidance, and encouragement throughout this journey.

ABSTRACT

Surgical Phase Recognition (SPR) is essential in supporting intraoperative decision support, minimally invasive surgery training, and operating room workflow. However, the correct identification of phases in laparoscopic cholecystectomy videos is a difficult task due to the weak visual features between phases, commonly occurring occlusions, imbalance in classes, and unavailability of large-scale annotated surgical videos. More recent approaches have been exploiting large vision-language models like the Contrastive Language-Image Pretraining (CLIP), where the fusion mechanisms are trained on large general-purpose datasets and might not be transferred to a small-data, domain-specific medical task. To overcome these shortcomings in the present paper, we recast CLIP-style multimodal fusion under low-data settings of surgical video analysis by comparing lightweight fusion strategies such as additive fusion, concatenation-based multilayer perception (MLP) fusion, gated fusion, and shared-projection fusion—on a frozen CLIP-based backbone. Frame-level phase classification experiments on the Cholec80 laparoscopic cholecystectomy data set indicate that simple fusion, where additive fusion yields optimal robustness by trading off a less complex number of parameters and reduced overfitting, and gated fusion show competitive generalization ability in adaptive modality integration. Meanwhile, due to data scarcity, they are more sensitive, and concatenation-based or shared-projection fusion exhibits inferior unstable behaviors. The error-flow analysis also reveals the deep-seated confusions across visually similar stages, e.g., Gallbladder Dissection and Cleaning and Coagulation, and this illustrates the physical constraints of the visual evidence when data are scarce. Comprehensively, this thesis proves that effective multimodal surgical phase identification in clinical practice must reconsider the design of fusion rather than further increase the model complexity, and lightweight fusion schemes are a feasible and practical solution to low resource setting surgical video analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

z_v	Vision embedding (output of the Vision Encoder)
z_t	Text embedding (output of the Text Encoder)
z_f	Joint embedding in the shared latent space
z_vis	Final visual feature (output of the Vision Encoder)
z_t	Final text feature (output of the Text Encoder)
I	Input image (frame from a laparoscopic surgery video)
X_"patches"	Flattened image patches
W_"patch"	Trained projector matrix for image patches
b_"patch"	Learned bias for image patches
Q	Query matrix for self-attention mechanism
K	Key matrix for self-attention mechanism
V	Value matrix for self-attention mechanism (image patches)
d_k	Dimensionality of query and key vectors
g	Gate value in gated fusion (learned weighting between vision and text)
α_v	Learnable weight for vision modality in weighted-sum fusion
β	Learnable trade-off parameter between vision and text in weighted-sum fusion
τ	Temperature parameter in contrastive loss
σ	Sigmoid function used in gating mechanism
W_g	Weight matrix for the gating function
R@k	Top-k accuracy (where k is 1, 5, or 10)
C_ij	Confusion matrix element (number of samples of class i classified as class j)
W_v	Projection matrix for vision embeddings
W_t	Projection matrix for text embeddings
z_v" Linear"	Linear projection of vision embeddings

z_t" Linear"	Linear projection of text embeddings
g⊙	Element-wise multiplication (gate-weighted fusion)
[z_v; z_t]	Concatenation of vision and text embeddings
"sim"(z_v, z_t)	Cosine similarity between vision and text embeddings
z_f	Final fused embedding of vision and text
N	Total number of test samples

LIST OF ABBREVIATIONS

AdamW	Adam with Weight Decay
ALBERT	A Lite BERT
AS	Auto-Regressive / Applied Science
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapping Language-Image Pretraining
CLIP	Contrastive Language-Image Pretraining
CNN	Convolutional Neural Network
CNN-LSTM	Convolutional Neural Network - Long Short-Term Memory
Concat-MLP	Concatenation Multi-Layer Perceptron
FPS	Frames Per Second
InfoNCE	Information Noise Contrastive Estimation
IRCAD	International Research Center for Advanced Disease
LSTM	Long Short-Term Memory
MDPI	Multidisciplinary Digital Publishing Institute
MLP	Multi-Layer Perceptron
MMA	Mixed Martial Arts
QK	Query-Key (Used in Attention Mechanisms)
ResNet	Residual Network (Deep Learning Architecture)
SMOTE	Synthetic Minority Over-sampling Technique
SPR	Surgical Phase Recognition
ViL	Vision-and-Language
ViLT	Vision-and-Language Transformer
ViT	Vision Transformer
VLM	Vision-Language Models
VRAM	Video Random Access Memory

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

A basic issue of the surgical data science problem is Surgical Phase Recognition (SPR), which attempts to automatically compute the various phases during an intraoperative video to enhance the safety, efficiency and clinical effectiveness of the surgery activities (Golany et al., 2022). Using correct SPR, many downstream applications can be achieved that can comprise running room workflow; surgical examination systems; auto-generated reporting and real-time choice frameworks that could be used as second-opinion in minimally invasive operation. Among such procedures (Kirtac et al., 2022), the laparoscopic cholecystectomy is regarded as the best researched reference application as it has a standard (anatomical and technological) structure that is typically broken down into a series of relatively clear steps such as preparation, Calot triangle dissection, clipping and cutting, dissection of gallbladder and cannulation as well as cleaning and coagulation (Twinanda et al., 2016). These stages should be diagnosed correctly, which is particularly important in procedures with high safety aspects such as the triangle dissection by Calot, which with an erroneous diagnosis can cause serious complications.

Cholec80 data set has been made into one of the most used public datasets in the study of SPR, which contains 80 annotated laparoscopic cholecystectomy videos of seven surgical phases. Despite its popularity, Cholec80 is a model of a number of challenges in the real world such as variations in lighting conditions, anatomic appearance, level of expertise of the surgeon and the distribution of instrument usage and perturbations such as smoke and bleeding light. It is known that all these factors along with high frequency occlusions and noise, cause a high severity of recognition performance, of course when models are tested- across institutions or in case of a domain change (Kirtac et al., 2022). It is also problematic in the frame-level prediction because surgical phases are significantly different in time and also introduce a temporal imbalance between the frames that can be classified using our proposed method in which our method depends heavily on them.Traditional approaches to SPR have been constructed largely based on convolutional neural networks with recurrent or time-varying models e.g. LSTM and GRU designs, to learn spatial and temporal dependencies. These techniques have allowed more accurate results, but they are highly influenced by unique visual

resemblance amidst frames specifically when association of surgical tools and anatomical structures and camera viewpoints clash. In order to solve this ambiguity, multimodal systems that use other types of information (e.g., natural descriptions, procedural semantics) have been suggested (J. Zhang, Barbarisi, Kadkhodamohammadi, Stoyanov, & Luengo, 2023). Multimodal Fusion Models Surgical processes are often semantically directed and structured in nature as to be described textually, in which case it is normally possible to describe the purpose of the procedure though often with reference to uncertainty as a function of text alone when distinguishing visual cue is no longer possible.

Recent VLMs, such as Contrastive Language-Image Pretraining (CLIP), have demonstrated excellent representation learning by aligning image embeddings and text embeddings in a common latent space under the influence of large-scale natural images and the corresponding texts (Radford et al., 2021). Nonetheless, the ability of these models to assist in surgical video analysis is limited by domain shift because of variations between natural and endoscopic scenes, or data-intensive pre-alignment. CLIP-style dual-encoder architectures do not perform as well in surgery that they are not trained on the domain. Some of these limitations are alleviated using specialized surgical VLMs, e.g. trained on surgical lectures or procedure-specific narration, at the cost of new computational and data demands differing (Yuan, Srivastav, Yu, et al., 2025).

Due to the restrictions of clinical set-up and the lack of annotated surgical data, most recently it is gaining popularity to simplify minimalistic multi-modal fusion. Instead of involving large cross-modal transformers, they research the small means of fusion e.g., additive fusion, gated fusion and low dimensional projection layer - with which it is allowed to interact efficiently (parameter overhead) in its cross-modality interaction (Mungoli, 2023). Empirically, it has been proved that such lightweight fusion designs are highly effective in enhancing the cross-modal information flow and are robust and memory/compute efficient (Q. Zhang et al., 2024). We believe it is interesting to note that returning to concept of CLIP-style fusion of lightweight analogs is one of the options in scalable and deployable SPR in clinical practices with low-data in low-weight clinical practice.

1.2 Problem Statement

Surgical Phase Recognition (SPR) is meant to discover the specified procedural steps in laparoscopic cholecystectomy, such as preparation, Calot triangle dissection, clipping and cutting, gallbladder dissection, cleaning and coagulation, and other organized steps, but not anatomical-phase (EAP) differences. Two basic factors that mostly challenge the application of SPR process are (1) high visual variability among operating room settings, and (2) low visual difference between successive and gradually changing surgical phases (Abiyev, Altabel, Darwish, & Helwan, 2024). Public datasets (like Cholec80) make these problems even more complicated as they have severe imbalance in phase duration and have shared intraoperative artifacts, such as tool occlusions, motion-induced blur, smoke, bleeding, and variation in illumination. These visually induced disturbances are also major limitations to correct phase discrimination, especially at safety-important steps like Calot triangle incision and transition maneuvers to gallbladder incision and cleaning and coagulation (Abiyev et al., 2024).

Canonical SPR pipelines that are convolutional, recurrent or transformer models of temporal data are limited in the aspect that they only use visual data in order to learn spatial and temporal associations. Even though they have contributed to making the recognition more precise, these methods remain constrained by the characteristic vagueness of surgical observation, particularly when frames between successive stages appear similar or sub-frame characteristics are impaired. Thus, vision-only models in most cases do not have the ability to generalize to low-data or cross-domain conditions (Kondo, 2025a).

To minimize the visual ambiguity, less old investigations have also been conducted on multimodal SPR models based on multimodal information like textual phase descriptions with visual properties. Despite the advantage of this type of multimodal solution, the majority of them rely on calculation which involves fusion mechanisms like cross-attention transformers, multi-stage encoder or a massive domain-specific vision-language model. The resulting approaches use memory and latency heavily in addition to training overhead and this inhibits their use in clinical environments with real-time or resource-constrained environments.

Antithetical vision-language models, e.g. CLIP, have also been demonstrated to provide an escape mechanism of cross-modal encoder-decoder cross-modal learning. Nevertheless, the cosine-similarity-based alignment of CLIP that is particular to large-scale natural image-text setups does not perform so well in the fine-grained surgical phase recognition when visual and textual accounts might be weakly related or with misalignment in domain. This in turn obliges us to propose fusion mechanisms that will be functional to introduce the procedural semantics that are entrenched in text as well as the uncertain way that visual evidence functions to make use of CLIP in the operation rooms (Radford et al., 2021).

In spite of the growing momentum on the topic of multimodal SPR, this research gap can be identified: so far, no research has delved into the systematic exploration of which lightweight fusion strategy proves to be fruitful in CLIP-based surgical phase recognition with low-data support. In particular, the available research lacks controlled ablation research in simple fusion processes e.g. additive fusion, gated fusion, concatenation-based fusion and shared-projection based fusion under a single experimental condition. Moreover, the fusion design consideration is absent on the performance of generalization and the other phase classification (Q. Zhang et al., 2024).

It seems to be an issue that needs to be studied in more detail, and this is why this thesis will go ahead and do just that but propose, perform and test a series of lightweight heads of fusion using a frozen CLIP lottery ticket backbone in a way that identifies surgical phases in the Cholec80 dataset. The research question will be to identify fusion techniques, which create the best trade-offs between classification, robustness and computational cost and thus contact multimodal SPR systems feasible and useful under low-data clinical environments.

1.3 Motivation

The motivation for this study arises from growing clinical need within realistic operational rooms, which requires the correct and efficient analysis of surgical data. As the deriving X-COM systems gradually gain popularity at hospitals, automatic surgical phase recognition has been assuming a significant role in supporting the efficiency of the workflow, assisting in the training of minimally invasive surgical procedures as well as monitoring the quality of the real-time state of the surgeries (Abiyev et al., 2024). However, due to the high cost of manual annotation, which needs time, skills, and finances, and the vast amount of intraoperative video data accessible to medical researchers, a significant percentage of them remain uncoded. One way out of this problem is automated surgical

phase recognition systems, which require fewer manual annotations and can be used to provide real-time analytics e.g., estimation of procedure progress, perception of resource use, and skills-based deviations during surgery (Yuan, Srivastav, Yu, et al., 2025). Moreover, consistent and objective determination of the phases will foster productive, organized surgical training to facilitate assured feedback back to the trainee, as well as facilitate a comparable review procedure across the institutions (Abiyev et al., 2024).

Technically, there are severe software limitations on the practical use of AI in surgery. During surgery, inference should be low latency; nevertheless, high-end Graphics Processing Units (GPUs) are usually limited in clinical practice. Broadly speaking, the existing multimodal strategies are divided into a broad group of procedures employing computationally intensive cross-attention transformers or large task-specific backbone models that are difficult to apply in real-time applications and in low-resource settings. As an appealing computational cost, fixed dual-encoder models, like CLIP, have a well-developed, modular, and reproducible starting point (Yuan, Srivastav, Yu, et al., 2025). In the case of CLIP as a fixed backbone, compression techniques can be experimented with separately, resulting in comparatively controlled experiments with architectural simplicity and reproducibility (Radford et al., 2021).

Scientifically speaking, the trade-offs between accuracy of classification, multimodal integration quality, and computational efficiency in lightweight fusion strategies have not been investigated fully at least in the low-data case of surgical video analysis. Though there have been reported that there are some more sophisticated multimodal fusion methods, no systematic comparison research is reported on simple fusion heads (e.g., additive, gated, concatenation-based, and shared-projection) with the CLIP pipeline in the analogue of unified architecture. These trade-offs should be learned because they will play a key role in developing deployable surgical AI systems that can balance performance and robustness, generalization and efficiency (Q. Zhang et al., 2024). It is the attempt to address this gap with this thesis in reimagining CLIP-style fusion techniques that are effective and viable in resource-limited clinical environments.

1.4 Significance of the Study

This study is significant in advancing both the clinical applicability and methodological understanding of surgical phase recognition (SPR) under low-data constraints. The idea has been confirmed through initial vision-based SPR systems, which have suggested that robotized phase identification may prove beneficial in surgical workflow assessment, training and quality control, however (Golany et al., 2022), they were crippled by the visual ambiguity and non-generalization in the crowded laparoscopic setting (Twinanda et al., 2016). After research revealed that the public datasets such as Cholec80 are being plagued by class imbalance, occlusions and variations of illumination that impose further constraints on the correct phase discrimination in safety critical tasks (Kirtac et al., 2022).

Recent literature calculates multimodal representation that determines the relationship of visual and semantic parts to surmount these constraints using textual or semantic data besides visual characteristics. Although large multimodal transformer architectures and domain-specific vision-language models have recently achieved success with improvements in computational, memory, and inference latency requirements too expensive to run in real time in resource-constrained clinical settings (Y. Li, Zhao, Li, & Li, 2024). The shortcomings of such a system drive the studies of useful and implementable multimodal systems, which have potential practice effectiveness with little annotated data (S. Li & Tang, 2025).

The contribution of this thesis is demonstrated by the fact that using pre-trained a single clipped-style two-encoder backbone, with lightweight aggregation mechanisms, it is possible to effectively recognize surgical phases without necessarily using heavy-weight cross-attention transformers or large task-specific models (Radford et al., 2021). Viewing CLIP as a frozen and modular representation backbone, the given framework provides composable experimentation of fusion strategies with regard to computational efficiency, architectural reproducibility and simplicity.

Methodologically, the work is among the initial systematic and controlled comparisons of a number of light-weight fusion models (additive fusion, weighted summation, concatenation based MLP fusion, gated fusion and shared-projection (Hadamard) fusion within the Unified CLIP based SPR pipeline. Multimodal networks that exist now are often complex structures with fusion operations incorporated and so the individual performance of each functionality is hard to disaggregate. Conversely, the current thesis formulates fusion design as the most

important manipulation variable, which yields fresh understanding in trade-offs between classification performances and its soundness, generalization and computation costs (Mungoli, 2023).

Additionally, the projection-based pre-fusion alignment tool that is introduced and discussed in this paper is useful in multimodal integration when visual and textual features are projected to a shared latent space by light-weight linear projections, along with normalizing layers. It is an architecture that supports efficient cross-modal communication with low-data requirement, which can maintain interpretability and low parameter burden - a problem that has been observed in existing works on multimodal fusion (Faray De Paiva, Yuan, Srivastav, & Padoy, n.d.). The given overall assessment model - through macro averaged performance measures, phase-wise recall provides feedback on the model performance in clinically critical stages.

Overall, this paper provides usable design recommendations on deployable surgical AI systems as it demonstrates that it is not making the existing models more complicated but reconsidering fusion designs is vital in developing resilient multimodal SPR in low-resource clinical environments. The findings suggest the extension of light-weight CLIP-style fusion techniques to real-time surgical video - understanding and introduce a saleable embodiment, leading to additional studies on effective, interpretable, and sustainably (surgical) data science (Yuan, Srivastav, Navab, & Padoy, 2025).

1.5 Research Questions

Multi-modal learning is a relatively new direction to the surgical phase recognition problem; however, it remains a significant issue to well combine the visual and textual information under low-data-limiting conditions. Such models as CLIP-style models are based on large-scale cosine-similarity-based alignment, which is difficult to replicate in domain-specific surgical conditions with limited labelling data and limited computing power. Therefore, this paper examines the methods of lightweight multimodal fusion that could be employed to promote the effectiveness and usability of surgical phase recognition system. The questions of the research are as follows:

- What is the impact of different multimodal fusion models that are lightweight (such as additive fusion, concatenation-based fusion, gated fusion and shared-projection fusion models) on the accuracy stability and generalization of surgical phase recognition models on small scale domain-specific surgical video data?

- To what degree are lightweight CLIP style fusion mechanisms potentially useful to replace large-scale cosine-similarity-based alignment in low-data surgical video analysis not only in terms of computational efficiency, but also in terms of feasibility to deploy to clinical issues?

1.6 Research Objectives

- To design and integrate four lightweight multimodal fusion mechanisms—additive fusion, concatenation-based fusion, gated fusion, and shared-projection fusion—within a CLIP-inspired architecture adapted for surgical phase recognition.

- To build and test both hybrid fusion schemes on a low-resource surgical video database to establish its performance, stability, and the capacity to generalize in the presence of data-scarce conditions.

- To compare the tested fusion methods in a systematic way and to decide on which designs would be most successful in alleviating the absence of large-scale multimodal registration but at the same time offer solid surgical phase recognition performance.

- To provide clear and practical suggestions to serve as an informational basis in the future in order to select the most suitable multimodal fusion strategies to be used in medical AI applications that are highly limited by datasets size, computation resources, or clinical facilities.

1.7 Research Scope and Limitations

In this section, the primary limitations of the study that are due to Cholec80 data, methodological decisions and flow of the evaluation process are taken into consideration. It talks about the implications of frame-level analysis, frozen CLIP encoders, single-dataset dependence, class-imbalance, and prompt design on findings and describes the scale on which the results may be interpreted.

1.7.1 Research Scope

- The dataset of 80 annotated cholecystectomy videos, Cholec80, is used in the experiment, which is categorized into seven stages of the surgery. The data set is scaled down on the original video in 1 FPS, such that it shares a consistency on the frames used to be classified (Abiyev et al., 2024). The paper deals with frame-level recognition in comparison with the temporal dependencies or sequence modelling (Faray De Paiva et al., n.d.).
- We are using the CLIP (Radford et al., 2021) two-encoder design. The CLIP encoders are also deterministic and thus provide consistency over models to the pipeline as well as remove the bias of domain adaptation (Yang, Zhang, Wang, & Xie, n.d.). That is, it is not the study of CLIP encoders in surgical images maximised, but one that is intended to examine the impact of lightweight fusion methods (Faray De Paiva et al., n.d.).
- The text prompts at the phase level can extract semantic information about each phase of surgery (Faray De Paiva et al., n.d.). These anchors work as anchors of the text embedding space, which restrains the multimodal correspondence of visual properties and textual expression of phrases (Radford et al., 2021).
- We contrasted five fusion heads in this work additive, concat-MLP, gated, weighted-sum and projection shared to ascertain the impact of different fusion methods on the matching of text and image embeddings (Yu et al., 2025).
- The retrieval models are evaluated according to R 0.5 k (R 0.1/5/10) and Top-1 RM, and we use macro-averaged precision/recall/F1 to measure the indexing system. The PR/F1 scores and the confusion matrices of the per-phase are also presented in this study to determine the performance of our model in individual surgical stages and whether there is any stage our model is failing to work properly (Abiyev et al., 2024).
- We also come up to allow new-video object recognition and retraining of the model in a manner that it can adjust to new surgical data in the future. This works in compliance with the need for the long-term model deployment and management of clinical applications (Yu et al., 2025).
- The work does not use the time modelling outside the frame-sampling (no sequence learning, including LSTM models or Transformer models). Moreover, CLIP encoders fail to map fine-tuning, and therefore domain adaptation and encoder fine-tuning are down the line (Faray De Paiva et al., n.d.).

1.7.2 Limitations

- The study assumes frame-level processing, and hence, a model does not reflect the smoothness of the surgical stages through time. The steps between phases of interphases, including Calot Triangle Dissection – Dissection, could, therefore, not be determined except by sequential context or motion patterns that are associated with sequential relations of actions to visual interactions modelling (S. Li & Tang, 2025).
- The CLIP encoders in the experiments are frozen to enhance more reproducibility and to eliminate the complexity of domain adaptation. However, even that could be compromising optimum performance since it is possible to selectively train the CLIP encoders to match surgical data more closely and result in a more accurate model (Abiyev et al., 2024).
- one operation (laparoscopic ilectomy of the cholecystis) and one institution. In this paper, therefore, we were not able to confirm the usability of the model for other methods of operation (e.g., robot surgery and open surgery) or in other hospitals (Faray De Paiva et al., n.d.).
- The out-of-balance between phase lengths and the phase frequencies in Cholec80 are train/validation split-wise with stratified splits and class weights loss. However, there would still be some relatively rare to extremely infrequent stages with performance issues of being under aligned in the set. Less available examples of phase classes are going to be identified less accurately (Abiyev et al., 2024).
- The stage-specific textual cues that are provided in the present study are artificial; therefore, the quality of textual cues and uniformity wherever it exists may influence the results. The article discusses a limited number of phase prompts and does not compare different prompt engineering methods and bigger collections of phase prompts. Variations in timely phrasing are also capable of influencing the functioning of the multimodal model significantly (Abiyev et al., 2024).

1.8 Thesis Organization

This thesis is separated into six chapters. Chapter 1, Introduction, presents the research problem, describes the motivation of this work, and mentions the purpose of this research. It offers research questions and also delineates what is of interest and what it is unable to discuss further. Discussion The value of this study is discussed based on its academic and practical value in the recognition of surgery phases.

Related Work (Chapter 2) This chapter presents a literature survey of the surgical phase recognition and multimodal learning in general in the focus of VLMs. The following section will elaborate on how the SPR methods have evolved along with three issues based on the limitation of the datasets and the inequitable modalities gaps and several fusion strategies that previous works have followed. It highlights the gap that this piece of work attempts to cover, that is of lightweight fusion substances of CLIP-based models.

In Chapter 3, Methodology, the methodology that was taken in this work is explained: dataset (Cholec80), CLIP-based pipeline implementation, and design of lightweight fusion heads. The chapter also defines the measures of evaluating the results and diagnosing the performance of the model and the clear description of the experimental setup.

In Chapter 4 Experiments and Results Experimental rig the results of the work on four various fusion heads are reported. We compare the findings regarding retrieval measures, nostalgia classification and diagnostic measures (confusion matrices and per-phase precision/recall measures) using each combination of the fusion strategy. The competency of the suggested fusion techniques for the achievement of improving SPR of Cholec80 is contrasted in the competing chapter.

The results are explained in detail in Discussion Chapter 5. It discusses the influence of these fusion heads on the model accuracy, efficiency and its potential clinical application and the limitations of the currently conducted study. Suggestions on areas of future research are provided.

The conclusion, chapter 6, Conclusion, reflects the key conclusions of the work in addition to its contributions in the sphere of surgical phase recognition. It then concludes by giving remarks on what we can do with our work and the future of research in multimodal fusion and surgical AI.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction to Surgical Phase Recognition

The study reviews the literature on surgical phase recognition (SPR) and the multimodal learning attempts to enhance the accuracy and effectiveness of phase detection. It is focused mostly on limited fusion methods in multimodal models that integrate visual and textual data to improve SPR, and laparoscopic cholecystectomy on Cholec80 dataset. The objective is to identify gaps that exist in present past and then aim at creating a niche in this thesis in the form of effective fusion strategies that can be applied with real time clinical applications. Focusing on light-weight fusion heads, such as additive-, weighted-sum-, gated- and projection-shared-heads, this survey preconditions addressing such shortcomings found in current SPR models (Yuan, Srivastav, Yu, et al., 2025).

The importance of Surgical Phase Recognition (SPR) in enhancing surgical workflow, patient safety and surgical education is great. Real-time phase detection enables feedback in the middle phases to be automated, which is significant to training and decision-making so that it informs the surgical team whether they are about to make a possible error or an adverse event occurs. This Automatic phase recognition plays a highly important role in laparoscopic cholecystectomy because human beings may find it hard to identify phase accurately like calot triangle dissection in the normal operation room setting. This thesis is based on the combination of CLIP (contrastive language-image pretraining) and lightweight fusion heads, where it is motivated to combine textual phase description and visual information to achieve higher performance on real-time phase detection (Radford et al., 2021).

Surgical Phase Recognition (SPR) Surgical Phase Recognition Surgical Phase Recognition is the problem of identifying and categorizing different phases of a surgical operation using visual data in surgical video footage, typically in comparison to textual explanations. SPR is required to further enlighten the information of the surgical processes, and to offer the real-time feedback to the surgery teams, and to have patient safekeeping. An example of this is cholecystectomy in laparoscopic surgery which can help reduce human error through accurate identification of phases, increase surgical training and use of automatic phase identification

systems to track the progression of surgeries. They allow phases like Triangle Dissection of Calot and Extraction of the Gallbladder that are considered important in efficient workflow of the OR to be quickly detected in the operating theatre. Another way SPR can increase the training of surgical residents is to provide real-time feedback in the form of guiding during live operation and learning phase classification automatically (Golany et al., 2022).

Despite the importance of SPR there are several challenges in SPR. Among the more difficult is the issue of class imbalance: certain surgical phases are less well represented in datasets including Cholec80 Busch, causing biased models to not learn such phases (Abiyev et al., 2024). Also, phase recognition is challenging because of the visual changeability of surgical videos to differences in lighting conditions, alteration of viewing angles, tool occlusions and motion artifacts. Complex problems like Calot triangle dissection are made worse by these issues as phase boundaries are not well defined and hard to see visually. Such challenges can be likely to occur in the traditional versions of CNN-based or LSTM based models that are vision-only-informed, in which the information would hardly suffice to capture changes between visually similar stages. Additional information provided by the spoken description can prove quite useful in model capturing and modeling performance in such cases (Zhou, Yang, Loy, & Liu, 2022).

2.2 Multimodal Learning for Surgical Phase Recognition

'Multimodal learning' is a term applied to refer to the integration of various types of data or modalities to enhance model accuracy. Multimodal learning can be applied to mean the combination of modalities (visual data learnt in surgical videos and the instruction in a text) in the context of SPR. Multimodal learning involves visual cues and textual characteristics that enable the model to have a comprehensive picture of the phases and, therefore, enhance the performance of phase classification. Only visual information is usually faced with the issues of tool occlusions, motion artefacts and vague environmental conditions, particularly in vital parts of surgery. Textual information also contains contextual cues in the form of phase-specific descriptions or purposes of surgery and enhances the detection of phase transitions and model robustness.

Multimodal models are a number of models that have contributed significantly to the research field besides giving alternative options of aligning text and image data. Contrastive Language-Image Pretraining (CLIP) is an OpenAI computing machine that has a dual-encoder structure; it is a system in which images and text are processed separately and projected to an overlapping embedding space. Zero-shot learning is possible with this method since the model can understand and classify images based on a text description without incurring a lot of fine-tuning on task-specific data. The effectiveness of CLIP on a variety of tasks, such as image captioning and zero-shot image classification, has prompted us to use it on the task of suture recognition in surgery, where the descriptions of phases are coupled with visual feedback to enable correct prediction of the surgical phases (Y. Zhang et al., 2022).

Besides LAVA, vision-language (VLM) embeddable models have also been suggested, including ALIGN and BLIP. Large collections of images with natural language descriptions have been exploited to develop text-image alignment by ALIGN (A Large-scale vision-language pre-training model) (Xing et al., 2025) and BLIP (Bootstrapping Language-Image Pre-training)(J. Li, Li, Xiong, & Hoi, 2022). The models are the state-of-the-art results in terms of ranked image-text retrieval and also in terms of providing descriptions of images, and they attest to the success of learning the tasks in a joint way whenever context between the different modalities was required. A different type of instant applicability can be shown by ViL T (Vision-and-Language Transformer), which can be deployed as a concurrent percent to show that transformers are more than capable of performing the multimodal task, and doing this effectively aligns text blustery (Kim, Son, & Kim, 2021). These models and CLIP represent big advances in multimodal learning and thus have been impactful on use cases that are thought about in SPR – where one would like to combine the description of textual phases with visual modalities to better discern the surgical phases.

Other necessary elements of multimodal models are the text encoder that encode text-based information just as much as surgical stage descriptions and procedural instructions, patient and clinical setting information are, to meaningful embedding vectors, which can be added to visual features to enhance recognition accuracy. Transformer-based text encoders, such as BERT, DistilBERT and ALBERT, are used as the standard text encoders in multimodal learning tasks in the modern world.

- BERT is a pre-training deep neural network model, which has demonstrated the state-of-the-art in various natural language processing tasks. It operates by operating on a text in two directions, both left to right and right to left, - in the process producing embeddings with contextual information about each word in the line. This enables the model to capture the fine interactions between the surrounding words and attains superb performance in the natural language understanding tasks like question answering and sentence classification. The biggest weakness of BERT is that BERT models are usually large and/or computationally intensive, so they cannot be readily utilized in resource-constrained systems (Radford et al., 2021).

- Thus, DistilBERT is a distilled variant of BERT intended to cut model sizes, and inference time without significantly sacrificing the performance of BERT by applying the above-mentioned knowledge distillation trick, which decreased model sizes by approximately 60 percent without losing the knowledge trained into BERT. Therefore, when real-time processing is involved, especially in the context of surgical stage detection when speed is a priority, DistilBERT appears to be a better option (Sanh, Debut, Chaumond, & Wolf, 2020).

- ALBERT (Lightweight BERT) is another variant of lightweight BERT that shares weights across layers in order to decrease the number of parameters. It would allow to make the model more memory-efficient and at the same time would be capable of performing the tasks of language understanding at the state-of-the-art. ALBERT is beneficial especially where large models like BERT cannot be used due to a reason of computation (Lan et al., 2020).

The tradeoff between large (BERT) and small (e.g., Distil BERT or ALBERT) models consists of having both performance and computational efficiency. BERT produces better quality text representations but is infeasible on computers in real time like in surgical phase recognition, low-latency processing is essential. Distil BERT and ALBERT can trade computational costs, and still maintain a good performance, hence it would be more appropriate to be used in the clinic.

The model choice, in image encoding, is of significance in the extraction and representation of visual features. In multimodal models, image encoder takes visual information (e.g., surgical video frames) and generates embedding which can be compared to textual information. There are some widely used popular models that are currently used in image recognition applications that include:

- ViT (Vision Transformer) is a transformer-based network that processes images as patches of fixed size sequences a model that is analogous to text modeling by transformers. ViT is superior to traditional CNN-based models because it is dependent on the worldwide features that are found throughout the image. This is of relevance to hard visual tasks that comprise of longer-range contextual data across images (e.g. surgical stage identification). Nevertheless, ViT requires massive data and computing power to achieve good performance, and it is not feasible to be implemented in the resource-constrained environment (Hu, Jia, & Rostami, 2024).

- Swin Transformer enhances the ViT architecture by incorporating a sliding window self-attention scheme that has more detailed information of local and global context. It allows it to extract gradually fine-grain and high-level features and achieve encouraging results in problems like object detection, image segmentation etc. (Liu et al., 2021). In SPR applications, the model derives complex features of surgical videos and, therefore, increases the accuracy of pathological staging. Its benefits are that it is computationally efficient compared to other models based on Transformer, such as ViT which is more feasible in clinical practice.

- Residual Network (ResNet) is a convolutional neural ResNet of skip connections that is famous about its ability to overcome the issue of the vanishing gradient in extremely deep networks. Such network facilitates ResNet that does well in feature extraction in the process of image recognition and more specifically in edge detection as well as texture analysis which is also a local feature. Although (ResNet) (Wang et al., 2025)demonstrates strong results without as many computational resources, the algorithm to acquire worldly image associations might be inefficient relative to Transformer-based models e.g., ViT or Swin Transformer.

The tradeoff between local and global context modeling is the reconciliation between CNN-based (ResNet-like) models and transformer-based (ViT or Swin-Transformer like) models. Convolutional neural networks are also more effective and are effective in the local extraction of features, whereas transformers are able to extract long-range correlations across the images - a property required in the understanding of complex scenes in surgical videos.

Fusion strategy will be employed in order to combine and integrate the nature of text modalities and Image modalities. The fusion approach determines the capability of textual-visual information to be combined and hence it influences the overall performance of the model. Ordinarily, the fusion mechanisms are:

One of the simplest methods of fusing is concatenation in which text and image embeddings are concatenated into a vector and processed in this form. Despite being computationally efficient, concatenation is unable to capture entirely the interaction between linguistic and visual features in a complex task as in estimating an operation stage.

- Additive fusion Embedding text and image additively, by addition of the element values, to add more balanced contributions of both modalities. This method is very computationally efficient, and it is especially effective when modalities are in some correspondence (Abiyev et al., 2024).

- Attention mechanisms enable models to attend to the most significant aspects of a text, image features and combine them. The models dynamically control the intensity of the various modalities by assigning weights of attention to the text sections or the image areas, thereby encouraging cross-modal alignment. The primary tools that are applied to match fine grained textual and visual features are self-attention and cross-modal attention (Abiyev et al., 2024).

- Gate based fusion eliminates the fixed association involving modalities and directly regulates the information flow of various sources by gate mechanism (learnable parameters) so models can make decisions on how each modality can contribute to the chart due to task requirements. This is better still when the data are noisy, i.e. the irrelevant information in either of the modalities can be removed in order to maximize the quality of alignment (Abiyev et al., 2024).

- In projection fusion, the text and image features are projected to a shared latent space, and they are fused. The assurance is that both modalities are projected onto the same semantic space that results in the fact that it is easier to align and reason in cross-modal scenario. Heterogeneous data (e.g. text and image) is one area where the power of projection-based fusion is especially effective since the various forms of data occupy the same representational space before fusion (Zhou et al., 2022).

Such fusion architectures shall be aimed at aiding the state-of-the-art through improved alignment and incorporation of text and of image data. Some light weight strategies of incorporating fusion and projection fusion may prove quite useful in applications, where the efficiency of computation and real-time performance are essential, such as surgical procedure in real-time recognition.

2.3 Review of Surgical Phase Detection

One of the most famous standard datasets to solve surgical phase recognition (SPR) tasks, especially during laparoscopic cholecystectomy, is called the Cholec80 dataset. The data set includes 80 videos of laparoscopic cholecystectomy with labels of 15 surgical procedures, such as, among others, Calot's Triangle Dissection, Gallbladder Extraction and Cleaning Gallbladder. Each video is frame-level annotated with information about the transition between these stages, thereby providing a detailed observation of the surgical process of work (Faray De Paiva et al., n.d.). The Cholec80 data also offers a number of challenges, although it is large and detailed. One of the primary impediments is the visual similarity of various phases. An example is that it is possible to have dissimilar surgeries that have similar visual appearances, like Calot's Triangle Dissection and Dissection, which are different in vision; thus, models cannot distinguish them based on their visual appearance only. Moreover, the dataset is highly imbalanced by the number of classes, with certain phases (e.g., cleaning the gallbladder) being under-represented and numerous biases against phase discovery. This is exactly what causes vision-only models not to be able to provide stable performance throughout all the stages, as phase transitions can be rather sensitive.

Table 2.1 several representative papers in the field of surgical phase recognition (SPR)

Paper	Model/Method	Dataset	Fusion Approach
ReSW-VL: Representation Learning for Surgical Workflow Analysis Using Vision-Language Mode (Kondo, 2025b).	CLIP-based, Vision-Language Model (VLM)	Cholec80	CLIP-based fusion (Vision + Textual Phase Descriptions)
Surgical Video Workflow Analysis via Visual-Language Learning (P. Li et al., 2024).	Cross-Attention Transformer	Cholec80	Cross-attention fusion (Vision + Text)
Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends (Y. Li et al., 2024).	CNN-based	Cholec80	Vision-only (No textual fusion)
HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition (Yuan, Srivastav, Navab, & Padoy, 2025).	Additive Fusion, Gated Fusion	Cholec80	Lightweight fusion mechanisms
ARST: Auto-Regressive Surgical Transformer for Phase Recognition from Laparoscopic Videos (Zou, Liu, Wang, Tao, & Zheng, 2022).	Transformer (Auto-regressive)	Cholec80	Vision-only with temporal fusion

Multimodal solutions have been suggested in recent years to overcome the limitations of vision-only models; around vision-language models (VLMs), it is of particular interest to consider vision-language models (VLMs), including CLIP. To solve this, such models combine textual descriptions of the phases of surgery with the visual data of the video since additional context enables a better classification of the phases. Switching between unimodal and multimodal fusion models is therefore an important pattern in SPR because one can learn the relationship between text and image data in a common embedding space, like in the case of CLIP. The contrastive learning, CLIP, positions the textual phase representation and the visual one in a unit space that allows zero-shot learning as well as improving the accuracy of phase detection without high-level task-specific fine-tuning. We show that this multimodal strategy is superior to vision-only models by showing that it overcomes numerous constraints of the latter, including the separation of visually similar phases and enhancement of phase transitions with textual information. The fact that these models incorporate lightweight fusion schemes also increases their efficiency; therefore, they can be potentially used in real-time to deploy them in clinics depending on surgical theatres. Other VLMs, including ALIGN and BLIP (J. Li et al., 2022), contributed to this paradigm shift because they showed the strength of using large-scale image-text pairs databases to identify phases in surgeries better.

The recent few papers in SPR using Cholec80 dataset have explored the various forms of deep learning models each with their own merits and also concentrate on the various facets of phase detection. The previous ones were largely based on CNN-LSTM structures, with CNNs having been applied to extract spatial related information in the discrete video frames and the LSTMs having been trained to capture time related information between the individual video frames. These models had worked well to identify local visual features but failed in long range dependencies and context adversaries, especially on challenging steps like the Triangle Dissection of Calot and Dissection.

The replacement of Vision Transformers (ViT) then the Swin Trans- formers (Liu et al., 2021) models of recent research studies, cross attention transformers are powerful replacement models that can be trained end-to-end to perform vision tasks by representing an input image as a sequence of patches, in which they can capture all global dependencies fully. Such transformer-based models are more adapted to learn intricate spatial relations and have worked well at detecting surgical phases in conjunction with textual phase descriptions. The dual-encoder model of CLIP (Contrastive Language-Image Pretraining)(Radford et al., 2021) where text data and image data are fed through an embedding space trained to make the projection is

one of the most influential models and as such the one which is able to support strong cross-modal reasoning. Zero-shot learning characteristics of CLIP is appropriate in surgical phase recognition due to the limited or unavailable pre-defined labels (Yuan, Srivastav, Yu, et al., 2025).

Several fusion strategies have been proposed that can be utilized to merge the visual and text features in the most optimal way. Additive fusion combines both text and image content on an element-by-element basis in such a way that both modalities equally take part in the final representation (Abiyev et al., 2024). It is a simple and light-weight technique but perhaps it cannot fully model complex text and image feature correlations. Gated fusion, in its turn, makes use of gates (learnable parameters) to control the flow of information that is tied to each modality. This approach is able to alter the contribution of every modality in an adaptive manner when fusing and only utilize suitable information. This is particularly useful in cases where the quality of data is bad or unclear (e.g. when changing during surgical phases).

The other strategy is to use projection fusion, in which the text features and image features are projected to a shared latent space, and then the fusion is done. This may stimulate an alignment of the two modalities within the same semantic space to do cross-modal reasoning. The fusion strategies are critical in models so that they can be capable of integrating both the textual descriptions of phases and pictorial-based features to enhance better phase classification (Zhou et al., 2022).

The evaluation of models is mostly followed by a set of evaluation criteria. Popular measures include top-k accuracy that determines whether the model can in its k best samples rank the correct phase. More specifically, this is true when the model must specify a phase out of a series of potential phases. Other meaningful measures include precision, recall and F1-score that give a more complete view of how the model performs in phase identification and also balancing between false positives and false negatives. Precision and recall can be particularly useful to test a model on phases that are underrepresented, which is challenging to learn due to the skewed nature of the data. Lastly, confusion matrices do well to find misclassifications and know what sort of the model is amiss with (e.g. a mix of similar phases or a lack of phase transitions). These metrics of evaluation may be employed to compare the performance of the models of interest and provide information on how the model can be improved in future particularly in the case of the imbalanced data like Cholec80 (Yuan, Srivastav, Yu, et al., 2025).

2.4 Limitations of Existing Approaches and Research Focus

Deployment of Large Models in Real Time. They are based on using large computationally intensive models like cross attention transformers and multi-level encoders. Large memory and long latency are typical of such models and are not quite relevant in a real-time system of clinical practice. Surgery demands real-time feedback, but resource-intensive models have had difficulties in the provision of low-latency decision support in surgery. Thus, researchers require the demand of lightweight models with precise detection of the stage.

Data Imbalance

The possible imbalance of the depicted skill data has been the primary issue in the SPR investigation. Data In the Cholec80 dataset, certain actions such as cleaning adhesions have a very small number of samples as compared to other more common steps. Imbalance in the dataset causes bias in the model: models are effective on the fairly rich training stages but are less effective at identifying fresh ones. The consideration of imbalanced data was not an issue in previous studies, and the possibility of data augmentation to equal representation and detectability of all stages was ignored. These researchers might want to employ more intricate methods (e.g., SMOTE: synthetic minority oversampling technique) or stratified sampling in future work in order to handle the problem (Abiyev et al., 2024).

Absence of Domain-Specific Pre-Training.

The other shortcoming of the past models is the inability to pre-train domain-specifically, in particular, medical text and surgical images. Most systems, like CLIP, tend to be generally trained on some sort of generic sample and thus might not be able to remember medical words or even surgical practices adequately. Surgical images are characterized by a different nature than overall image recognition, and the medical vocabulary is not similar to the everyday language. Thus, general-purpose dataset-trained models might not be successful in the identification of the dynamics of surgical intent or switchable transition during laparoscopic cholecystectomy and thereby undermine performance in SPR tasks (Yuan, Srivastav, Yu, et al., 2025).

Specific Research Focus

Fusion Head Research Focus: The premise underpinning this paper is that different combinatorial fusion head methods will probably help increase the text-image matching through to surgical phase recognition. It is particularly based upon four lightweight fusion plans:

- **Additive Fusion:** It is a summative method of adding up the embeddings of text and image pair.
- **Weighted Fusion:** Trains visualization of texts.
- **Gated Fusion:** Modalities learn and explicitly regulate their information exchange by a learning process.
- **Projection-Shared Fusion:** It executes the process of projecting the text and image components to shared latent space so that fusion effect along with the maximum alignment occurs.

In each fusion operation, the impact of the fusion operation on both cross-modal and computational complexity will be measured. The efficiency of these procedures in enhancing the classification of stages and real-time application of surgery stage recognitions will be established through this project.

Due to the fact that fusion heads based on CLIP are optimal in SPR in real time, the following is the reason why:

The most suitable ones to be used to obtain surgical phase recognition by mobile devices during the online mode are CLIP-based models, which have a dual-encoder architecture at their input stage and their ability to learn zero shots. So, in this way the model has a chance to balance the visual and textual ones in an effective way with CLIP and lightweight fusion heads. New as opposed to the bulky models, which are difficult to calculate and latency-free.

2.5 Summary and Literature Gap

The literature displays that multiform learning in surgical phase recognition (SPR) is well-grown as well as it has also been demonstrated that it performs well and, in this case, the integration of visual and textual source can enhance the accuracy of the classification significantly as compared to vision solely approach. Convolutional and recurrent neural network-based architectures and transformer-based temporal models and vision-language models (VLMs) such as CLIP have been shown to minimize ambiguity in the visual space and better understand the context of the surgical procedure. Nevertheless, despite the success of such approaches, various critical limitations are still not solved, and which involve data imbalance, computational efficiency and the ability to implement such approaches in practice in time-critical clinical environments using limited annotated data and limited hardware.

Another gap in research is that there is limited exploration of lightweight multimodal fusion systems in CLLIP-based systems in the direction of surgical phase recognition. Despite the high level of correspondence of CLIP and other VLMs to the text and image representation through the large scale of pretraining, they are built on computationally costly fusion or alignment techniques, making their deployment to cost sensitive healthcare systems like surgery impossible. The existing methods largely concentrate on the effective cross-attention transformers or domain-specific large models and often overlook the power of simple fusion designs. Consequently, a systematical study of lightweight fusion mechanisms capable of effectively combining the visual and textual information with strong generalization on a small amount of data is immediately needed; such mechanisms also help to learn the surgical phase in real time without an additional cost in computation.

CHAPTER 3

METHODOLOGY

3.1 Overview

Here we outline the methodology of our work and explain the experimental conditions on which the lightweight multimodal fusion strategies of SP R will be conducted with the Cholec80 dataset. The proposed method is designed to give the solutions to the research questions and fulfill the objectives in a logical way by integrating multi-mode visual-textual information on the basis of CLIP-like environment and test the various lightweight fusion mechanisms under low-data regime.

Research Objectives Review

This research has four objectives which formed the methodological design of this project:

- To create and test a modularized CLIP-based surgical phase recognition pipeline with replaceable lightweight fusion heads, e.g. additive fusion, weighted-sum fusion, gated fusion and shared-projection fusion.

- To project visual and textual data to a common latent feature in light linear maps and normalized, to enable meaningful multimodal integration and enhance phase classification resiliency.

- To examine stage wise diagnosis to measure behavior of the model on various phases of surgery and to determine the most typical misclassification patterns particularly in visually similar and transitional stages.

- Another line of research would be plans to attain incremental video-based identification and in-site retraining to sustain model changes with the accrued surgical data, which can be deployed over the long run without the necessity of complete retraining.

Methodological Framework

The suggested approach applies a dual-eca CLIP architecture as the underlying multimodal representation backbone. The visual encoder receives surgical video frames then the text encoder which encodes the phase descriptions. Then it combines these representations using lightweight fusion mechanisms at the embedding level. Method The researchers have compared the additive fusion, weighted-sum fusion, gated fusion and shared-projection fusion systematically in the parameters of recognition accuracy, stability and computational efficiency.

Real-time surgical phase recognition is highlighted, and it is based on a focus in fusion designs to maintain a low number of parameters and training and inference time required. Informal Results (Top 1 most popular topic Do they make any difference? Each and every method of fusion is experimented under identical experimental environment to establish a fair comparison. In this way, the research will attempt to identify the fusion techniques that can create a perfect compromise between performance and deploy ability, thus rendering multimodal SPR systems viable in the clinical settings in low resources.

Process Overview:

The procedure will include control steps in the following manner:

- Cholec80 dataset preparation Preprocess the Cholec80, annotated per-frame and phase-specific. Balance the unbalanced address using a stratified sample or any other balance.

- A slim fusion model based on CLIP and a combination of visuals of surgical frames and the textual description of the respective surgical phases will be fine-tuned.

- The training approach will be trained on supervised learning by following the conventional methods, and the performance will be tested with precision, recall, F1 score and R@k. Model generalization over varying partitions of data can be acquired by cross validation.

- Some of the key performance metrics will be used to evaluate the performance of the trained model, i.e., the accuracy of the classification at each stage, the confusion matrices and the latency of the inference in real time. The diagnostic analysis is expected to help in establishing misclassification at the level of the stage and variation in the performance of the various fusion procedures under various conditions.

This approach allows the quantitative comparison of the choice of fusion, which gives a complete picture of how this affects the determination of the surgical stage. It aims at improving the clinical usefulness of viable real-time models.

3.2 Dataset Description

The Cholec80 dataset, which is described in the study of this paper, is described in this section. It is on this dataset that we develop surgical phase recognition (SPR) of laparoscopic cholecystectomy. It is important to understand the characteristics and challenges of the dataset, as well as the motivation for using the dataset for this work.

Data Set Origin:

Public availability Cholec80 data is publicly availed and the outcome of the Cholec80 research team in collaboration with the University Hospital of Strasbourg/IRCAD (Strasbourg, France). It consists of 80 videos of laparoscopic cholecystectomy surgeries with 7 phases being marked. The tools of the dataset are also annotated, indicating that seven surgical tools were used by the operations. It is also published freely under the Creative Commons (CC-BY-NC-SA 4.0) license, which allows using the data in the non-commercial process and at the same time, it is possible to transform it or republish it, but with the references to the original authors.

Dataset Characteristics:

Number of Videos: The proposed dataset comprises of 80 Full length laparoscopic cholecystectomy videos. These videos are indispensable for training SPR models as they have complete narration of all the phases of the surgery.

Number of Phases The dataset Cholec80 has been annotated with 7 phases of surgery:

- Preparation
- Calot's Triangle Dissection
- Clipping and Cutting
- Gallbladder Dissection
- Gallbladder Packaging
- Cleaning and Coagulation
- Gallbladder Retraction

These stages are included into the essential operative laparoscopic cholecystectomy; therefore this dataset is a good candidate for surgical stage recognition experiments.

- **Instrument Annotation:** Annotations for instruments are also given at 1 fps indicating whether seven instruments (grasper, bipolar forceps, finger retractor, scissors and monopolar curvature) exist. Such annotation can provide additional hints for models of stage identification, in particularly the association of instruments with surgical phases.
- **Video Length:** The average length of Cholec80 dataset videos is 38+-16 minutes. The length of the video poses a challenge to the stage detection models to retain their accuracy on long video sequences and to process the temporal behavior of surgery.

- **Number of Frames:** Extracted 25 frames of videos per second, with approximately 1000 frames per video. The surgical phase is offered per each frame. The proposed frame extraction technique enables a detailed investigation and provides frame-level annotated data to train a model.
- **Phase Durations:** Table 1 below lists each of the seven surgical phases in the Cholec80 dataset along with their average durations across all 80 videos. The values are presented in seconds with standard deviations.

Table 3.1 Phase Durations

Phase	**Duration (s)**
Preparation	125 ± 95
Calot's Triangle Dissection	954 ± 538
Clipping and cutting	168 ± 152
Gallbladder Dissection	857 ± 551
Gallbladder Packaging	98 ± 53
Cleaning and Coagulation	178 ± 166
Gallbladder Retraction	83 ± 56

These times of phase show how much longer some phases are than others, some phases such as the Triangle Dissection of Calot and the dissection of the gallbladder being much longer than others such as Gallbladder Retraction. This time variation should be considered during construction of models to identify the phase in real time.

Challenges:

The Cholec80 dataset is not an exception as it also faces several challenges, which make it a difficult yet still useful tool in developing SPR models:

- **Imbalanced Phases:** Phases in surgery of Cholec80 are not as represented as others. As an example, cleaning and coagulation stage and gallbladder retraction may have smaller numbers than other more popular stages like the Calot Triangle Dissection, Gallbladder dissection. This causes the imbalance in classes making the models inclined more in identifying more common phases and less common phases are identified by the model.
- **Motion Artifacts:** Laparoscopy surgery has been prone to motion artifact due to the movements of camera and movements used by the surgeon. The phase detection can also be complicated by these artifacts that introduce noise to the visual channel that is difficult to classify phases based on the video frames when applying the models.
- **Tool Occlusions:** It is worth mentioning that in laparoscopy surgery, surgical tools could also block sections of the field of view that is vital in the specified operation. It can cause the neglect of visual information, potentially causing the model to be unable to determine some of the stages in a correct way.
- **Visual Similes:** It is due to the fact that a lot of steps in Cholec80 are visually similar, e.g. Triangle Dissection and Dissection, where the instruments and the environment around them are visually similar. This is why it is hard to draw a line between the phases using the visual stimulus only, and it is required to resort to the use of the text that will assist in the phase's classification.

Why this Dataset:

The Cholec80 data is particularly appropriate in this research because the data is a large and well-labeled collection of laparoscopic surgical videos, and thus an ideal data to evaluate the method of surgical phases recognition. The experimental settings of multimodal fusion strategies can be made under realistic settings because of the dataset heterogeneity in regard to the surgical stages, interactions of the tools, and visual challenges. In addition, it possesses frame-level annotations, thereby making it an ideal data to be used to train deep learning models, particularly in cases where the objective is the fusion of multimodal as in the case of CLIP based model in this paper.This research paper aims at exploring how the issue of data imbalance, visual similarity, and motion artifact in real time surgical phase detection can be addressed through dissimilar lightweight fusion techniques.

3.3 Preprocessing and Data Augmentation

Video Preprocessing

- Each laparoscopic cholecystectomy video in the Cholec80 dataset is sampled to 1 fps, and the video frames are then extracted using the aid of video.
- It is the reason why they become the most significant stages, which are covered without overloading the model with irrelevant data.
- Frames extracted are trimmed down to 224x224 (CLIP model size).
- The frames are then standardized to remove the meaning and then divided by the standard deviation of the values of the pixels that stabilizes the training process.

Text Preprocessing

- The CLIP tokenizer is used for phrased phase descriptions tokenization (Ivor-Lewis).
- The tokenized text is stretched out or cut to a certain length (in our case we cut off the end) in our example, 32 tokens.

Frames are stored in the form of .pt files (PyTorch tensors) for a number of reasons:

- Reduction of frames to tensors makes sure that the I/O overhead is minimized and makes the training data faster to access. It does not require recurring preprocessing either as frames are already readily available.
- pt); therefore, the preprocessed frames can be easily combined into the training pipeline without any preprocessing.
- Tensors are less memory-consuming as compared to the image files, and this implies that you can store the visual information in totality in full visual quality without necessarily filling up your drive.
- PT files can be used to introduce a certain amount of randomness during the pre-processing of the image, e.g., random cropping, flipping and color jittering, as well as enhancing generalization of the model.

Data Augmentation

- In order to overcome the data imbalance problem and make the data more robust:
- Coping with Class Imbalance: Under-represented phases are oversampled to bring about the equilibrium of all classes when training.

- Random Cropping: It adds spatial invariance, like the model can be tolerant to changes of viewpoint and the location of tools.
- Flipping: Horizontal flipping is done in random fashion; this can be performed during training in order to mimic random camera orientation as well as tool side that would assist the network to generalize better between different conditions in the surgery.
- Other Augmentation Rotation, jittering of color, zoom-in, etc., are other types used to assist the model to operate in different conditions of lighting and camera angle in surgery.

3.4 Model Architecture

The following section defines the model of surgical phase recognition, with particular focus on the CLIP based multimodal pipeline. It has a paragraph that reports the vision and text encodings, the fusion process of combining features of both modalities and the loss function contemplated during training.

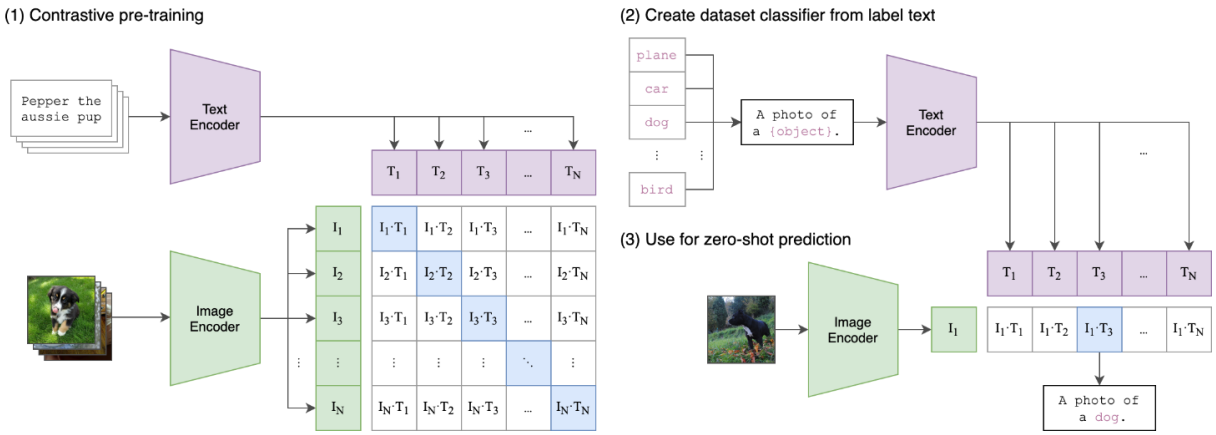

Figure 3.1 CLIP Architecture

CLIP Model:

CLIP (Contrastive Language-Image Pretraining) is a type of model which also learns to comprehend text captions and images through optimizing a common embedding space. Visual and textual information are processed by a dual encoder architecture with each processing information independently and then aligning them in a common space to perform further tasks such as phase recognition.

Vision Encoder:

The vision encoder of CLIP takes as its input frames of laparoscopic surgery videos, and the engine is a Vision Transformer (ViT). It is the task of the encoder to capture elevated-level visual characteristics that appear to depict surgical contentedness, like the movement of tools, deformation of the anatomy, and the interactions during the different surgical stages. The principal processes of the vision encoder are as follows:

Image Processing:

Let **I** be an input image (the frame of a surgical video). The picture is divided into disjointed patches, and every patch is linearly projected to a fixed dimensional representation:

Where:

$$\mathbf{X}_{\text{patches}} = \text{Flatten}(\mathbf{I}) \cdot \mathbf{W}_{\text{patch}} + \mathbf{b}_{\text{patch}}$$

Where:

$\mathbf{W}_{\text{patch}}$ is the trained projector matrix

$\mathbf{b}_{\text{patch}}$ is the learned bias.

Self-Attention:

It uses a multi-head self-attention mechanism to learn the long-range interdependence of patches across the entire world:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Where:

Q is a query matrix (patches of the projected image).

K is the matrix of the key (projection of image patches).

V is the image patches, and V is the reconstructed image patches.

$d_k$is the query and key vectors dimensionality.

Final Visual Features:

The final result of the vision encoder $\mathbf{z}_v$is a fixed-size image of the image following a number of layers of transformers. This property is applied in the fusion process.

$$\mathbf{z}_v = \text{ViT}(\mathbf{I})$$

Text Encoder:

The textual descriptions of the surgery stages (“Calot's Triangle Dissection" and "Gallbladder Removal") are transformed as well by a Text Encoder and the procedure used is also based on a transformer architecture. It uses these descriptions to generate semantically equivalent embeddings to the visual features that have been generated by the vision encoder.

Text Tokenization:

The textual input **T**is tokenized into a sequence of tokens using a tokenizer.

$$\mathbf{T}_{\text{tokens}} = \text{Tokenizer}(\mathbf{T})$$

Text Embedding:

This is a computation of the representation of each token in the input sequence as a high-dimensional vector. The token embeddings are extracted by another layer called a transformer layer that captures the context of tokens.

$$\mathbf{z}_t = \text{Transformer}(\mathbf{T}_{\text{tokens}})$$

The output $\mathbf{z}_t$is the final text embedding that captures the meaning of the phase description.

Contextual Embeddings:

The tokens are contextually embedded with the help of the transformer. Such embeddings are then summed up (e.g. " mean pooling " or special token [CLS] embedding) to a single embedding which summarizes every entire phase description.

$$\mathbf{z}_t = \text{Pool}(\mathbf{z}_t)$$

Fusion Mechanisms

The model makes a comparison of different light-weighted fusion approaches to vision and text embeddings alignment. These hybrid mechanisms aim at the improved alignment and classification performance at the same time cost of computation is minimized. The fusion strategies that were explored in this thesis are: Fusion Strategy is selected randomly.

Additive Fusion:

Under the latter approach the visual and textual embeddings are mapped to the same space of vectors and the two added. The resulting embeddings are normalized using Layer Norm. This simplistic method ensures the image and text representation make equal contributions to the resultant feature representation.

$$\text{Fused} = \text{LayerNorm}(\mathbf{z}_v + \mathbf{z}_t)$$

Where $\mathbf{z}_v$and $\mathbf{z}_t$ the vision and text inscriptions.

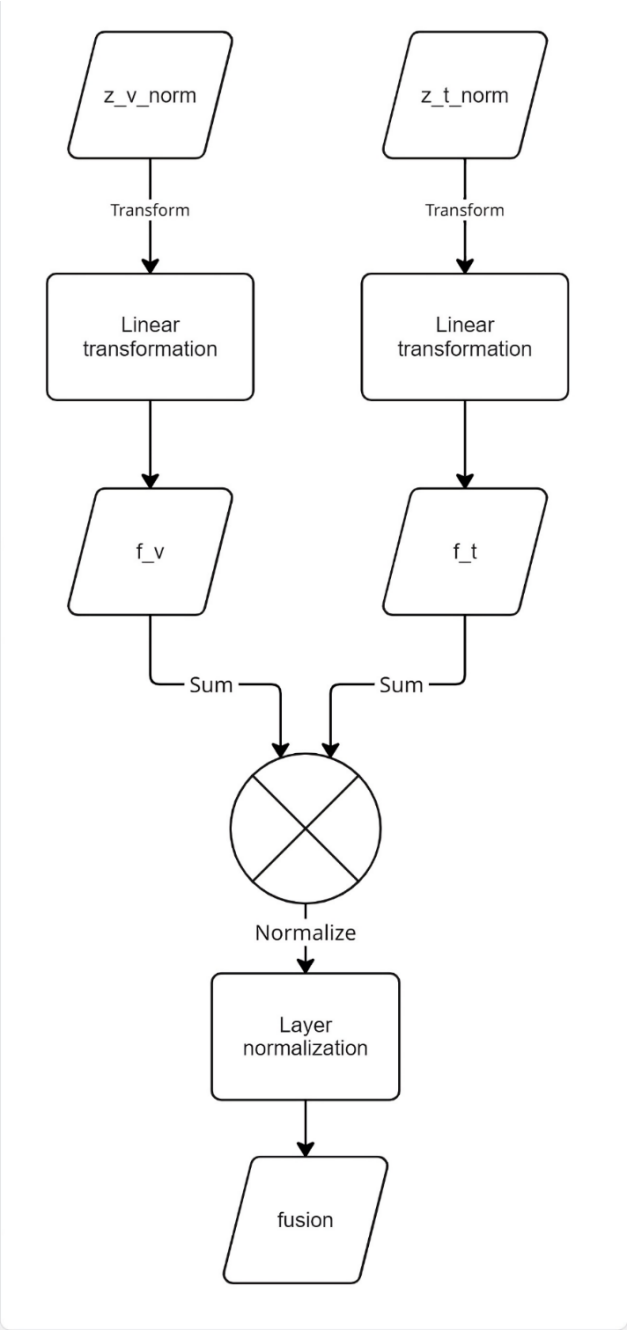

Figure 3.2 Additive Fusion

Gated Fusion:

It is a hybrid method, with a gating mechanism to place dynamic weights of vision and text embeddings. The gate is trained (through training) by allowing the model to attend more to the modality which has the most information regarding a certain task. The gated fusion mechanism is used in aggregating the weighted embeddings to form the final representation.

$$\text{Fused} = \text{LayerNorm}(\mathbf{g} \odot \mathbf{z}_v + (1 - \mathbf{g}) \odot \mathbf{z}_t)$$

Where $\mathbf{g} = \sigma(W_g[\mathbf{z}_v, \mathbf{z}_t])$ gating function that learns to weight the visual representation and textual representation.

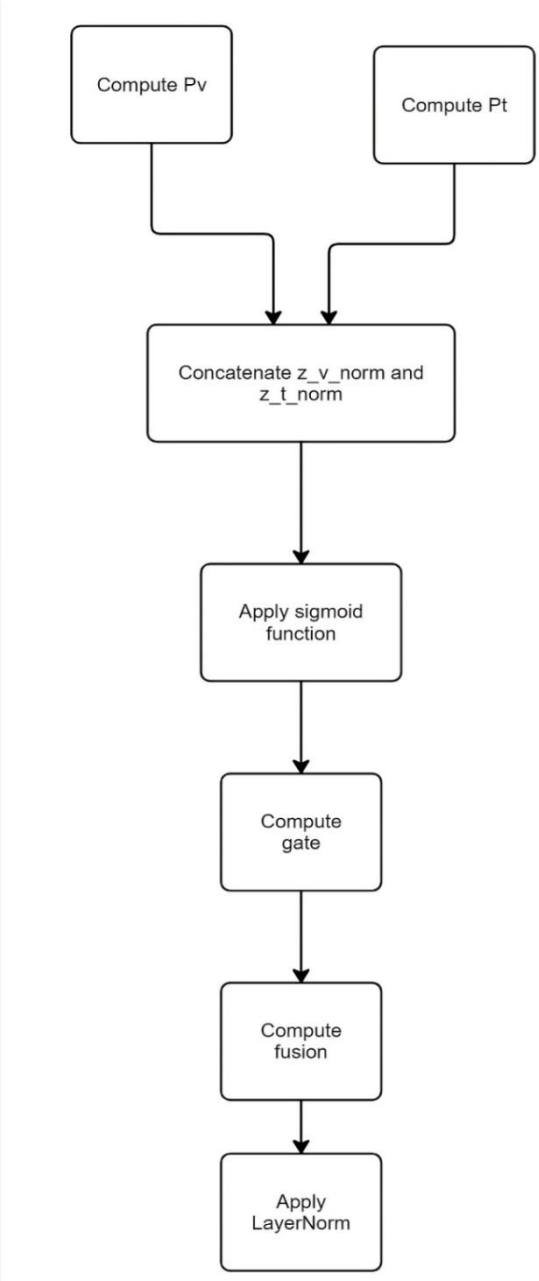

Figure 3.3 Gated Fusion

Concatenation-MLP Fusion:

In this scheme, concatenation of visual and textual embeddings is fed into an MLP. The MLP will learn interaction of the two modalities. whereas we will learn to get long range correlation out of the vision & text features.

$$\text{Fused} = \text{MLP}([\mathbf{z}_v; \mathbf{z}_t])$$

Here $[\mathbf{z}_v; \mathbf{z}_t]$ is the concatenation of vision and embeddings of the text.

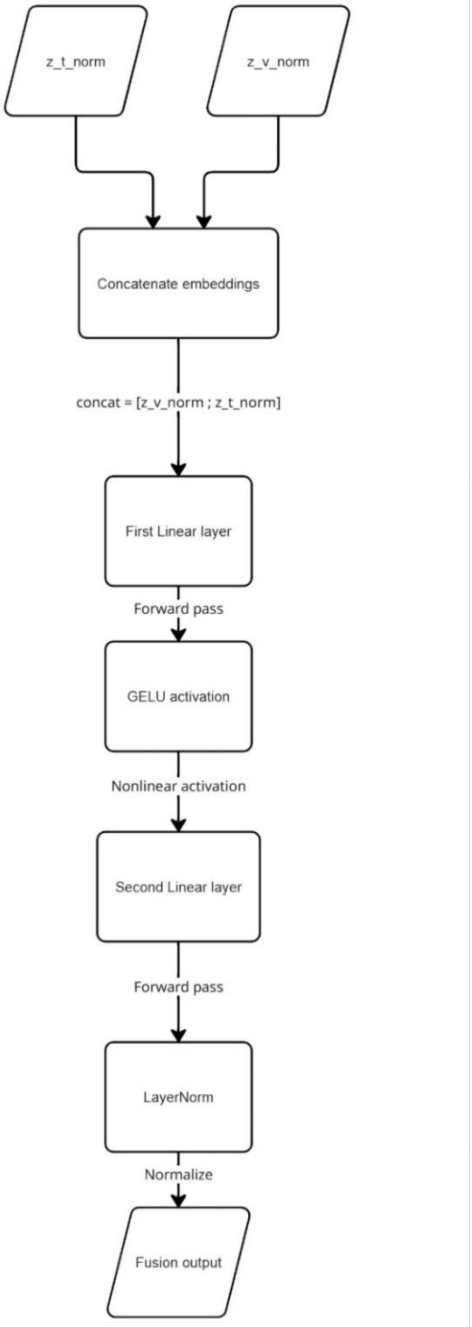

Figure 3.4 Concatenation-MLP Fusion

Projection-Shared Fusion:

Projection share fusion There is a shared latent space to which independent linear projection layers are mapped for pre-projection of the image text embeddings, followed by multiplying the two modalities point-wise thereafter.

$$\text{Fused} = \text{LayerNorm}(\mathbf{W}_v \mathbf{z}_v \odot \mathbf{W}_t \mathbf{z}_t)$$

Where $\mathbf{W}_v$and $\mathbf{W}_t$are the projection matrices for vision and text embeddings, and $\odot$denotes element-wise multiplication.

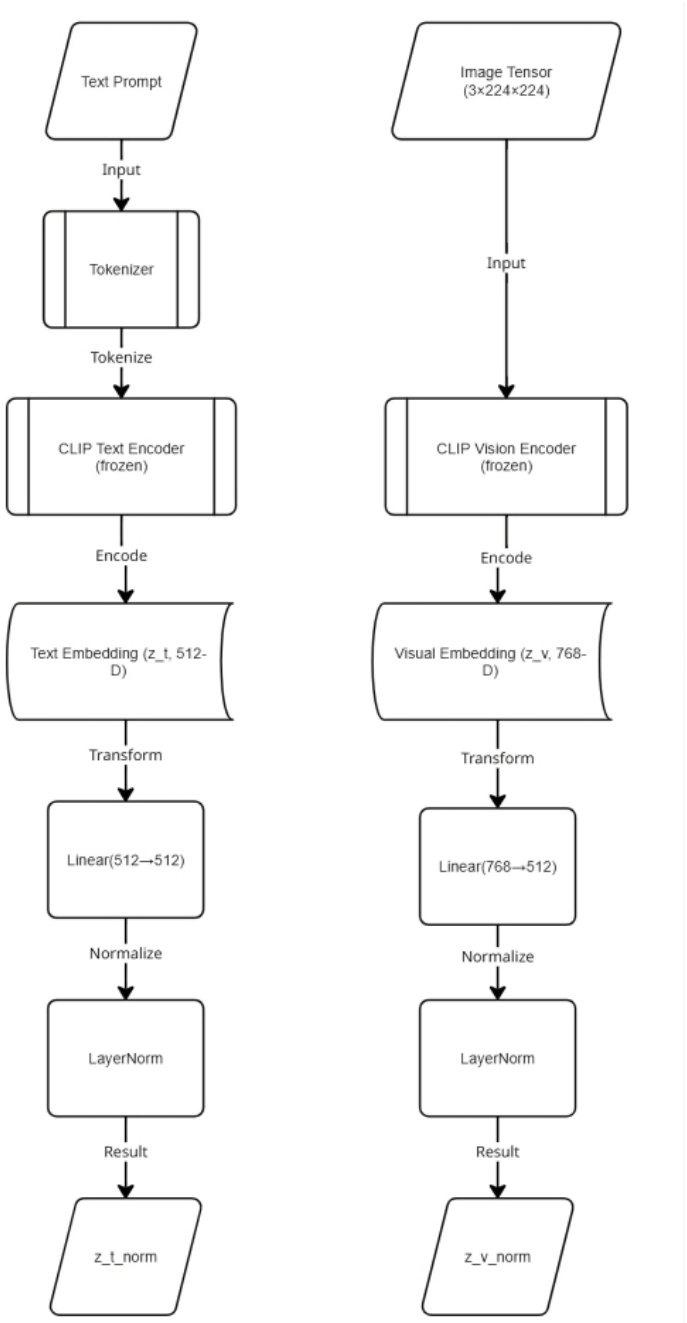

Figure 3.5 Projection-Shared Fusion

Weighted-Sum Fusion:

Although this procedure maps embedded both visual and text based into the same dimensional space. Then a weighted sum is computed depending on weights learned in the training The fusion is learned by a parameter, beta, that controls the degree to which every modality within the modalities contributes to generate the final representation.

$$\text{Fused} = \text{LayerNorm}(\alpha_v \odot \mathbf{W}_v \mathbf{z}_v + (1 - \alpha_v) \odot \mathbf{W}_t \mathbf{z}_t)$$

Here $\alpha_v$is a learnable trade-off parameter between the vision and text.

Projection Layer

One building block in this structure is projection layer which addresses the visual- and textual features into same latent space prior to fusion. The above processes take place in two steps:

Linear Projection:

The vision and text embeddings are processed separately, i.e. using linear layers. These transformations are such that the both embeddings of modalities have the same dimensionality, for what it's worth in terms of being able to fuse them. The projection is responsible in order to allow the model to handle different feature spaces produced by each encoder.

$$\mathbf{z}_v = \text{Linear}(\mathbf{z}_v), \mathbf{z}_t = \text{Linear}(\mathbf{z}_t)$$

Where $\mathbf{z}_v$and $\mathbf{z}_t$ the embeddings of vision and text.

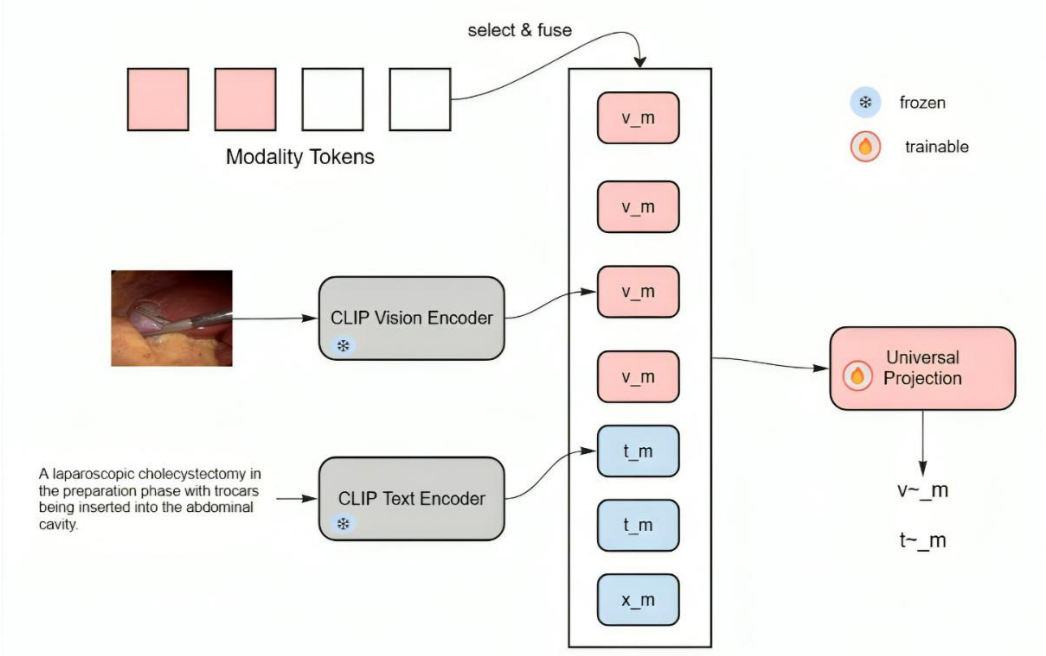

Figure 3.6 Linear Projection

Layer Normalization:

Layer Norm is used after projection embeddings in order to scale their features. This helps in stabilizing the training process as it normalizes the activations such that the model would be learning representations that are effective from both modalities.

$$\mathbf{z}_v, \mathbf{z}_t = \text{LayerNorm}(\mathbf{z}_v), \text{LayerNorm}(\mathbf{z}_t)$$

Shared Latent Space:

From each extracted feature, and for each feature modality (vision and text), we produce the corresponding embeddings with which they are then projected to a common shared latent space where they can be better fused and aligned. This is required to have some high-dimensional space for both the modalities for better end model performance.

$$\mathbf{z}_f = \mathbf{W}_v \mathbf{z}_v \odot \mathbf{W}_t \mathbf{z}_t$$

Where $\mathbf{z}_f$ is the joint embedding of the space of the joint embedding of the shared latent space.

Frozen Encoders

The vision and text encoders are frozen when - which helps to accelerate computation Which implies that the weights of the encoders are not changed during the training stage and only the fusion head and the Projection layers are optimized. Freezing of the encoders is good for several reasons:

- **Efficiency:** Freezing the pre-trained/encoders of CLIP helps to ease the computation of training. As the encoders of CLIP have been pre-trained on a large-scale dataset, they directly generate fine-grained features, which are customized on the surgical phase recognition. The frozen encoders help the model to focus on how to learn the best fusion strategy for surgical phase recognition task at hand.
- **Utilizing Pretrained Knowledge:** The encoders are pre-trained on large datasets and have learnt rich feature representations. It keeps them frozen so we can still enjoy the benefit of this pre-trained knowledge without having to retrain the humongous models from scratch.
- **Decreased Complexity of Training:** Freezing the encoders makes training easy since this is only the fusion mechanism that needs to be learned by model. This is more efficient when we are in the training phase but helps to keep overfitting under control.

Loss Function

It is learned for a contrastive loss function (e.g. infoNCE loss or CLIP loss) during the training. This loss allows models to learn this visual and text embeddings alignment as it tries to maximize the similarity between pairs of embeddings that go together and minimizes the similarity between the non-corresponding embedding pairs. The effect of contrastive loss is to ensure that for each image-text-pair in the data set, while the true matching is closer to each other in latent space than individual wrong matchings.

CLIP Loss:

The contrastive objective in CLIP seeks to maximize the similarity of positive pairs (i.e. correct image-text pairs) and minimize the similarity for negative pairs (i.e. incorrect image-text pairs). This supports the learning of meaningful correspondences between features of text and image that is critical for successful multimodal alignment.

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp\left(\mathbf{z}_v \cdot \mathbf{z}_t/\tau\right)}{\sum_{i=1}^{N} \exp\left(\mathbf{z}_v \cdot \mathbf{z}_i/\tau\right)}$$

Here $\mathbf{z}_v$and $\mathbf{z}_t$ the embeddings of vision and text, respectively and to scale the similarity since is a learnable temperature parameter.

Training and Inference

In the training process, the encoder of the vision and text are frozen, and we only optimize the fusion head (including projection layer). The model has to learn to project both modalities (vision and text) into one common Latent space, in which it can recognize and classify the phases of surgery with low error rates.

At test time, the model uses the learned fusion mechanism to make the pairs of images and text consistent and obtain a surgical phase label for each frame of video. The fusion mechanism ensures that the two modalities are considered during the classification process which is good for the model in order to make good predictions based on the synchronized multimodal features.

3.5 Training Setup

Loss Function:

The method is trained using a contrastive loss objective which plays an important role to align multimodal features from text and image encoder. In particular, the InfoNCE loss is used which computes a similarity between normalized image and text embeddings.

The formula used for contrastive loss is:

$$\mathcal{L}_{contrastive} = -\log\left(\frac{\exp\left(\text{sim}(z_v, z_t)/\tau\right)}{\sum_{i=1}^{N} \exp\left(\text{sim}(z_v, z_i)/\tau\right)}\right)$$

Where:

- $\text{sim}(z_v, z_t)$is the cosine similarity between the image (z_v) and text (z_t) embeddings.
- τ is a trainable temperature parameter which weighs the similarity score.
- N is the number of samples in the batch; and the sum in denominator numerates all the (image, text) pairs.
- This loss function ensures that matching image-text pairs are close in the shared embedding space while non-matching pairs are pushed apart.

Optimizer:

The model is trained by the AdamW optimizer which is the improved version of the Adam optimizer having a weight decay regularization. This decision makes the efficient optimization of large models easy. The parameters are tuned into as follows:

- **Learning Rate**: 3×10^{-4}
- **Weight Decay**: 1×10^{-4}

These values also contribute towards the regularization of the learning process and in the overfitting issues a problem when handling large-scale data.

Learning Rate and Weight Decay:

We choose a learning of 3×10^{-4}, which is a normal choice for transformer models and give smooth convergences. Furthermore, a weight decay is set to 1×10^{-4} in order to prevent overfitting, by punishing large weights.

Abandoning explicit learning rate scheduling in this environment, because the optimizer is able to take advantage of the native control of the learning rate decay in AdamW.

Epochs and Batch Size:

- **Batch Size**: 8
- **Epochs**: 3

The batch size is set to be able to comfortably fit in the memory of the GPU; and to achieve a trade-off between efficiency of the training and stability of the convergence. The epochs parameter is 3 which provides enough time for the model to train, without causing the model to overfit within the relatively small dataset (Cholec80).

Gradient Clipping:

Gradient clipping is done to prevent exploding gradients to backpropagating. In particular\we use Max Norm 1 (clip by their L2), this attractive feature is well-performing and meaningful when applying to larger number of tasks.

Checkpointing and Incremental Training:

A checkpointing strategy is adapted to support the incremental retraining strategy and a model recovery strategy in long training processes. The model checkpoint where the best model with regard to evaluation metrics is serialized at the end of every epoch such that it can be retrieved later for use. Here is the check point:

- For every epoch the model saves its weight, so in the end, once training is finished, we take the last, the best model as best fusion model. pt.
- Retraining or continuation is possible in which case the training process can be continued from the last checkpoint saved and not start from scratch, optimizing resource utilization.
- In the event of new data or additional fine-tuning, models could be retrained in incremental fashion using the existing checkpoints enabling long term improvement of the model as more data is obtained.

This is extremely helpful for the use of this model in a real-time surgical situation, because retraining the model is required in many cases when more video data is gathered or when new steps/phase/anomaly is present.

Hardware:

The proposed model has been trained on an Nvidia Geoforce RTX 3050 Ti Laptop GPU which has been sufficient for the management of the large amount of parameters of CLIP model, and training in an efficient way the multimodal fusing mechanism. This configuration of GPUs makes training reasonably fast, and the accuracy of the model over many training iterations is guaranteed.

3.6 Evaluation Metrics

This section describes the main evaluation criteria followed to measure the model performance in terms of SPR. These measures are useful in interpreting how/good the model is in ordering the right surgical phase and also managing the class imbalance and direction for improvements.

Content:

Top-k Accuracy (R@k):

Top-k accuracy assesses the quality of the model's ranking of the given instances which belong to the positive surgery phase or negative surgery phases. In our case we evaluate Top-1 (R@1), Top 5 (R@5) and Top 10 (R@10) accuracy, which is the frequency of the true label in the top 1, 5 or 10 vibrational phases predicted.

The Top-k Accuracy can be defined as:

$$\text{R@k} = \frac{1}{N}\sum_{i=1}^{N} \quad \mathbb{I}(y_i \in \hat{y}_i^{(1:k)})$$

Where:

- $N$is the total number of test samples,
- y_i is the appropriate label of sample i,
- $\hat{y}_i^{(1:k)}$is the top-k predicted labels for sample i,
- $\mathbb{I}(\cdot)$ is the indicator function 'if away from home wins' is in the top-k which is 1 if the true label is of the top-k predictions and 0 otherwise.
- **Top 1 Accuracy (R@1)**: Measures if the top predicted phase is the correct one.
- **Top 5 Accuracy (R@5)**: Measures if the correct phase is within the top 5 predictions.
- **Top 10 Accuracy (R@10)**: There is information of predicting the correct phase on top of the top 10 prediction.
- The two-evaluation metrics were used as a way for the model's capability of ranking the right phase in surgical phase recognition scenario, with ambiguousness on distinctions between phases. **Precision, Recall, and F1-score:**
- Precision is the ratio TP/ all predicted positives (TP + FP) in which FP is false positive. It can be formulated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- Recall is the percentage of true number of positive pixels (TP) appear in all true and false positive truth pixels (TP + FN and FN, false negative). It is given by:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- F1-score is the harmonic means of precision and recall so as to obtain a good balance between them:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

They are used for both macro-averages (in which performance is averaged over a set of classes) and per-phase metrics.

Per-phase Precision, Recall, and F1-score:

The per-phase metrics enable analysis of the performance of the model for each single phase, in surgery. By comparing precision, recall F1-score per phase we can add statistics for following point analysis which allows assessing models strengths/weaknesses on specific phases especially for those which are underrepresented in the dataset (rare or transitional phases). This would be used to determine if one, or more, phase is being misidentified more often.

Confusion Matrix:

The confusion matrix is a visualization of the model performance with the number of correct and incorrect predictions for each stage. It provides the possibility to look in detail at, for example, the misclassifications and helps in identifying patterns of phase confusion. The matrix is also helpful in learning what are the phases that spend most time to be confused with other and help model improvements.

The confusion matrix $C$is defined as:

$$C_{ij} = \text{Number of samples of class } i \text{ that were classified as class } j$$

Where:

- C_{ij}represents the count of samples where the true label was $i$and the predicted label was j.

The first row from the confusion matrix help us to get honesty, truthfulness, recall and F1-score of every phase accuracy.

3.7 Experimental Setup

Training Procedure:

The training process of the CLIP-based surgical phase recognition model can be summarized as the following:

- **Splits:** data set is divided into 70% train 15% validation and 15% test. This creates an assurance that there is enough data to train on, but separate information sets to validate and ultimately test in order to answer the question: how often is the model generalizing? The splits are on video level (based on video IDs) and the training, validation and testing do not overlap.
- **Random Seed for Reproducibility:** Here one specific random seed is used for all the non-deterministic parts of training (e.g. dataset splitting, data shuffling and model initialization).
- **Training Process**: It first trains the model with a pre-defined number of epochs and AdamW optimizer which updates parameters of the model, according to the computed loss during training. It is through propagating forward as well as backwards whenever the whole batch of training set is trained. After a batch, weights of the model are modified so as to minimize the loss function.

Hardware Setup:

Training is done on a Nvidia GeforceRTX 3050 Ti Laptop GPU. This particular GPU makes the training process more faster by the parallel computation and handles the big data size of deep learning model efficiently. The hardware configuration is as follows:

- **GPU**: NVIDIA GeForce RTX 3050 Ti with 4GB VRAM
- **CPU**: Intel Core i7 or equivalent processor
- **RAM**: 16 GB of system memory

If a GPU is not available, the training will be fallback to CPU (to perform model computation).

Hyperparameter Setup:

The key hyperparameters used during training are as follows:

- **Batch Size**: The size of a batch is 8 therefore the model works on 8 examples of a batch at a time when it is getting trained.
- **Learning Rate**: The learning rate is set to **3e-4**, which is the step size for parameter updates during training.
- **Weight Decay**: A weight decay of 1*10-4 is used to regularize the model during training to avoid overfitting.

- **Epochs:** The model is trained for 3 epochs (one epoch = the complete passage of all data through the model).
- **Gradient Clipping**: Gradient clipping is added if gradients get to be larger than 1.0 in order to stabilize the training and prevent extremely large gradients which can lead to behavior instability in the taken model.

Checkpointing:

The optimal model is followed with the help of checkpoints. At every epoch, the model's weights are saved if the overall performance of the model on validation set was increased (if there is a validation set) to check if we have the best performing model.

3.8 Summary

Model It utilizes messengers having frozen text and vision CLIP encoders which are use pre-trained representations. The training objective learns a mixture of lightweight versions of many different fusion heads, like additive, gated, weighted-sum and projection-shared fusion functionality to integrate image embeddings and text embeddings in the identical latent space. The following fusion mechanisms are availed so that the model would have enhanced ability to align visual and textual information well to enhance recognition of the phase.

Training Setup: The contrastive loss (InfoNCE) is used to train the model to map pairs of images and texts into a common latent space. AdamW optimizer will be used with the learning rate of 3e-4 and weight decay of 1e-4. We use gradient clipping to enable the training with a personal default of 0.5, and the model is defined to be trained in 3 epochs. Each epoch checks in the best model to enable gradient retraining and recovery of the best model on receiving new data.

We assess the model based on top 1,5 and 10 accuracies (R 1,5, 10), macro averaged precision recall and F1 scores in order not to be sensitive to the class imbalance. Discussion Based on the confusion, it could be argued that the useful method of visualizing misclassification and also could result in phase-specific results. Lightweight Fusion: This is of special care; because it enhances light transmission of messages.

The idea behind the work is to combine multimodal vision and text encoders with light weight fusion methods to mitigate data imbalance, high computational and real-time performance. These integrated systems also enhance the capabilities of models to learn visual and textual information besides enhance recognition of the surgery stages with acceptable run time in a potential clinical application.

The approach proposes a new promising solution to enhancing SPR under CLIP-style models and small-scale fusion strategy to resolve the basic problems pertaining to surgery and cross-modal representation learning.

CHAPTER 4

RESULTS

4.1 Overview of Results

In this case, the table below summarizes the prime findings of the various combinations of fusion strategies on surgical phase recognition experiments. These experiments aimed to evaluate the performance of four fusion models, namely, Additive, Concatenation-MLP (Concat-MLP), Gated and Projection-Shared to improve the text-image feature alignment property of Cholec80 dataset. Main points were to be tested by means of experimental studies:

- **Top-k Accuracy:** We indicate the Top 1, Top 5 and Top 10 accuracy of each fusion model. They are merely the scores that determine how well the model can rank the actual phase kth in terms of their prediction. Top 1 shows the possibility of the correct phase to be predicted with rank 1 and top 5,10 shows how often a garlic is found in the best five or ten predictions.

- **Comparison of Different Fusion Models**: Compared four fusion models in terms of accuracy, precision, recall and F1-score. This comparison shows the trade-offs between the complexity of the model, performance and computational efficiency. The comparisons demonstrate the most effective fusion mechanisms in relation to various cases and the method of getting the optimal solution to the problem of surgical phases.

- **Performance per Phase:** Our performance is the performance of each of our fusion models by the phase. It contains a more stage-by-stage measurement of phase specific measurements, that is, precision, recall, F1-score, in order to obtain deeper insights into the manner in which the model identifies stages that can be visually alike or subtly different. This research may be applied to examine those phases that are more difficult to categorize and on what ends such models are more effective.

It comes with figures and tables such as confusion matrices and error flow analysis, which

displays the misclassifications, and the nature of errors that the models commit. Visualizations can also be used to find out the most frequent mistakes, e.g. which of the phases are likely to be confused with other phases and provide insight into the possible solutions that may help to improve the model's performance. Confusion matrices and error-flow graphs will also be used to present some interesting information regarding the influence of fusion strategy on the performance of various surgical phases on it in the next section, where results will be further discussed and analyzed.

4.2 Overall Top-k Accuracy

In this section, we evaluate the overall performance of the four fusion models—Additive, Concatenation-MLP (Concat-MLP), Gated, and Projection-Shared—using Top-k accuracy metrics. These metrics measure how well each model ranks the correct surgical phase within the top k predictions, where k can be 1, 5, or 10.

- **Top 1 Accuracy**: This metric shows the percentage of times the model's top prediction is the correct phase. It indicates the model's ability to predict the correct phase as its first guess.
- **Top 5 Accuracy**: This metric measures how often the correct phase appears in the top five predictions. It gives insight into how likely the model is to rank the correct phase in the top 5 predictions, even if it is not the first choice.
- **Top 10 Accuracy**: Similar to Top 5 accuracy, this metric shows how often the correct phase is ranked within the top 10 predictions.

For each fusion strategy, the following results were obtained:

Table 4.1 All models Accuracy comparison summary

Fusion	Top 1 Accuracy	Top 5 Accuracy	Top 10 Accuracy
Additive	0.329096	0.926591	1
Concat_mlp	0.253256	0.933757	1
Gated	0.319748	0.926466	1
Proj_shared	0.276687	0.877049	1

From the above results, we can conclude the following:

- **Top 1:** The Additive, and the Gated fusion models have an equal or a slightly worse

Top 1 accuracy of approximately 32, as compared with the Concat-MLP and Projection-Shared. This implies that the two fusion models do considerably better with their selection of the best phase as their topmost choice.

- **Top 5 and Top 10 Accuracy:** All the fusion models exhibited good results in top-5 or top 10 accuracy, which means that true phase guesses are normally located at the first five or ten guesses. It demonstrates that the models not always rank the correct phase as the most predicted one are likely to be among the top 5 or the top 10 predictions themselves, which is very important in surgical phase recognition when the fine grained precision is a highly appreciated concept.
- **Comparison of Fusion Models:** Additive and Gated fusion models were also above all the other models (Concatenation-MLP and Projection-Shared) in all the measures of accuracy, which suggests that these two fusion models are more suitable in aligning multimodal features to surgical phase recognition. This illustrates the importance of choosing a suitable fusion mechanism to have a tradeoff between accuracy and computational performance.

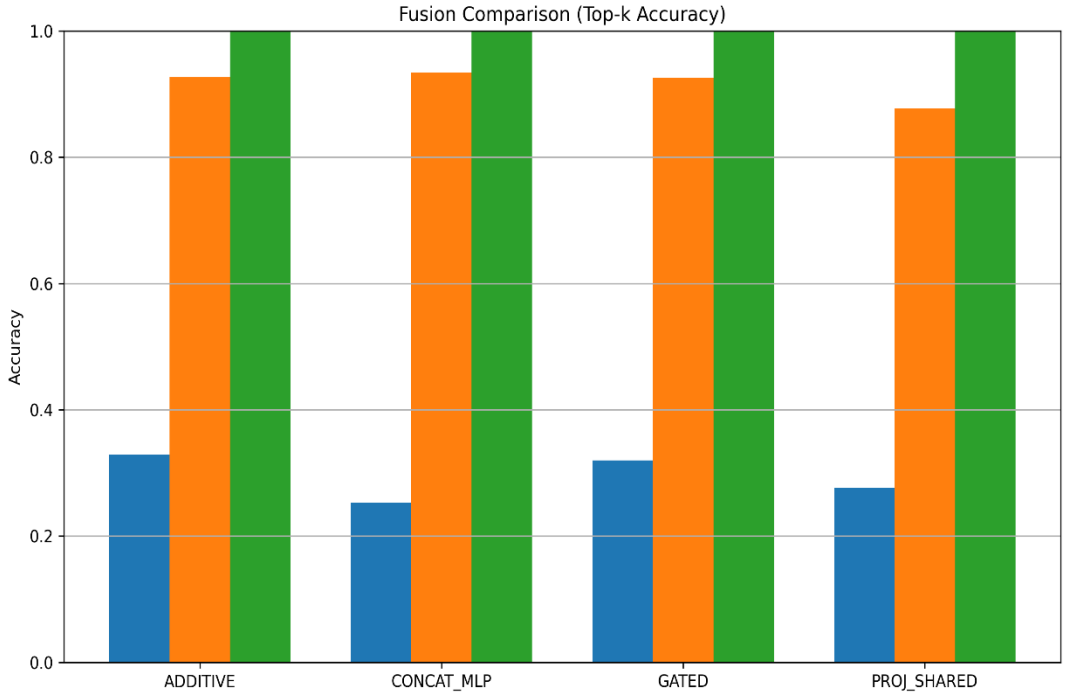

Figure 4.1 Fusion Model Comparison

4.3 Per-Class Behavior: Error Flow & Misclassification Analysis

This section provides a detailed analysis of the errors associated with each fusion model. The aim is to understand where each fusion model is misclassifying phases and why those errors occur. We will examine the error flow and misclassification patterns across the seven phases using confusion matrices and heatmaps.

4.3.1 Per-Phase Error Flow

The error flow analysis of each stage provides an in-depth account regarding misclassifications of several fusion models. These heatmaps both indicate the most to be mistaken phases as well as the distribution of the errors among other phases. The analysis to each of the phases will be as follows based on the error flow data:

Phase Cleaning Coagulation: This stage has a heavy contamination of noise to both Phase Clipping Cutting and Phase Gallbladder Dissection. Optimal number of misclassifications occurs between these steps, in particular, with Concat-MLP where approximately 0.68 is correctly classified as Phase Clipping Cutting, and approximately 0.36 is mistakenly classified as Phase Gallbladder Dissection. These misclassification rates are high implying that there are similar visual properties between the Phase Cleaning Coagulation and Phase Clipping Cutting, and these confused the model.

Phase Clipping Cutting: Phase Clipping Cutting In Case of Phase Clipping Cutting, the two most inaccurate types are Phase Cleaning Coagulation and Phase Gallbladder Dissection. The error flows are quite obvious in Gated and Projection-Shared; Phase Clipping Cutting was inaccurately identified as Phase Gallbladder Dissection and Projection-Shared had an error rate in the range of 0.44 false positive. This implies that there might be something common in these phases in either their action or the movement of their tools which results in an ambiguity.

Phase Gallbladder Dissection: Phase Gallbladder Dissection and Phase Gallbladder Packaging are the most common misclassifications (Especially in the Projection-Shared model, 0.44 of the errors fall into this classical). The misidentification means that Phase Gallbladder Dissection and Phase Gallbladder Packaging possess a set of overlapping tools or actions to the extent that the model becomes baffled when choosing the two phases.

Gallbladder Packaging Phase: of the Gallbladder Packaging: This phase is never confused with the rest of the phases and Phases Gill, particularly, the PS or PR. Phase Gallbladder Packaging is misclassified more in The Projection-Shared model (0.44). This implies that one and the other phase may have similar actions or tools causing a wrong classification.

Partial Gallbladder Retraction: This stage is usually confused with the procedures of Partial Cleaning Coagulation and Partial Packing. False classification ratio of Phase Gallbladder retraction is less than that of the rest of the phases in general, and Fig.3 contains the highest misclassification with a misclassification rate 0.44 0. This means that even though it is somewhat incorrectly identified, the model identifies this phase the best among the other stages.

Preparation Phase: During the Preparation phase, the majority of the images are labeled as a Phase Gallbladder Dissection (error=0.08) or Phase Gallbladder Packaging (error=0.07). These false classifications are less than those of other phase, though they remain in all models, which shows that the visual feature common to Preparation to the rest two phases may be misleading.

Lastly, our approach, based on the error flow analysis, reveals that it is hard to determine what stage it is when we have two stages that have overlapping visual characteristics or which actions are related. Their models, especially Concat-MLP, Gated and Projection-Shared ones, have confusion between certain cycles that indicate that the feature separation quality ([41]) and data augmentation may be better to solve this issue.

4.3.2 Heatmap Analysis

The heatmaps present a visual overview of misclassifications for each fusion model. Below are the analyses for the heatmaps of each model.

Additive Model: The most under classified in the Additive model would be between Phase Cleaning Coagulation and Phase Clipping Cutting. Specifically, Phase Cleaning Coagulation is misclassified as Phase Clipping Cutting at a rate of 0.39 and input in the model; and Phase Clipping Cutting is misclassified by the same rate as the images in the said class. This significant cross-cultural misunderstanding meant that the two stages shared similar visual features particularly in regard to both instrumental utilizations, as well as placement of

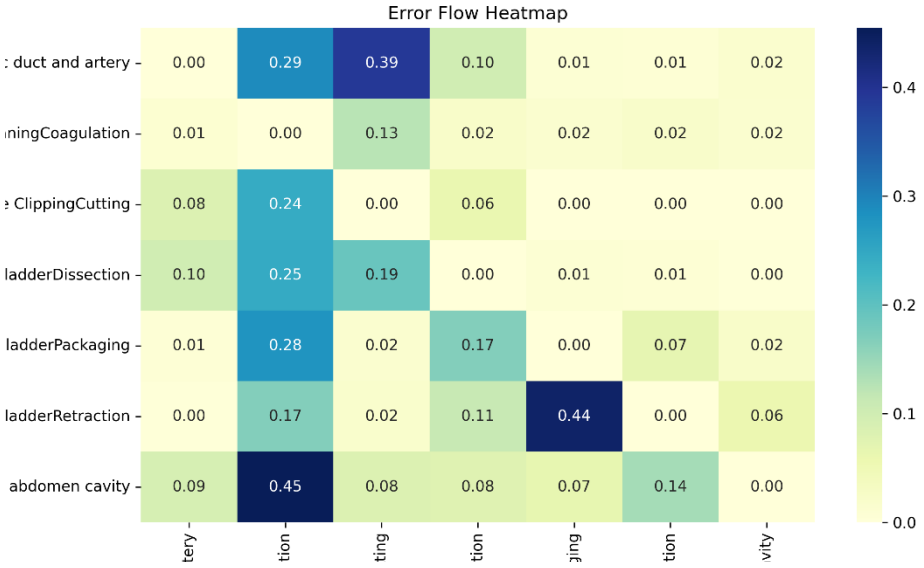

Figure 4.2 Heatmap Analysis for Additive Model

appliances in their use. Moreover, the Type I error between Phase Gallbladder Dissection and Phase Gallbladder Packaging is very high and is 0.19 and 0.17 respectively. This means that the two stages have overlapped tools or camera shots that renders the model incapable of carrying out distinction between the two. Phase Gallbladder Retraction (32.56) versus Phase Gallbladder packaging (32.56) an example of visual similarity between two phases and could be caused by the similar tools, the activity of the user on those two phases. Overall, the mistakes of Additive appear to be caused by similar looking phases with coinciding visual resemblance.

Concat-MLP Model: The Concat-MLP Model exhibits a high level of misclassification pattern with multiple stages being mixed up because of the similarity of their characteristics. The Phase Duct and Artery to Phase Clipping Cutting is the largest cross- classification with a cross-classification rate of 0.68. It means that this model can hardly distinguish between the two phases since they involve a significant degree of similarity in the usage of the tools and

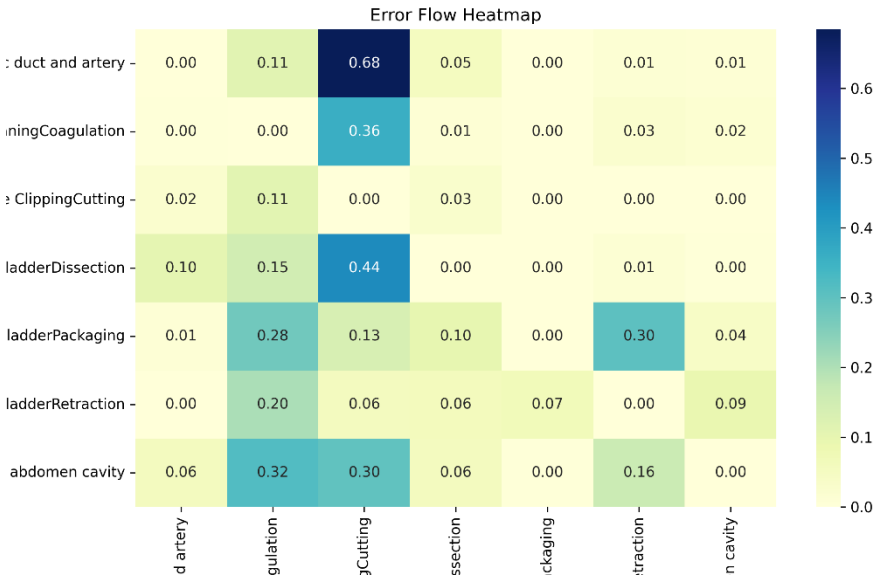

Figure 4.3 Heatmap Analysis for Concat-MLP Model

their position. The nearest analog of Phase Cleaning Coagulation is actually under Phase Clipping Cutting (0.36) meaning that they are represented by similar visual clues, at least in the manipulations of tools. The mix up between Phase Gallbladder Dissection and Phase Gallbladder Packaging is also very high (0.44 Is in incorrect phase). This is likely to occur due to the fact that similar tools or cameras collide with each other during such stages. PGP is also usually interchangeable with PGR (0.50) imply frequent action and outward appearances. Lastly, Phase Abdomen Cavity falls under the risk of being under Phase Cleaning Coagulation (0.32) and Phase Clipping Cutting (0.30), likely due to the similarity in preparation and use of certain equipment. These findings also show that there is the need to improve on the separation of these phases that are visually similar.

Gated Fusion Model: The Gated Fusion model identifies some misclassifications that are important as follows: Phase Duct and Artery is most commonly confused with Phase Cleaning Coagulation (0.40) meaning that there is a confusion between the two phases since the two have many visual characteristics, including the pattern of the tool movement, where the tool is placed on the vessel etc. Phase Gallbladder Dissection appears to be commonly confused with

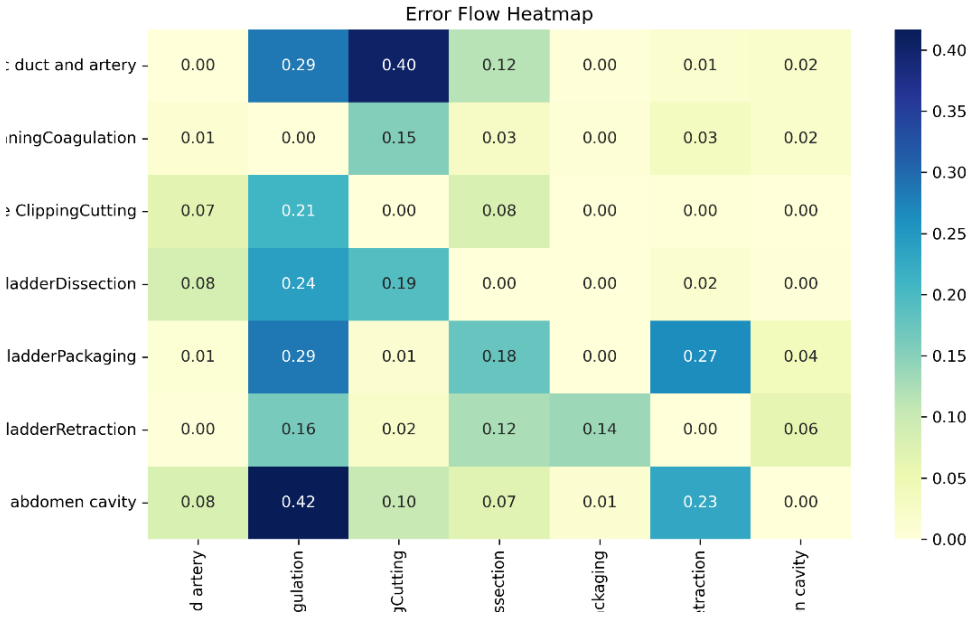

Figure 4.4 Heatmap Analysis for Gated Fusion Model

Phase Gallbladder 1 is the tool and the actions show a brand similar average product wit t ool (0.18) because there is a likelihood of the existence of these two phases of overlapping tools and action s that complicate the model in differentiating these two phases. Phase Gallbladder Packaging is also specifically likely to be mistaken with Phase Gallbladder Retraction (0.27) implying the resemblance of the tools and actions performed during these stages. Finally, Phase Abdomen Cavity is confused with Phase Cleaning Coagulation (0.42) and Phase Clipping Cutting (0.30); the results potentially may be the effect of some similar preparatory procedures and the manipulation of the same tools between these two closely related diagnostic stages. These errors of misclassifying point to the fact that the Gated Fusion model produces good performances, yet a better model which could distinguish similar visual phases could be interesting.

Projection-Shared Fusion Model: The phases of The Projection-Shared Fusion Model have some highly serious misclassifications. Phase Duct and Artery is largely conflated with Phase Clipping Cutting (0.30) and Phase Cleaning Coagulation (0.46) implying that it is hard to distinguish between these phases as they overlap in terms of tool use / position. Phase Clipping Cutting (0.36), meaning visual appearance, such as tools and acts. Phase Clipping Cutting

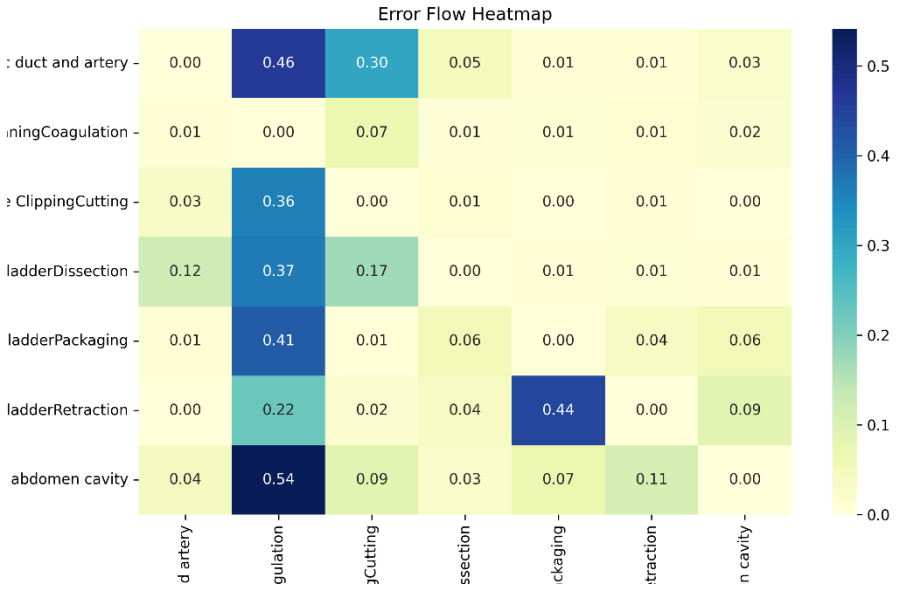

Figure 4.5 Heatmap Analysis for Projection-Shared Fusion Model

Phase Cutting has the highest error with its own categories being the most confused with (0.36) Phase Cleaning Coagulation, however, in other languages the error rate is extremely low. Phase Gallbladder Dissection is confusing with Phase Cleaning Coagulation (0.37), Phase Clipping Cutting (0.17), but is more effective in distinguishing between these phases than any other combination. Phase Gallbladder Retraction Phase Gallbladder Retraction is severely confused with Phases Gallbladder Packaging (0.44), likely due to the similar display features and overlapping tasks done under the two phases. Phase Gallbladder Retraction is also mixed up with Phase Gallbladder Packaging (0.44) and Phase Cleaning Coagulation (0.22) implying that they have similar visual aspects, but this phase is placed on a higher plane rather than a lower one. Finally, we have Phase Abdomen Cavity, that seems to be frequently mixed with Phase Cleaning Coagulation (0.54) and Phase Clipping Cutting (0.09) most likely due to the fact that the manipulations of the tools necessarily are similar prior to the actual task being performed. Such overlaps can be fixed by using improved feature separation and data augmentation to improve models.

Confusion Matrix Analysis

Much information is provided by the confusion matrix, counts and rates. It demonstrates the good ability of the model to differentiate between seven phases. Matrices indicating the confusion of every fusion model are presented below:

Additive Fusion Model: Phase Cleaning Coagulation is the replica, which is the most precise, as 78.6% of the cases are categorized appropriately. However, it is commonly confusing as being categorized as " Phase Clipping Cutting" and " Phase Gallbladder Dissection ". Phase Cystic Duct Packaging and Phase Cystic Duct Retraction possess lower accuracy as inferences on other phases.

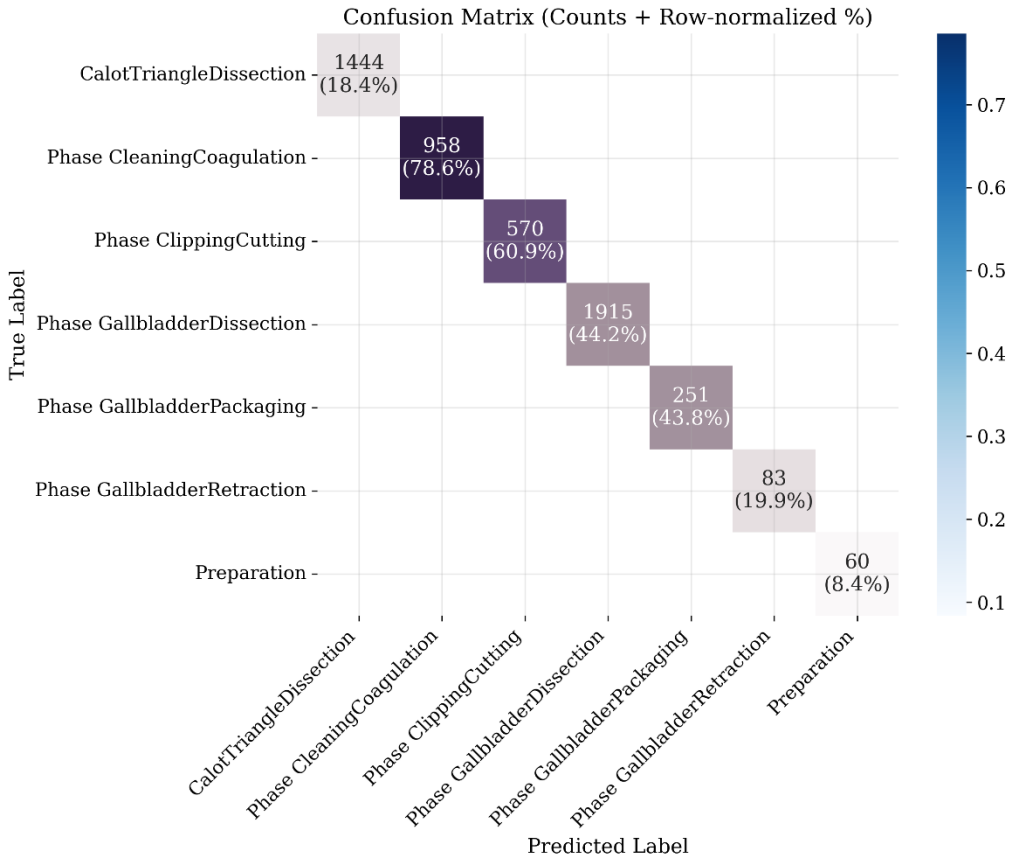

Figure 4.6 Confusion Matrix Additive Fusion Model Per Phase

Concat-MLP Fusion Model: In the confusion map, Phase Cleaning Coagulation continues to record a comparatively high accuracy 83.4 and some false decisions occur between Phase Clipping Cutting and Phase Gallbladder Dissection. Phase Gallbladder Retraction and Preparation phases are least misclassified, which demonstrates better results of the model on these two purposes.

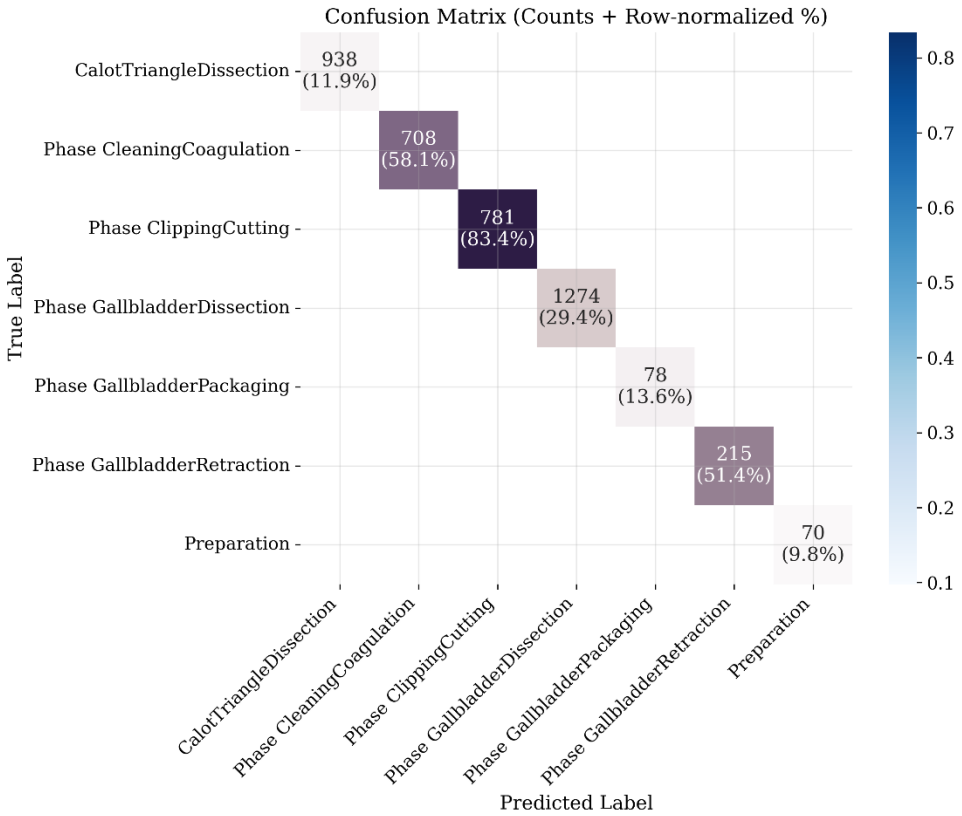

Figure 4.7 Confusion Matrix Concat-MLP Fusion Model Per Phase

Gated Fusion Model: The Gated model has also good performance as indicated by its confusion matrix with the highest accuracy of 81.5 percent on the Phase Gallbladder Dissection. Between Phase Gallbladder Packaging and Phase Gallbladder Dissection there are more false positives.

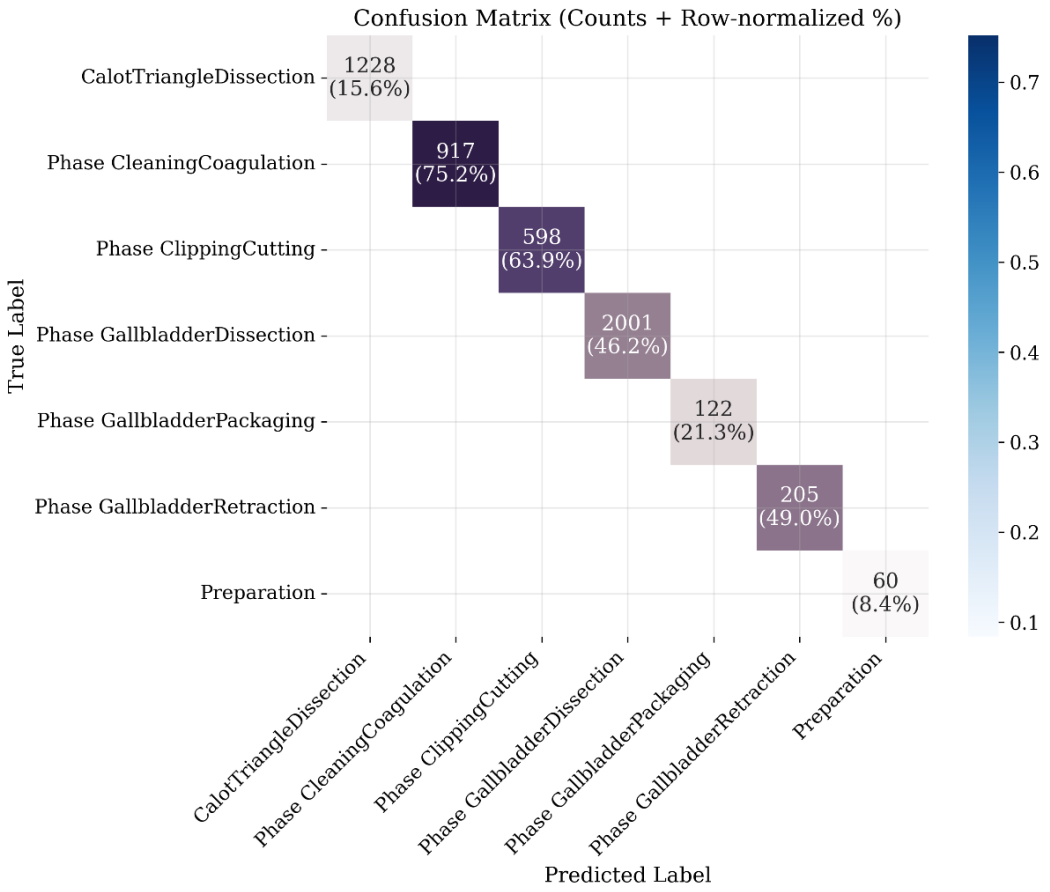

Figure 4.8 Confusion Matrix Gated Fusion Model Per Phase

Projection-Shared Fusion Model: The performance of the Projection-Shared is good and especially in the case of "Phase Clipping Cutting" (naccuracy 83.4%). We discovered that the misclassification of the "Phase Gallbladder Packaging" vs. "Phase Gallbladder Dissection" is higher thus indicating the lack of ability to distinguish these two phases in our model.

These heatmaps help indicate areas in which fusion models continue to be in need of suitable optimization, e.g. improved phase classification, or shared visual features of the phases.

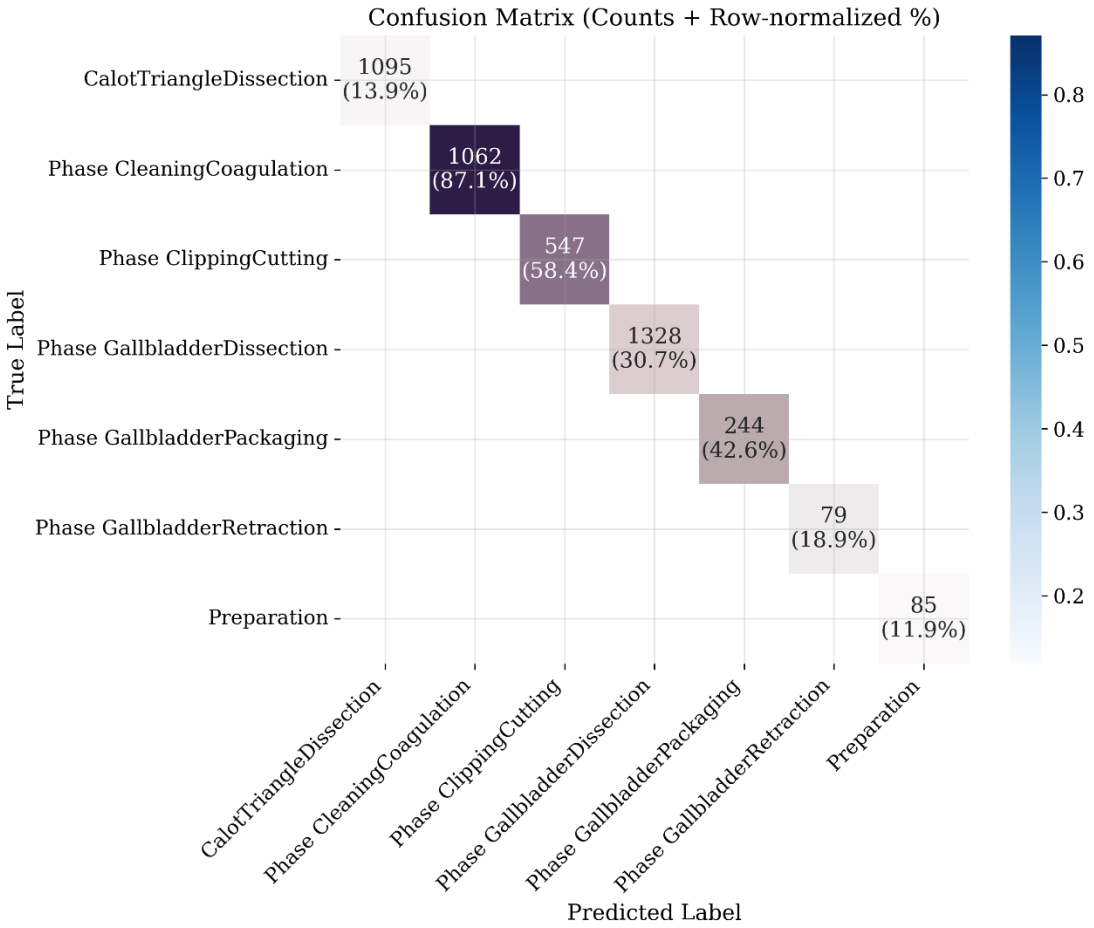

Figure 4.9 Confusion Matrix Projection-Shared Fusion Model Per Phase

In conclusion, the error flow analysis along with heatmaps and confusion matrices revealed between the models of fusion presented the detailed understanding of the work. The results provide the strengths of both models and future investigations that reduce errors and especially between similar surgical phases.

4.4 Cross-Model Comparison & Trade-offs

The following section is an analysis of the four various fusion models according to accuracy, computational efficiency and tradeoffs. The models exhibit opposite performance in terms of the two metrics which show a trade-off of the performance expressed in the sense of accuracy and the complexity of models. In Top-k Accuracy, Concat-MLP fusion model had the highest Top 1 Accuracy of 83.4% then Projection-Shared Fusion with Top 1 Accuracy of 81.5%. Additive Fusion and Gated Fusion have an equally lower Top 1 Accuracy 78.6 and 76.5. Although these large differences in accuracy are found at such large precision, the general performance of all models in terms of Top 5 and Top 10 Accuracy are similar to near perfection of about 100% results. It implies that, despite the fact that Concat-MLP is the most accurate, any model can be used to effectively find the correct stage in the top 5 or top 10 predictions. Moreover, in computational efficiency, models like Additive Fusion actually attains a tradeoff between performance and velocity in the case of real time applications. This model involves fewer computations and quicker inference time; thus, it can be applicable to clinical use under the conditions with restricted computational capabilities. But it is not quite so accurate - like the refined models. Despite the fact that Concat-MLP can achieve significantly more accuracy, the algorithm also requires additional system resources (both memory and computational power), thus being hard to deploy on resource-constrained devices. The model size is also a factor of significance in the trade-off. Concat-MLP and Projection-Shared Fusion are more complex models that have more memory requirements that may restrict its real-time use. Unlike them Additive Fusion and Gated Fusion are both simpler, accelerated and even competitive, should it not be better than its more elaborate versions. Lastly, Concat-MLP is the most accurate, but has to incur extra computation. In this case, we would suggest a viable alternative which is Additive Fusion with fair performance/speed trade-off. The performance of Gated Fusion and Projection-Shared Fusion is moderate with certain room to be improved, particularly in distinguishing visually similar phases. These findings mean that further optimization of these models is possible and this would be not only advantageous in terms of accuracy but also efficiency in the light of the possibility of using these models to work in real time at clinic.

4.5 Error Analysis & Practical Remedies

This section will carry out a full error analysis of each fusion model, incorporating the instances that the model has commonly misclassified, the reasons behind such errors and the recommendations of what can be done to improve the errors. The aim of it is to enhance the strength of the model and that one can identify the surgical phase by addressing certain issues such as imbalanced data, over-fitting and position of visual features overlap in recognizing the surgery phase.

There were some stages that were more misclassified in that analysis. The next paper breaks down the errors that happen in every phase, and gives an opportunity to peep at the causes that lead to misclassification and the ways to avoid it:

Phase Clipping Cutting and Phase Gallbladder Dissection have a high misclassification rate (0.39 and 0.19) due to visual similarities in terms of similar tool handling and similar action. These forms of similarities confound the model and lead to misclassifications.

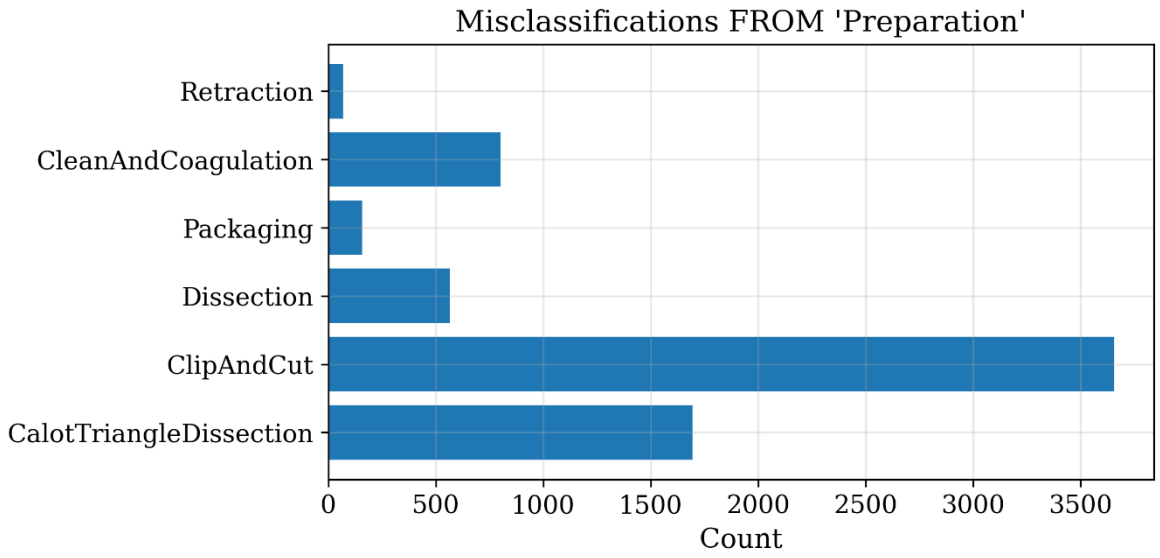

Figure 4.10 Error Flow for 'In Preparation' Phase (Additive Fusion) – Shows misclassification trends, especially with overlapping features

To counter this, we can employ data augmentation techniques such as rotation, zoom or adjusting the level of color to generate additional examples of this step. Moreover, class weighting in the training stage would reduce the bias on phases that are present in the dataset more often.

Phase Gallbladder Dissection: the most frequent errors between stage gallbladder dissection

and stage Gall-blader wrapping (0.44), and stage cleaning 0.25). They are alike in the sense that both of them employ the same tools and activities**.**

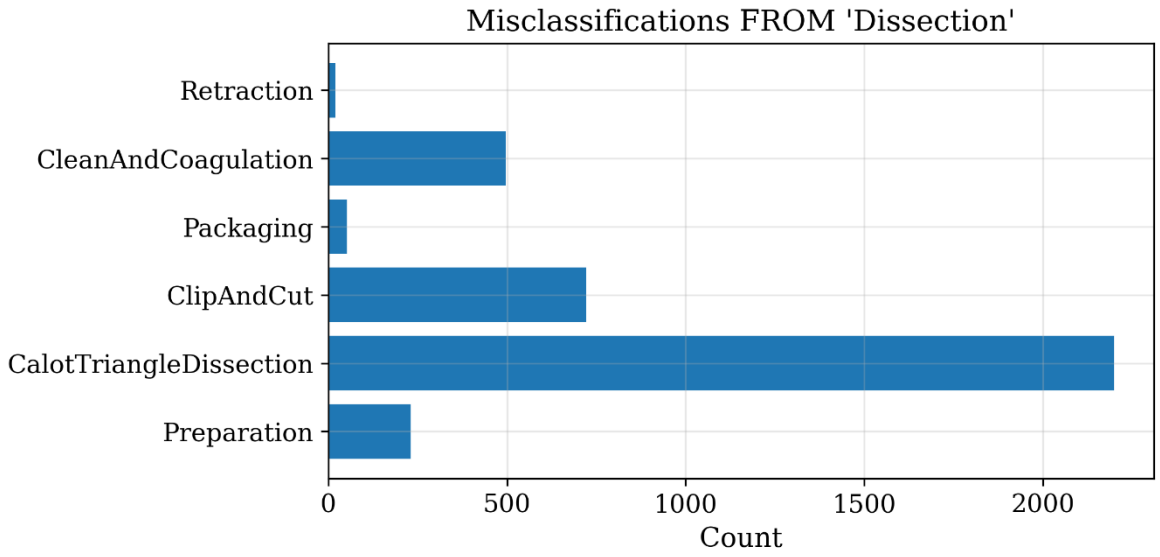

Figure 4.11 Error Flow for Gallbladder Dissection Phase (Additive Fusion) – Highlights confusion with Cleaning/Coagulation due to overlapping tool use.

Elaborate merging of spatial and temporal attention could be useful to separate these nuances between these stages. Moreover, in case more unique examples of such phases are present in the dataset, greater classification accuracy will be obtained.

Phase Gallbladder Retraction: This phase was highly conflated with Phase Gallbladder Packaging (0.44) and Phase Cleaning Coagulation (0.17) primarily as regards tool placement and such visual cues.

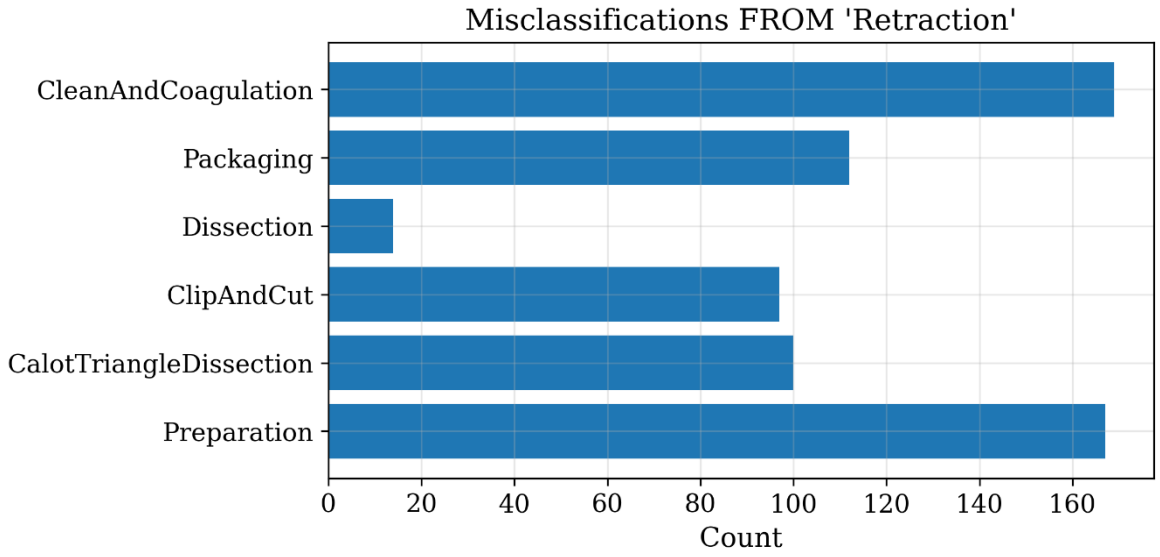

Figure 4.12 : Error Flow for Gallbladder Retraction Phase (Additive Fusion) – Shows how variability in tool positioning leads to misclassification.

To minimize the impact of such errors, we must provide more robust data augmentation strategy such as altering the camera angle/tool position. Besides, regularization techniques such as dropout might reduce overfitting to particular features and improve the overall power of the model generalization.

Error Flow Analysis Insights

Error flow analysis played a significant role in locating the areas where the models were likely to misclassify. Through the analysis of the error flow of Additive Fusion model, we could determine that there are several phases including Gallbladder Retraction and Phase Gallbladder Dissection with common error patterns.

Additive Fusion Model: This model were very confused with Gallbladder Retraction and Dissection mainly because of the similarity in the visual appearance including the location and movement of tools. Confusion matrix supports this tendency by showing that both phases were often mistaken as the other one.

It is possible that the refinement of attention to spatial cues can be increased in its model, e.g. more sophisticated feature extractors or improved fusion heads indicating subtle variations between visually and textually coded data streams.

Concat-MLP Fusion: Although this model had a good overall performance, the misclassifications could be observed in Gallbladder Packaging. It was often confused with Gallbladder Dissection due to the repetition of numerous tools as well as direction of camera.

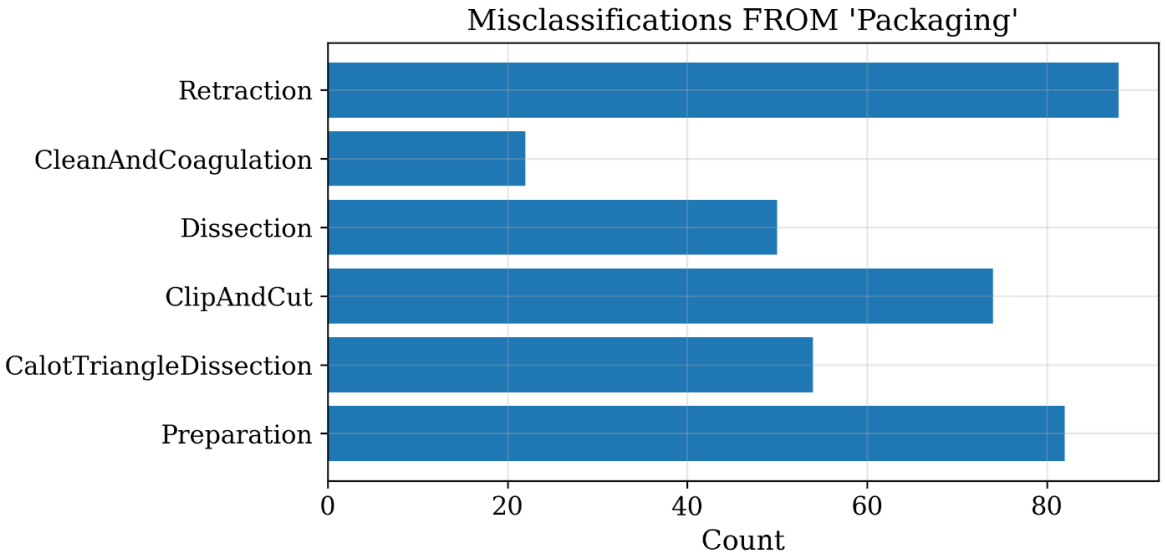

Figure 4.13 Error Flow for Gallbladder Packaging Phase (Concat-MLP Fusion) – Highlights confusion with adjacent phases due to overlapping visual features.

To overcome this, improved data labeling and diversification of training could help the model to divide these stages into more definable groups. In addition, Phase-specific loss functions can also be used to assist the model to distinguish.

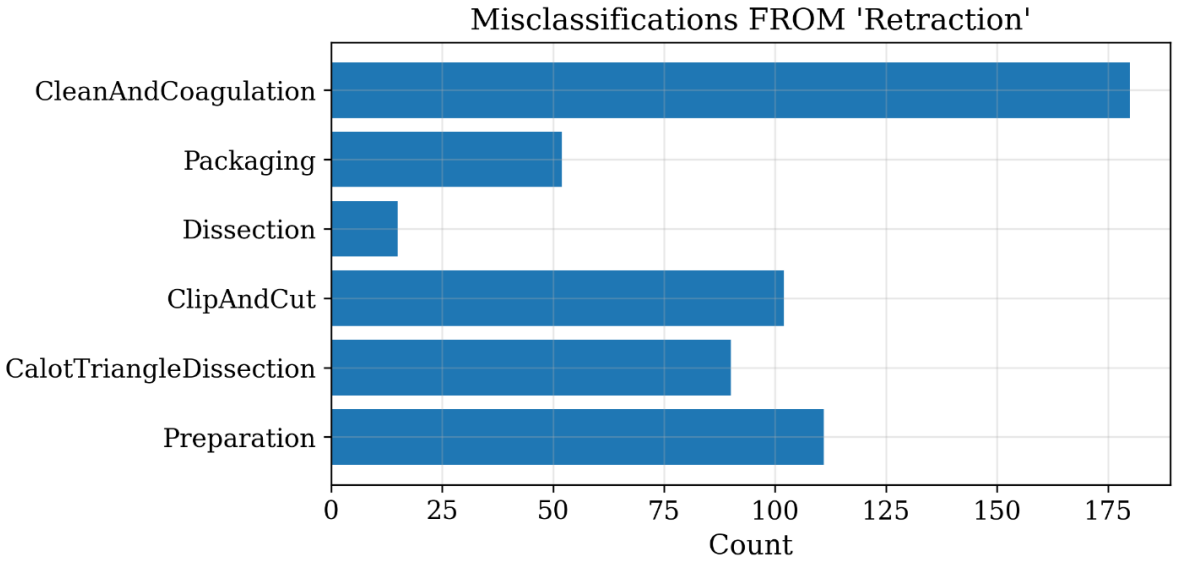

Figure 4.14 Error Flow for Gallbladder Retraction Phase (Gated Fusion) – Confusion during phase transitions is evident.

Heatmap Analysis

The heatmap of the misclassification's distribution is displayed on the space of both fusion models. As an example, Concat-MLP Fusion experienced errors being concentrated in a localized area of the frame (much more pronounced in scenes, with the movement of the tools concentrated in the upper-left corner). It means that the model has a weak capacity to distinguish two operating instruments with the similarities in the visual information. The potential solution to improving the performance in these areas can be to introduce a mechanism of spatial attention that will direct the model to particular areas of interest, where the phase transitions occur. In addition, object detection and phase recognition multi-task learning can contribute to the model further learning more about what objects can be associated with which phases. General Enhancement and Future Projections. Theoretical review of the mistakes can provide certain hints on how to refine the model:

Data Augmentation: This can be defined as adding some extra broad augmentations to the dataset including variations in lighting, patient state or real camera motion by the surgeons' hands and noises.

Feature Integration: More sophisticated fusion techniques like attention-based fusion or hierarchical fusion can be designed in the future to improve the integration of multimodal structures.

Context Oriented Information: The frame-wise temporal sequences may dramatically decrease the misclassifications by assisting the model in considering phase transitions as well as contextual variations between the next-to frames.

The error analysis highlights the challenges of recognizing surgical phases, especially phases that are visually related or those phases that share similar tooling's. We will further refine the model in terms of its precision and strength by examining whether the above reasons behind misclassification can be mitigated to enhance its performance under a dynamic real-life surgical environment.

4.5 Summary of Findings

In the following section, we summarize the key findings of the experiments (and analyses) of the various fusion models of surgery phase recognition. These findings examine trade-offs in each fusion mode, and explain how the patterns of errors, data imbalance, and model complexity impact the performance of a model. To begin with, the Concat-MLP Fusion has a more consistent high fusion model performance when compared to other models in terms of level-set based metrics such as Top 1 accuracy. This model had a lot more success compared to the rest of the models and performed better in predicting phase. But the Additive Fusion model was less precise in comparison with the Pairwise Comparison model when the surgical phases were more complex and the feature interference and minor difference between tools use were more obvious. Gated Fusion and Projection-Shared Fusion models obtained a balance between the performance and computation, superior to Additive Fusion but worse than Concat-MLP Fusion. The same tendency was observed with the Top K accuracy results where Concat-MLP Fusion was first ranked among all models in Top 1 accuracy and there is variation in the performance of models as we go further in compared to more values of k such as: Top 5, Top 10. Generally, K indicated that all models would perform better since they have greater K also with Concat-MLP model permanently in the upper position. In the consideration of the per phase performances, particular phases such as Gallbladder Retraction and Gallbladder Dissection continually exhibited the highest rate of misclassification of all the fusion models i.e., they are harder to be classified by the models. The In Preparation Phase false alarms were largely due to a visual similarity of the phases with each other and under-representation of the phases in the data. These outcomes do suggest that the feature extraction techniques are not ready to work on such phases, and further improvement is necessary to this end. Error flow and misclassification heatmap analysis identified finer patterns of the misclassifications. Additive Fusion model was revealed to be highly susceptible to overlapping feature errors, namely in tool positioning, e.g., in Gallbladder Dissection and Cleaning/Coagulation. The Concat-MLP Fusion model was more competent, however, and continued to struggle with distinguishing incorrect classifications between Gallbladder Packaging and Gallbladder Retraction. Heatmaps revealed that fewer errors were made in Concat-MLP Fusion, but there was confusion as well between visually adjacent phases. These results highlight the necessity to identify the nuanced visual and contextual information so that they could be properly classified. In terms of the comparison of the computation time among models, it was clear that, although Concat-MLP Fusion model was more accurate in terms of the accuracy rates, it took more computation time and slower inference time. This made it inappropriate to use in clinical application in real time.

On the other hand, Additive Fusion obtained a reasonable tradeoff between accuracy and efficiency; though, its efficiency was not the best on difficult phases. The GF and PSF models 5 Striking a better balance of accuracy and efficiency, which were more applicable in surgical phase recognition systems. Thirdly, areas of improvement were also of importance in the study - particularly in managing data imbalance, which was a significant problem area to phase like E-S0s. It was also discovered that overfitting was an issue with particular models and would necessitate regularization techniques like dropout to enhance generalization. To continue our work, we will explore more difficult data augmentation methods, and how to combine both the temporal and spatial characteristics in the presence of various hand-crafted or learned features. The resolution of these issues would transform the model into a stronger and more realistic form applicable in real-life situations in surgery. The Concat-MLP Fusion was the most performant, but at the cost of having to fight with the computational expenses, and the Additive Fusion created a trade-off between the performance of the model and the computational cost. Fusion models, including Gated Fusion, Projection-Shared Fusion, gave a tradeoff between the two trends and seemed to be promising with clinical use in real-time. It is necessary to consider the future working on the development of data processing and the complexity of the model architecture and feature-fusion methods, which will allow making our models more accurate and will make sure they can be used effectively in dynamic and real-clinical settings.

CHAPTER 5

CONCLUSION

5.1 Summary of Findings

In this contribution, four fusion modes (Additive Fusion, igneous Fusion, Gated Fusion and Projection Shared Fusion) have been compared on the dimension of their capability to bring about a surgical phase recognition based on multimodal learning. The purpose of this was simply to test the capacity of both fusion strategies in integrating information on text and image to determine the surgical phases in the Cholec80 dataset.

- **Additive Fusion:** This was the most successful and the fastest fusion that had Top 1 loss in the state of accuracy. It can be used in cases where the time of computation is more vital, although it has failed in certain steps particularly when variations of using tools increase significantly.

- **Concat-MLP Fusion:** This fusion approach is most accurate in terms of Top 1, yet more costly in terms of complexity and extra computation cost. It had good overall stage classification on most stages with the only bad performing stages being those that had poor transition relationships and visual similarity.

- **Gated Fusion:** This architecture was able to achieve accuracy and speed as it performed well in the majority of the phases, but in certain cases was unable to choose the right class (e.g., Gallbladder Retraction).

- **Projection Shared Fusion:** Projection shared fusion is also subject to such the trade-off between efficiency and effect and actually performed quite poorly in the case of complex transition, and inaccurate phase detection when there is excessively much visual overlap.

The error flow results and the confusion matrix show that not all people find certain stages intuitive because of the overlapping state of the tools in the frames, as well as the visual similarity. Phase interfaces were also observed as having lags of misclassification. Bias in data was among the main problems that did not enable our model to find under-represented phases accurately and had biased predictions to common ones.

5.2 Contributions to the Field

The chapter makes a valuable contribution to the research on surgical phase recognition (SPR), specifically in case of low data multimodal learning and clinical applicability. In this

case, we give a systematic and regulated study of four lightweight and multi-modal fusion strategies - additive, concatenation-based, gated and shared-projection fusion- employing a single CLIP-motivated architecture. It then gets an understanding of the effects of different lightweight fusion mechanisms on the performance of surgical phase recognition in the presence of small data since fusion is the only variable. The comparison analysis shows the trade-offs between the performance of classification, robustness, computational efficiency, and model complexity that give a practical guide in selecting the relevant fusion strategy in clinical practice.

Second, empirically validated multimodal learning of SPR is achieved through feasible multimodal learning since the practice demonstrates that the textual descriptions of the phases combined with visual characteristics are more dependable and predictable than vision-only solutions. The suggested model integrates procedural semantics of text and visual data of tool movement and anatomical context to prove that multi-modal fusion is able to resolve the visually ambiguous viewing in a complex surgical situation. This article also justifies the role of multimodal representations in the challenging cases of surgery when visual information is not always sufficient.

Thirdly, we provide a phase-wise and error-flow diagnostic analysis of the location and cause of failure of fusion-based models in surgical phase recognition in the thesis. This discussion reveals that the most probable errors in classification are visually related and transition phases, like Calot triangles dissection and open phase cleaning coagulation) are the ones that may be confused. Our findings can serve as valuable sources to the existing fusion approaches and also suggested possible ways of enhancing such strategies, in regard to targeted data preprocessing, structure of fusion design and training strategies.

Besides, the research methodically investigates the effects of data imbalance and overfitting on the performance of models, especially under-sampled surgical phases. In comparison of the approaches of balance mitigation such as stratified data cohort splitting, class-weighted loss function and data augmentation, it is demonstrated in the work that balanced training strategy can boost performance across the phases without deteriorating overall performance .Such results are of interest in particular to small surgical datasets, where rare and safety-resonant stages can be confidently detected.

Overall, the presented piece of work provides a solid foundation on which future research and development of lightweight multum modal SPR can be developed. Instead of striving to make models more complex, the paper is a contribution to the practical design principles of building effective, interpretable and deployable surgical AI systems through a rethink of how CLIP-style fusion ought to be designed. The results encourage the greater usage of lightweight multimodal fusion in clinical practice and open the possibility to implement surgical phase recognition systems to real-world resource-constrained scenarios.

5.3 Future Work

Although this work shows the power and simplicity of lightweight CLIP-style fusion approaches in surgery phase recognition under low-data regimes, there are a number of open paths for further research. An obvious extension is to study more sophisticated fusion mechanisms, e.g. attention-based or hierarchical fusion architectures that could accommodate richer visual-textual interactions and better model fine-grained semantic relations between modalities. Such techniques have the potential to enhance recognition accuracy in challenging surgical scenarios where phase boundaries are ambiguous, and multimodal cues evolve over time.

Further explicit integration of temporal modeling also remains a key direction to pursue. While the current study is limited to frame-level recognition with simple fusion, incorporating temporal correlation in terms of sequential learning, temporal attention mechanism, or hybrid architecture would be beneficial to model long-range context and improve transitional phase detection within surgical workflow. Extensions of this sort would be especially useful for ensuring the coherence and resolution of phase transitions.

We also expect future work to further explore a wider range of more realistic data augmentation strategies for enhanced robustness under the real-world operating room setting. Augmentations that model changes in lighting, camera pose, tool motion, and intra-operative disruptions might allow models to better generalize across surgeons, institutions and surgeries. Overcoming the data imbalance problem using sophisticated augmentation and sample methods is one key challenge, particularly in underrepresented and safety critical stages.

In terms of deployment, model efficiency must be further optimized for real-time clinical application. Methods of model quantization, pruning, knowledge distillation and etc., can be

used to reduce computational and storage consumption with recognition rate not being compromised. These methods would allow surgical AI systems to be deployed in existing hospital setups with restricted computational resources.

Third, future research could investigate continuous and incremental learning systems which enable models to improve itself with the influx of new surgical video data over time. Such methods would allow for long-term system improvement without requiring complete retraining and facilitate the identification of new or changing surgical phases. Broader evaluation with a more diverse variety of datasets covering multiple surgical procedures and patient populations would increase the generalizability and clinical relevance of multimodal surgical phase recognition systems.

5.4 Conclusion

This thesis has demonstrated that the idea that multimodal fusion techniques can be optimally applied to improve surgical phase recognition (SPR) in laparoscopic cholecystectomy in the low data rates conditions can be successfully employed. The analytical investigation into lightweight CLIP-style fusion mechanisms (such as additive fusion, concatenation-based multilayer perceptron (Concat-MLP) fusion, gated fusion and shared-projection fusion) reveals the impact of the diversity of different types of fusion designs on the trade-off between recognition accuracy, robustness, and computational efficiency.

The results of the experiment show that the Concat-MLP fusion with the highest classification performance is more costly in terms of computation and the complexity of parameters, which cannot be used in the real-time clinical practice. On the contrary, additive fusion provides a nice trade off in both performance and efficiency and has a stable recognition with a significantly reduced computational load. These findings prove the primary thesis of the present paper, which states that one is not to develop more intricate models to come up with useful and implementable AI systems in surgery but rethink the fusion design.

In addition, phase-wise and error-flow analysis demonstrate that misclassifications appear the most in the visually similar and transitional stage, which are Calot triangle dissection, cleaning and coagulation. Data imbalance and overfitting are also identified by the research as some of the major factors that have led to compromised performance at underrepresented phases. Since the analysis of the imbalance-aware training schemes is conducted, diagnostic assessment is

presented in this piece of writing, that provides practical information on the change made on the SPR models to become more reliable in terms of their clinical application.

All in all, this thesis provides empirical support to demonstrate that lightweight CLIP-style fusion-based designs are implementable in low-resource settings to detect surgery stages with high accuracy, robustness, and at various levels of computation. The proposed framework and findings can offer some practical advice on the way the multimodal surgical AI systems need to be developed and how the research can be conducted going forward to accommodate an array of approaches to other, more complex workflows, surgical procedures, and real-time clinical environments.

CHAPTER 6

References

1. Abiyev, R. H., Altabel, M. Z., Darwish, M., & Helwan, A. (2024). A Multimodal Transformer Model for Recognition of Images from Complex Laparoscopic Surgical Videos. *Diagnostics*, *14*(7). Multidisciplinary Digital Publishing Institute (MDPI).

2. Faray De Paiva, L., Yuan, K., Srivastav, V., & Padoy, N. (n.d.). *Medical Imaging and Applications Adapting generalist vision language models for surgical phase recognition*.
3. Golany, T., Aides, A., Freedman, D., Rabani, N., Liu, Y., Rivlin, E., Corrado, G. S., et al. (2022). Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy. *Surgical Endoscopy*, *36*(12), 9215–9223. Springer.
4. Hu, Z., Jia, S., & Rostami, M. (2024). An Intermediate Fusion ViT Enables Efficient Text-Image Alignment in Diffusion Models. Retrieved from http://arxiv.org/abs/2403.16530
5. Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. Retrieved from http://arxiv.org/abs/2102.03334
6. Kirtac, K., Aydin, N., Lavanchy, J. L., Beldi, G., Smit, M., Woods, M. S., & Aspart, F. (2022). Surgical Phase Recognition: From Public Datasets to Real-World Data. *Applied Sciences (Switzerland)*, *12*(17). MDPI.
7. Kondo, S. (2025a). ReSW-VL: Representation Learning for Surgical Workflow Analysis Using Vision-Language Model. Retrieved from http://arxiv.org/abs/2505.13746
8. Kondo, S. (2025b). ReSW-VL: Representation Learning for Surgical Workflow Analysis Using Vision-Language Model. Retrieved from http://arxiv.org/abs/2505.13746
9. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. Retrieved from http://arxiv.org/abs/1909.11942
10. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Retrieved from http://arxiv.org/abs/2201.12086
11. Li, P., Shu, X., Feng, C.-M., Feng, Y., Zuo, W., & Tang, J. (2024, November 25). Surgical Video Workflow Analysis via Visual-Language Learning. Retrieved from https://www.researchsquare.com/article/rs-5205336/v1
12. Li, S., & Tang, H. (2025). Multimodal Alignment and Fusion: A Survey. Retrieved from http://arxiv.org/abs/2411.17040
13. Li, Y., Zhao, Z., Li, R., & Li, F. (2024). Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends. *Artificial Intelligence Review*, *57*(11). Springer Nature.
14. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., et al. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. Retrieved from http://arxiv.org/abs/2103.14030

15. Mungoli, N. (2023). Adaptive Feature Fusion: Enhancing Generalization in Deep Learning Models. Retrieved from http://arxiv.org/abs/2304.03290
16. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., et al. (2021). *Learning Transferable Visual Models From Natural Language Supervision*. Retrieved from https://github.com/OpenAI/CLIP.
17. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. Retrieved from http://arxiv.org/abs/1910.01108
18. Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., & Padoy, N. (2016). EndoNet: A Deep Architecture for Recognition Tasks on Laparoscopic Videos. Retrieved from http://arxiv.org/abs/1602.03012
19. Wang, M., Lu, Y., Nie, J., Wang, Z., Lin, Y., Xuan, Q., & Gui, G. (2025). ReStNet: A Reusable & Stitchable Network for Dynamic Adaptation on IoT Devices. Retrieved from http://arxiv.org/abs/2506.09066
20. Xing, S., Li, P., Wang, Y., Bai, R., Wang, Y., Hu, C.-W., Qian, C., et al. (2025). Re-Align: Aligning Vision Language Models via Retrieval-Augmented Direct Preference Optimization. Retrieved from http://arxiv.org/abs/2502.13146
21. Yang, L., Zhang, R.-Y., Wang, Y., & Xie, X. (n.d.). *MMA: Multi-Modal Adapter for Vision-Language Models*. Retrieved from https://github.com/ZjjConan/Multi-Modal-Adapter
22. Yu, X., Sun, H., Ling, Z., Niu, Z., Bai, Z., Qin, R., Chen, Y.-W., et al. (2025). EPIC: Efficient Prompt Interaction for Text-Image Classification. Retrieved from http://arxiv.org/abs/2507.07415
23. Yuan, K., Srivastav, V., Navab, N., & Padoy, N. (2025). HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2405.10075
24. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J. L., Marescaux, J., Mascagni, P., Navab, N., et al. (2025). Learning Multi-modal Representations by Watching Hundreds of Surgical Video Lectures. Retrieved from http://arxiv.org/abs/2307.15220
25. Zhang, J., Barbarisi, S., Kadkhodamohammadi, A., Stoyanov, D., & Luengo, I. (2023). Self-Knowledge Distillation for Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2306.08961
26. Zhang, Q., Wei, Y., Han, Z., Fu, H., Peng, X., Deng, C., Hu, Q., et al. (2024). Multimodal

Fusion on Low-quality Data: A Comprehensive Survey. Retrieved from http://arxiv.org/abs/2404.18947

27. Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., Langlotz, C. P., Zhang, Y., Jiang, H., et al. (2022). *Contrastive Learning of Medical Visual Representations from Paired Images and Text*. *Proceedings of Machine Learning Research* (Vol. 182). Retrieved from https://github.com/yuhaozhang/convirt
28. Zhou, K., Yang, J., Loy, C. C., & Liu, Z. (2022). Learning to Prompt for Vision-Language Models. Retrieved from http://arxiv.org/abs/2109.01134
29. Zou, X., Liu, W., Wang, J., Tao, R., & Zheng, G. (2022). ARST: Auto-Regressive Surgical Transformer for Phase Recognition from Laparoscopic Videos. Retrieved from http://arxiv.org/abs/2209.01148

APPENDICES

221-35-1045

ORIGINALITY REPORT

11%	8%	7%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Midlands State University Student Paper	1%
2	arxiv.org Internet Source	1%
3	umpir.ump.edu.my Internet Source	<1%
4	Submitted to Liverpool John Moores University Student Paper	<1%
5	Submitted to NCC Education Student Paper	<1%
6	Submitted to King's College Student Paper	<1%
7	Submitted to Daffodil International University Student Paper	<1%
8	link.springer.com Internet Source	<1%
9	Twinanda, Andru Putra, 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, 2016. Publication	<1%
10	Submitted to Universiti Malaysia Pahang Student Paper	<1%

11	deepai.org Internet Source	<1%
12	mdpi-res.com Internet Source	<1%
13	"Intelligent Strategies for ICT", Springer Science and Business Media LLC, 2025 Publication	<1%
14	Submitted to University of Greenwich Student Paper	<1%
15	Fatema Tuj Johora Faria, Mukaffi Bin Moin, Zayeed Hasan, Md. Arafat Alam Khandaker, Niful Islam, Khan Md Hasib, M.F. Mridha. "MultiBanFakeDetect: Integrating advanced fusion techniques for multimodal detection of Bangla fake news in under-resourced contexts", International Journal of Information Management Data Insights, 2025 Publication	<1%
16	Laure Berti-Équille. "AI for SDGs – A technical and illustrated tour", EDP Sciences, 2025 Publication	<1%
17	download.bibis.ir Internet Source	<1%
18	Submitted to Zhejiang University Center of Modern Educational Technology Student Paper	<1%
19	avesis.yildiz.edu.tr Internet Source	<1%
20	ruj.uj.edu.pl Internet Source	<1%
21	Submitted to University of Newcastle Student Paper	<1%

22	core.ac.uk Internet Source	<1%
23	amslaurea.unibo.it Internet Source	<1%
24	d197for5662m48.cloudfront.net Internet Source	<1%
25	pure.mpg.de Internet Source	<1%
26	impa.usc.edu Internet Source	<1%
27	www.mdpi.com Internet Source	<1%
28	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2020 Publication	<1%
29	Submitted to Tennessee Board of Regents Student Paper	<1%
30	Submitted to University College London Student Paper	<1%
31	events.iist.ac.in Internet Source	<1%
32	"Computer Vision – ECCV 2022", Springer Science and Business Media LLC, 2022 Publication	<1%
33	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2023", Springer Science and Business Media LLC, 2023 Publication	<1%
34	pmc.ncbi.nlm.nih.gov Internet Source	<1%

35	Submitted to Kennesaw State University Student Paper	<1%
36	Zipei Wang, Sitian Pan, Mengjie Fang, Ruofan Zhang, Jie Tian, Di Dong. "Chapter 11 CholecMamba: A Mamba-Based Multimodal Reasoning Model forCholecystectomy Surgery", Springer Science and Business Media LLC, 2026 Publication	<1%
37	dokumen.pub Internet Source	<1%
38	Submitted to Liberty University Student Paper	<1%
39	Submitted to Lebanese International University Student Paper	<1%
40	Submitted to National Institute Of Technology, Meghalaya Student Paper	<1%
41	Submitted to University of Sheffield Student Paper	<1%
42	discovery.ucl.ac.uk Internet Source	<1%
43	ebin.pub Internet Source	<1%
44	publications.polymtl.ca Internet Source	<1%
45	"Information Processing and Network Provisioning", Springer Science and Business Media LLC, 2026 Publication	<1%

46	Benjamin Riordan, Joshua Millward, Zhen He, Dan Anderson-Luxford, Samatha Pararath Salim, Maree Patsouras, Emmanuel Kuntsche. "How to analyze visual data using zero-shot learning: An overview and tutorial.", Psychological Methods, 2025 Publication	<1 %
47	Suman Lata Tripathi, Om Prakash Kumar, Allwin Devaraj Stalin, Tanweer Ali. "Innovations in Computer Vision, Communication Systems, and Computational Intelligence - Proceedings of the First International Conference on Computer Vision, Communication System and Computational Intelligence (CVCNCE 2025), 08–09 May 2025, Tirunelveli, India", CRC Press, 2025 Publication	<1 %
48	export.arxiv.org Internet Source	<1 %
49	Submitted to Northern Kentucky University Student Paper	<1 %
50	Zhuoyang Zou, Xinghui Zhu, Qinying Zhu, Hongyan Zhang, Lei Zhu. "Disambiguity and Alignment: An Effective Multi-Modal Alignment Method for Cross-Modal Recipe Retrieval", Foods, 2024 Publication	<1 %
51	Ajay Kumar, Sangeeta Rani, Krishna Dev Kumar, Manish Jain. "Handbook of AI in Engineering Applications - Tools, Techniques, and Algorithms", CRC Press, 2025 Publication	<1 %
52	Yunlong Li, Zijian Zhao, Renbo Li, Feng Li. "Deep learning for surgical workflow analysis:	<1 %

	a survey of progresses, limitations, and trends", Artificial Intelligence Review, 2024 Publication	
53	Daniela Onita, Matei-Vasile Căpîlnaș, Adriana Baciu (Birlutiu). "Distinguishing Human- and AI-Generated Image Descriptions Using CLIP Similarity and Transformer-Based Classification", Mathematics, 2025 Publication	<1 %
54	Tiantao Liu, JIangcheng Xu, Xinke Zhan, Shaolong Lin, Shirley W. I. Siu. "Enzyformer: a Two-Stage Pretrained Model for Enzymatic Retrosynthesis", American Chemical Society (ACS), 2025 Publication	<1 %
55	Submitted to Associatie K.U.Leuven Student Paper	<1 %
56	Submitted to San Jacinto College Student Paper	<1 %
57	Tanya Buddi, Rohit Kandakatla, Rameshrao Nitin Kotkunde, Upadrasta Ramamurty, Asma Perveen. "Multi-Disciplinary Research and Sustainable Development - Proceedings of 2nd International conference on Multi-Disciplinary Research and Sustainable Development (ICMED-2025), 7th and 9th March 2025", CRC Press, 2025 Publication	<1 %
58	Submitted to Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) Student Paper	<1 %
59	Submitted to University of Bradford Student Paper	<1 %

60	Wiera Bielajewa, Michelle Tindall, Perumal Nithiarasu. "COMPARATIVE STUDY OF TRANSFORMER- AND LSTM-BASED MACHINE LEARNING METHODS FOR TRANSIENT THERMAL FIELD RECONSTRUCTION", Computational Thermal Sciences: An International Journal, 2024 Publication	<1 %
61	researchwap.net Internet Source	<1 %
62	www.southlewis.org Internet Source	<1 %
63	Arnatchai Techaviseschai, Sansiri Tarnpradab, Vasco Chibante Barroso, Phond Phunchongharn. "A Real-Time Semi-Supervised Log Anomaly Detection Framework for ALICE O2 Facilities", Applied Sciences, 2025 Publication	<1 %
64	repository.up.ac.za Internet Source	<1 %
65	xin-xia.github.io Internet Source	<1 %
66	"Medical Image Learning with Limited and Noisy Data", Springer Science and Business Media LLC, 2022 Publication	<1 %
67	Chengcheng Ma, Yang Liu, Jiankang Deng, Lingxi Xie, Weiming Dong, Changsheng Xu. "Understanding and Mitigating Overfitting in Prompt Tuning for Vision-Language Models", IEEE Transactions on Circuits and Systems for Video Technology, 2023 Publication	<1 %

68	Ding-Qiao Wang, Long-Yu Feng, Jin-Guo Ye, Jin-Gen Zou, Ying-Feng Zheng. "Accelerating the integration of ChatGPT and other large-scale AI models into biomedical research and healthcare", MedComm – Future Medicine, 2023 Publication	<1 %
69	Hoard, Michelle T.. "Promoting Negative Gaussian Curvature in Lyotropic Liquid Crystal Systems with Oleate-derived Gemini Surfactants.", University of Minnesota, 2020 Publication	<1 %
70	Islam, Md. Zahidul. "Integrating Smart Sensing and Data-Driven Decision Making Toward an Intelligent and Resilient Cyber-Physical Power System", New York University Tandon School of Engineering, 2025 Publication	<1 %
71	raw.githubusercontent.com Internet Source	<1 %
72	shura.shu.ac.uk Internet Source	<1 %
73	"OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis", Springer Science and Business Media LLC, 2018 Publication	<1 %
74	Durgesh Kumar Mishra, Nilanjan Dey, Bharat Singh Deora, Amit Joshi. "ICT for Competitive Strategies", CRC Press, 2020 Publication	<1 %

75	Keumgang Cha, Donggeun Yu, Junghoon Seo, Hyunguk Choi, Taegyun Jeon. "Pushing the Limits of Vision-Language Models in Remote Sensing without Human Annotations", The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2025 Publication	<1 %
76	Submitted to University of South Florida Student Paper	<1 %
77	Yuqin Li, Guowei Zhao, Chuqi Li, Weili Shi, Zhengang Jiang, Ziyang Zhang, Guanyuan Feng. "STSANet: Spatial Temporal-Self-Aggregation Network for surgical phase recognition", Information Fusion, 2025 Publication	<1 %
78	dr.ur.ac.rw Internet Source	<1 %
79	etheses.dur.ac.uk Internet Source	<1 %
80	public-pages-files-2025.frontiersin.org Internet Source	<1 %
81	repositories.nust.edu.pk Internet Source	<1 %
82	repository.nwu.ac.za Internet Source	<1 %
83	www.theseus.fi Internet Source	<1 %
84	"Computer Vision – ACCV 2016", Springer Science and Business Media LLC, 2017 Publication	<1 %

85	"Intelligent Information and Database Systems", Springer Science and Business Media LLC, 2020 Publication	<1 %
86	"Natural Language Processing and Information Systems", Springer Science and Business Media LLC, 2020 Publication	<1 %
87	Chi, Zhixiang. "Towards Model Adaptation at Test-Time in Open-World Environments.", University of Toronto (Canada) Publication	<1 %
88	Lasse Renz-Kiefel, Sebastian Lünse, Rene Mantke, Peter Eisert, Anna Hilsmann, Eric L. Wisotzky. "Inter-hospital transferability of AI: A case study on phase recognition in cholecystectomy", Computers in Biology and Medicine, 2025 Publication	<1 %
89	Myint Swe Khine, László Bognár, Ernest Afari. "Future of Learning with Large Language Models - Applications and Research in Education", CRC Press, 2025 Publication	<1 %
90	Pethuru Raj, B. Sundaravadivazhagan, V. Kavitha, B. Narendra Kumar Rao, Hannah Vijaykumar. "Real-Time Artificial Intelligence (AI) - Key Motivations, Technologies, Platforms, and Use Cases", Apple Academic Press, 2026 Publication	<1 %
91	Saleem Ramadan, Mohammad Abu-Shams, Sameer Al-Dahidi, Ibrahim Odeh, Najat Almasarwah. "A data-driven approach for	<1 %

predicting remaining intra-surgical time and enhancing operating room efficiency", Journal of Industrial Engineering and Management, 2025
Publication

92	Witold Abramowicz, Marek Kowalkiewicz, Krzysztof Węcel. "AI-Driven Digital Transformation - Perspectives from a Business School", Routledge, 2025 Publication	<1 %
93	Yunfan Li, Himanshu Gupta, Prateek Prasanna, IV Ramakrishnan, Haibin Ling. "Surgical Phase Recognition in Laparoscopic Cholecystectomy", Procedia Computer Science, 2024 Publication	<1 %
94	Yuxin Peng, Zhaoda Ye, Jinwei Qi, Yunkan Zhuo. "Unsupervised Visual-Textual Correlation Learning With Fine-Grained Semantic Alignment", IEEE Transactions on Cybernetics, 2020 Publication	<1 %
95	assets.amazon.science Internet Source	<1 %
96	d-nb.info Internet Source	<1 %
97	epjdatascience.springeropen.com Internet Source	<1 %
98	files01.core.ac.uk Internet Source	<1 %
99	github.com Internet Source	<1 %

hal.science

100	Internet Source	<1 %
101	ijirt.org Internet Source	<1 %
102	indah.ump.edu.my Internet Source	<1 %
103	psasir.upm.edu.my Internet Source	<1 %
104	ulspace.ul.ac.za Internet Source	<1 %
105	www.lrec-conf.org Internet Source	<1 %
106	www2.mdpi.com Internet Source	<1 %
107	Mark D. Shermis, Joshua Wilson. "The Routledge International Handbook of Automated Essay Evaluation", Routledge, 2024 Publication	<1 %
108	Pushpa Choudhary, Sambit Satpathy, Arvind Dagur, Dhirendra Kumar Shukla. "Recent Trends in Intelligent Computing and Communication", CRC Press, 2025 Publication	<1 %
109	Qin Zhu, Zuzhi Jiang, Matt Thomson, Zev Gartner. "Revealing a coherent cell state landscape across single cell datasets with CONCORD", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
110	boris.unibe.ch Internet Source	<1 %

111	"Medical Image Computing and Computer-Assisted Intervention – MICCAI 2017", Springer Science and Business Media LLC, 2017 Publication	<1 %
112	Praveen SR Konduri, G Siva Nageswara Rao. "Surgical phase recognition in laparoscopic videos using gated capsule autoencoder model", Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2023 Publication	<1 %
113	Qin Zhu, Zuzhi Jiang, Matt Thomson, Zev J. Gartner. "Revealing a coherent cell state landscape across single cell datasets with CONCORD", Cold Spring Harbor Laboratory, 2025 Publication	<1 %
114	Yu Liu, Rui Xie, Lifeng Wang, Hongpeng Liu, Chen Liu, Yimin Zhao, Shizhu Bai, Wenyong Liu. "Fully automatic AI segmentation of oral surgery-related tissues based on cone beam computed tomography images", International Journal of Oral Science, 2024 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		