Analyzing the Behavior of Compact Transformer Encoders in Fixed Prompt Multimodal Learning

GAWSHIL RAHMAN RIFAT

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSIT

APPROVAL

This thesis titled on "**Analyzing the Behavior of Compact Transformer Encoders in Fixed Prompt Multimodal Learning**", submitted by **Gawshil Rahman Rifat (ID: 221-35-900)** 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	:	Gawshil Rahman Rifat
Date of Birth	:	25 August 2002
Title	:	Analyzing the Behavior of Compact Transformer Encoders in Fixed Prompt Multimodal Learning
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:

(Student's Signature)

221-35-900

Student ID

Date:27-12-2025

(Supervisor's Signature)

Mr. Musabbir Hasan Sammak

Name of Supervisor

Date: 27-12-25

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.

(Student's Signature)

Full Name : Gawshil Rahman Rifat

ID Number : 221-35-900

Date : 27-12-2025

Analyzing the Behavior of Compact Transformer Encoders in Fixed Prompt Multimodal Learning

GAWSHIL RAHMAN RIFAT

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

To all patients whose lives depend on safe and efficient surgery, and to the surgeons, nurses, and operating room teams who strive every day to deliver better care.

ABSTRACT

Multi-modal learning systems are increasingly being conditioned on pretrained text encoders to condition visual representations, although the behavioral implications of text encoder size and depth under fixed-prompt and low-data conditions remain poorly understood.The paper discusses the behavior of compact transformer encoders with a fixed prompt multi-modal learning system, and in surgery phase recognition as a controlled assessment task in seven stages including Preparation, Calot Triangle Dissection, Clipping Cutting, Gallbladder Dissection, Gallbladder Packaging, Cleaning Coagulation and Gallbladder Retraction.We analyze the issue of whether text size compact pretrained encoders (MiniLM-L3, MiniLM-L6, MiniLM-L12, and DistilBERT) can be helpful in preserving multimodal alignment in conditions of textual inflexibility. To make the behavior of text encoders isolable, it is frozen and only the text encoder (a lightweight 512-dimensional projection head) and a trainable temperature are learned in a symmetric contrastive (InfoNCE) objective. The dataset utilized in experiments is Cholec80 (80 videos), the frame rate is used, 1 FPS, phase prompts are fixed, pre-processing is light standardized, and the train/validation/test splits are video-wise. To measure the performance of the models and the confusion matrix analysis to understand the behaviour of per-phase alignment, the top-k (Top-1/5/10) accuracy based on the image-to-text nearest-prompt classification is applied. The results show that Top-1 is much more accurate in Encoder depths: MiniLM-L3 (approximately 44%), DistilBERT (approximately 39-40) and MiniLM-L6 (approximately 39) have the highest accuracy, and the accuracy of deeper MiniLM-L12 is much lower (approximately 24-25%). Despite these differences, Top-5 (approximately9496) and near-perfect (approximately100) accuracies of all models are strong indicating that correct prompts tend to be in close semantic similarity. Interestingly, encoders with higher levels of sensitivity are more sensitive to limited supervision but low sensitivity encoders are more stable to fixed-prompt constraints. These findings highlight the point that in cases where depth of the transformers is increased multimodal alignment is not invariably supported in low-data fixed-prompt cases. Instead, smaller encoders can perhaps act more strongly and reliably, offers useful empirical guidance to the choice of encoder of text in restricted multimodal learning situations. The limitations of the study are the unequal representation of classes in the database, visual overlap of certain stages of surgery, and the failure to refine the vision encoder.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

$\boldsymbol{x}$	Input image/frame
$\boldsymbol{t}$	Phase prompt (text)
$\{\boldsymbol{w_i}\}_{\boldsymbol{i=1}}^{\boldsymbol{L}}$	Token sequence
$\mathbf{h}_{\boldsymbol{i}}$	Token embedding
$\mathbf{z}_{\text{text}}$	Sentence/text embedding
$\boldsymbol{d}$	Shared embedding dimension
$\boldsymbol{d_v}$	Visual feature dimension
$\boldsymbol{d_t}$	Text feature dimension
$\mathbf{v_0}$	Raw visual feature
$\mathbf{P}_{\text{img}}$	Visual projection matrix
$\tilde{\mathbf{v}}$	Projected visual feature
$\mathbf{v}$	Normalized visual feature
$\tilde{\mathbf{u}}$	Projected text feature
$\mathbf{u}$	Normalized text feature
$\mathbf{W_1}, \mathbf{W_2}$	Projection head weights
$\mathbf{b_1}, \mathbf{b_2}$	Projection head biases
$\boldsymbol{\sigma}(\cdot)$	Nonlinearity (e.g., GeLU)
$\boldsymbol{\phi}(\cdot)$	Final activation (often identity)
$\boldsymbol{\alpha}$	Log-temperature
$\boldsymbol{\tau}$	Temperature
$\boldsymbol{s_{ij}}$	Cosine similarity (i–j)
$\mathbf{S}$	Similarity matrix
$\mathbf{L}$	Scaled logits matrix
$\boldsymbol{\mathcal{L}}_{\text{i}\to\text{t}}$	Image→Text loss
$\boldsymbol{\mathcal{L}}_{\text{t}\to\text{i}}$	Text→Image loss
$\boldsymbol{\mathcal{L}}$	Symmetric contrastive loss
$\boldsymbol{N}$	Batch size (effective)
$\hat{\boldsymbol{y}}_{\boldsymbol{i}}^{(\mathbf{1}:\boldsymbol{k})}$	Top-k predicted labels
$\boldsymbol{y_i}$	True label
$\mathbb{I}(\cdot)$	Indicator function

C_{ij}	Confusion matrix entry
$\hat{k}(x)$	Predicted phase index

LIST OF ABBREVIATIONS

SPR	Surgical Phase Recognition
OR	Operating Room
VLM	Vision–Language Model
VLP	Vision–Language Pretraining
CLIP	Contrastive Language–Image Pretraining
ViT	Vision Transformer
MiniLM	Minimal Language Model (distilled BERT family)
DistilBERT	Distilled Bidirectional Encoder Representations from Transformers
MLP	Multi-Layer Perceptron
InfoNCE	Information Noise-Contrastive Estimation (contrastive loss)
R@k	Recall at k (Top-k accuracy)
Top-1	Top-1 Accuracy
AMP	Automatic Mixed Precision
AdamW	Adam with decoupled Weight Decay
LR	Learning Rate
WD	Weight Decay
TP	True Positive
FP	False Positive
TN	True Negative
FN	False Negative
F1	F1-score (harmonic mean of Precision & Recall)
HMM	Hidden Markov Model
CRF	Conditional Random Field
TCN	Temporal Convolutional Network
MS-TCN	Multi-Stage Temporal Convolutional Network
U-Net	U-shaped Convolutional Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
BPE	Byte-Pair Encoding (tokenization)
CLS	Classification token (Transformer)

GPU	Graphics Processing Unit
CPU	Central Processing Unit
I/O	Input/Output
FPS	Frames Per Second
Cholec80	Laparoscopic cholecystectomy dataset (80 videos, 7 phases)
M2CAI	MICCAI “Workflow” challenge dataset
Cataract-101	Cataract surgery video dataset
GP-VLS	General-Purpose Vision–Language model for Surgery
HecVL	Hierarchical Video–Language pretraining (zero-shot SPR)
Endo-CLIP	Endoscopy-adapted CLIP variant
SurgLaVi	Surgical Large-Scale Vision–Language Dataset
BLIP	Bootstrapping Language–Image Pretraining
BLIP-2	BLIP-2 (frozen image encoder + LLM)
LLaVA	Large Language and Vision Assistant
ViLT	Vision-and-Language Transformer (minimal visual pipeline)
ALIGN	A Large-scale ImaGe and Noisy-text embedding model

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

Multimodal structures Multimodal learning structures are being progressively developed to incorporate visual representations in addition to an English language to facilitate the recognition, retrieval and reasoning processes. One such use is the Endoscopic video analysis as a form of surgical phase recognition (SPR), or a medical video recognition of the stage of procedure in progress in the video, which is an automated medical technique of identification. More lately, textual supervision Vision-language models have been shown to learn procedures, and can directly learn to perform zero-shot or weakly-supervised phase recognition at natural language prompts. Although this paradigm does improve the size of the multimodal systems, it is also associated with the problem of consistency of evaluation, reproduction and execution of the model with less supervision (Yuan, Srivastav, Navab, & Padoy, 2025a).

The relevance of SPR in the current work lacks a locus of application but an experimental test site on the investigation of the multimodal alignment behavior. Our dataset is named Cholec80, and consists of 80 laparoscopic cholecystectomy videos and annotated with seven stages of surgery, i.e., Preparation, Calot Triangle Dissection, Clipping and Cutting, Gallbladder Dissection, Gallbladder Packaging, Cleaning and Coagulation and Gallbladder Retraction. The dataset is highly imbalanced in classes since there is high level of longer time of given phases than others (e.g. Calot Triangle Dissection and Gallbladder Dissection). Original 25-30 FPS is subsampled to 1 FPS that is also common and also contributes to the biases of imbalance. The differences between the measures of data splitting and testing regimes of previous researchers have made the comparisons between the available findings directly complicated (Funke, Rivoir, Krell, & Speidel, 2025).

Multimodal approaches that were recently introduced are fundamentally a cosine similarity loss with temperature scaling, over an image encoder and a text encoder (a mapping of image inputs and text inputs into the same embedding space) (Radford et al., 2021)(Zhang et al., 2022). This architecture has been demonstrated to be effective, and has good zero-shot and few-shot transfer properties, although the capacity and depth of the text encoder, which is generally an opaque component, has a potent impact on its behavior.

As a matter of fact, especially in edge to resource constrained systems, giant text transformers impose memory as well as latency and thermal constraints, which can negatively affect the stability of the system. Freezing vision encoder, lightweight projection heads and temperature parameter tuning techniques have been demonstrated to freeze at significantly lower computation price (Roy et al., 2024). These observations lead to an even more basic question: what are the properties of compact transformer encoders when learning multimodally is performed in fixed-prompt and low-data circumstances?

The gap has implications on this research because it is a systematic investigation of the behavior of small pretrained text encoders; MiniLM-L3, MiniLM-L6, MiniLM-L12, and DistilBERT in the context of learning in a fixed-prompt multimodal learning setup. A frozen CLIP ViT image encoder is used in order to isolate the effect of text encoders, only the text encoder, a shallow projection head, and learnable temperature parameter are trained with a symmetric contrastive loss. The benefit of such carefully controlled combination is that the encoder depth, stability, and alignment behavior can be made to make an apples-to-apples comparison, and that it is not bound to vision encoder adaptations (Funke, Rivoir, & Speidel, 2023).

1.2 Problem Statement

The contemporary models of vision-language (VLMs) such as the CLIP-based one have been demonstrated to be characterized by a high quality of zero-shot and few-shot transfer performance in a large range of tasks. However, these models are typically founded on large text encoders based on transformers, the depth, and the number of parameters of the encoder that have a substantial impact on the multimodal alignment behavior. These encoders may have great representational capacity, but they also have great computational overhead and complexity and the action they produce even with limited supervision is difficult to interpret. Moreover, the existence of literature is largely due to the fact that the encoder of the text is a fixed system, and there has been very little research done to define how the size and depth of the text encoder affects the multimodal learning (Radford et al., 2021).

A major question that is yet to be answered is whether small pretrained text encoders such as MiniLM-L3, MiniLM-L6, MiniLM-L12, and DistilBERT can work effectively at multimodal alignment when texts are supervised exclusively with text. In particular, it is not quite clear how such encoders would act in case of multimodal learning performed in the context of fixed-prompt conditions under which linguistic variation is eliminated and performance depends heavily on the ability of the encoder to retrieve and compare semantic substance. To answer this question one must divide the contribution of text pathway and hold the other components constant(Zhang et al., 2022).

To allow an analysis that is based on fair and controlled analysis, this study employs an apples to apples experimental protocol in order to isolate the behaviour of text encoder. The CLIP ViT encoder of the images is kept fixed and only the text encoder which is a lightweight projection head into a common 512-dimensional embedding space and a learnable temperature parameter is optimized using a symmetric contrastive (InfoNCE) loss (Funke et al., 2023).This design eliminates confounding variables that are dominant in the prior surgical phase recognition work, such as the unbalanced domain of dataset partitioning, evaluation leakage, or metric difference (Kostiuchik et al., 2024).

Within this organized approach, the multimodal alignment is quantified as image-to-text retrieval (Recall@1, Recall@5, Recall@10) and fixed-prompt nearest-neighbor classification accuracy, as well as a few parameters-effort parameters, including the count of parameters and throughput. The methods can be used to relatively rank the size-accuracy-stability trade-offs of compact text encoders, trained in a similar environment (Kostiuchik et al., 2024).

Even though the multimodal and attention-based architecture have gained more and more popularity, they do not give any special analysis to the text pathway to examine the behavior of compact encoders in the fixed-prompt and low-data multimodal learning conditions. The issue on whether there is a greater alignment of deeper text encoders and whether supervision stability or instability are created by greater capacity even with limited supervision is not resolved. This study seeks to address this area of knowledge gap with empirical evidence concerning the behavior of text encoders that can be utilized to provide a recommendation on what models to apply in sparse multimodal learning environments, and use of surgical phase recognition as an example of evaluation task (Yuan et al., 2025a).

1.3 Motivation

The role of pretrained text encoders in conditional visual representations provided by multimodal systems of learning has not been studied sufficiently but conditioning such encoders during behavior is increasingly relying on pretrained text encoders. Most vision-language models often assume that the deeper the semantic grounding provided by the text encoder, the bigger the depth and capacity of the encoder is. However, this assumption can hardly be tested when the circumstances are limited by the fact that the linguistic variability is also limited or the size of the training data is also limited. Their behavior in such settings should be known so as to have reliable and interpretable multimodal systems (Roy et al., 2024).

An accumulating literature in contrastive vision-language learning has also shown that strong multimodal correspondence is possible when a range of simplifications of architecture, such as vision encoder freezing, lightweight and well-tuned temperature parameter projection heads are used in the contrastive objective (Roy et al., 2024). Despite the interest in this observation, most existing multimodal surgical systems focus on the quality of the alignment

that is determined according to the stability of the representation instead of the magnitude of the model itself, or the mediating role of the text pathway in this relationship (Yuan et al., 2025a).

Recent emergence of surgical vision-language pretraining including hierarchical and general-purpose models demonstrates that prompt-based zero shot learning is effective in procedures. Such systems are however confounding the effects of depth and capacity of the text encoder alongside the simultaneous change in the vision backbone, temporal modeling and the magnitude of training data. It is therefore not definite that better language representations lead to better performance, or that there was another consideration of architecture and training (Yuan et al., 2025a).

It is based on this gap that this paper uses a reductionist experimental configuration whereby CLIP ViT image encoder is frozen and the text encoder replaced systematically with variants of compact transformers (MiniLM-L3, MiniLM-L6, MiniLM-L12 and DistilBERT). The paper is able to directly measure the joint effect of encoder depth and multimodal alignment behavior by simply parameterizing the model by a combination of predefined textual prompts and only optimization of the text pathway and projection head parameter. This artificial setting offers the identification of surgical phases as a diagnostics problem of evaluating the representational power, steadiness and overfitting predisposition of weak supervision (Yuan et al., 2025a).

Finally, not the least, when multimodal learning is performed under the same preprocessing and evaluation conditions, there is little empirical advice available on how to trade off encoder capacity, alignment quality, and stability. The previous research has also revealed that differences in the partitions of datasets, measurement and reporting processes obscure the comparisons that can be undertaken meaningfully across the models(Funke et al., 2023). The study will attempt to explain the behavioral trade-offs of compact transformer encoders through a traditional apples-to-apples comparison, but will have principled implications beyond just one domain of interest and could be applied more generally to assist in making decisions about the use of text encoders in limited multimodal learning applications (Kostiuchik et al., 2024).

1.4 Significance of the Study

The study provides a controlled and methodical platform of testing behavior of tiny transformer encoders in fixed-prompt multimodal study.The proposed research design disconnects the role of the text encoder to learning cross-modal representations by freezing CLIP ViT image encoder and assessing multimodal alignment using the text route alone (Radford et al., 2021).The isolation is necessary to the interpretation of the dependence of the encoder depth and capacity on alignment behavior when linguistic supervision is limited, and an example is in low-data and prompt-limited learning (Zhang et al., 2022).

The most important of the works is the apples-to-apples with smaller pretrained text encoders, i.e., MiniLM-L3, MiniLM-L6, MiniLM-L12, and DistilBERT, which are trained under the same preprocessing condition, training schedules, and evaluation conditions as a CLIP-Text reference. This solves a comparability issue that has been present in previous multimodal and surgical phase recognition experiments in which splits, augmentation strategies and measures of evaluation varied across studies and thus it was hard to interpret their results (Funke et al., 2023).

The specific analysis presented in the paper also incorporates both measures of alignment (Recall1/5/10 and fixed-prompt nearest-neighbor Top-1 accuracy) and efficiency-related ones (such as the number of parameters and throughput). Such measurements are used to put size-accuracy-stability trade-offs into perspective, not by viewing it as a deployment problem but to put in perspective new best practice regarding contrastive vision-language benchmarking (Roy et al., 2024).

In addition to that, the outcomes demonstrate that multimodal alignment can be conducted with a very thin architectural dish, i.e. a lightweight projection head into a shared embedding space alongside a learnable temperature parameter, and no activities in the encoder of the vision. This finding is consistent with the theoretical results of the contrastive learning task and other past medical vision language tasks where fixed-prompt representational stability and calibration can be a determinant as scale-invariant as scale-invariant as model scale (Zhang et al., 2022).

Lastly, the rigor of the methodology and reproducibility is highlighted by the paper because it provides a program of the experiment, partition of data fixed, CLIP-consistent, preprocessing of the experiment, default optimization parameter, were run with seeds, hardware and runtime

environment report.The research and its result are valuable because they offer empirical reference to the past research on evaluation drift that can be reused and replicated in studies of multimodal learning and text encoder analysis in future (Funke et al., 2023).

1.5 Research Questions

Comparisons between small pretrained text encoders (in terms of size) (MiniLM-L3/L6/L12, DistilBERT) in effectivity in multimodal surgical phase detection with low-data performance, and what is the effect of encoder depth/size on down-stream classification results?

1.6 Research Objectives

- To encode a fixed-prompt condition to different text encoders, which are lightweight.
- To find the impact of the encoder depth on the minimum-data low-data performance. text conditions.
- To compare the encoder behaviors so as to determine the most appropriate models.
- To provide appropriate suggestions on a selection of effective text encoders in the medical arena multimodal AI.

1.7 Research Scope and Limitations

This section highlights the limitations of the study and the limitations arising from the dataset, chosen techniques, and evaluation process.

1.7.1 Scope

1. The only laparoscopic cholecystectomy videos we used in this paper were Cholec80: 80 laparoscopic cholecystectomy videos, with seven phases of annotations (Preparation; Calot Triangle Dissection; Clipping Cutting; Gallbladder Dissection; Cleaning Coagulation). They down sample images at 1 FPS , preprocess fixed (and re-size/crop like in the CLIP pipelines and ImageNet mean-std CLIP normalization). We divided into train/val/test (no leakage of patient/video) by video. We examine class imbalance (where some classes e.g., have longer dwell time in steps such as CalotTriangle Dissection or Gallbladder Dissection are over-represented) and where one can, apply naive countermeasures - e.g. uniform sampling of the phases, by class - which will describe all the selections vividly(Kostiuchik et al., 2024).
2. The CLIP ViT image encoder is fixed, and the only difference between MiniLM-L3/L6/L12, DistilBERT and an image baseline CLIP-Text is the text encoder. A thin projection head is used to project text properties to 512-D (CLIP space), and is also learned, and with an additional learnable temperature (τ) on the similarity scale during contrastive training(Radford et al., 2021).
3. To learn image-text and text-image features with constant hyperparameter (epochs, learning rate, weight decay, batch size) we firstly train all encoders with the same symmetric InfoNCE loss. We do not eliminate the augmentations and report whether gradient clipping and AMP are present or not. Reproductivity Runs are seeded(Radford et al., 2021).
4. These are alignment-Recall 1/5/10 (image-text retrieval) and zero-shot phase-prompt Top-1) and efficiency, parameters (M), throughput (images/s) and wall-clock time in seconds to key milestones, optional, but measured when rationale, is GPU memory. All encoders have the metric execution in order to compare models on Cholec80 imbalanced set(Funke et al., 2023).
5. We specify hardware, library version, checkpoints and random seeds of the tokenizer. Output will be provided in CSV logs and plot (e.g. accuracy vs parameters; accuracy vs throughput) so that it can be re-used and so that later analysis is possible. This is based on the community suggestions of transparent and reproducible SPR and contrastive VLM benchmarking(Roy et al., 2024).

1.7.2 Limitations

1. Its result is based on the Cholec80 dataset and has not undergone other processes or institute approval(Kostiuchik et al., 2024).
2. We have not optimized the visual text path- fine-tuning the CLIP ViT model only optimized the text path, which is not always as good as a pipeline model that is tuned fully(Hoque, Hasan, Emon, Khalifa, & Rahman, 2024).
3. Video recognition of the full gallery and long-time temporal reasoning. The paper only describes specialized temporal models (e.g., MS-TCN/link, temporal U-Nets), which are not experimentally tested(Park, Oh, Jeong, & Yu, 2023)(Funke et al., 2025).
4. The experimental research utilized English staged prompts testing and no cross-lingual prompts testing without studying the breadth of prompt engineering(Yuan et al., 2025a)(Perez, Nwoye, Kermani, Mohareri, & Jamal, 2025).
5. We have observed the fact that throughput and wall-clock time are dependent on hardware, and hence we should construct a correct system and relative benchmarking in this environmental context(Funke et al., 2023).
6. We learn our models in a relative sense of accuracy, efficiency, and trade-off budget, and not absolute SOTA prioritization on the Cholec80 dataset(Funke et al., 2023)(Roy et al., 2024).

1.8 Thesis Organization

Introduction: Introduction Presents problem context and motivation States problem and gaps in research formulates problem and objectives Defines scope and limitations Summarizes contributions

Background and Related Work: describes Surgical phase recognition (SPR) and workflow analysis, the Cholec80 dataset and the problem of class-imbalance in it, contrastive vision-text learning (e.g., CLIP-style dual encoders, InfoNCE, temperature scaling), medical adaptations (e.g., ConVIRT), surgery-aware VLP (e.g HecVL, SurgLaVi, GP -VLS), and evaluation techniques/metrics of SPR. This chapter reveals definition and terms and contextualizes the starting point of the study in literature of accuracy-efficiency.

Methodology: description of the experimental set up: data pipeline (Cholec80, 1 FPS samples, CLIP preprocess., and video-wise splits), model architecture (frozen CLIP ViT, lightweight text encoders MiniLM-L3/L6/L12, DistilBERT and base line with CLIP-Text) projection head of dimension 512-D, learnable temperature, symmetric infonce, fixed hyperparameters/schedules, augmentations, seeds and any imbalance mitigation (uniform sampling/loss reweighting).

Experiments and Results: Tables: alignment (R@top1,5,10): summary, efficiency (parameters, throughput, wall-clock; (memory) optional). It achieves this through the production of combined comparative tabular/plots of all text encoders with the same settings along with sensitivity analysis (pooling, prompts, batch size) showing effects size and variability amongst seeds. Gives examples of retrieval outside the curve and analysis of error due to imbalance in the phases.

Discussion and Conclusion: Gives results interpretations on: the accuracy-efficiency trade-offs, stabilization by projection.This paper gives results interpretations on the following: on the what was lost/gained relative to CLIP-Text; clinical implications/engineering implications to near-real-time deployment; and scientific implications to the text pathway under a frozen vision tower. They include a discussion of weaknesses (single big / dataset, frame-level, English prompts only, hardware independent), future work (multiple procedure / datasets, temporal / modelling, joint fine-tuning multilingual prompting / off-machine optimization).

CHAPTER 2

LITERATURE REVIEW

2.1 Background & Problem Context

Surgical phase recognition Surgical workflow analysis Surgical phase recognition Automatic identification of procedural phases of endoscopic video on two levels frame level (one label/frame) or segment level (continuous frames with the same label) is one of the first steps in surgical workflow analysis(Funke et al., 2023)(Kostiuchik et al., 2024). It provides workflow (phase schedules, performance and variability) intraoperative decision support (phase-oriented/phase-warned) reducing errors and quality assurance (objective auditing, standards meeting), training/education (phase-sensitive feedback and assessment)(Kostiuchik et al., 2024). The models are expected to be able to allocate across a change of distribution since the actual OR deployments involve surgeons, devices (scopes and resolutions) and domains (institutions, case-mix) otherwise no sensible comparison can be made without a practice of evaluations ontologically(Funke et al., 2023). Visual cues can be under-discriminative (identical when used in other situations using the same tools or views), and may be trained with multimodal training with language semantic context phases, tools and actions - to separate look-alike frames and could become prompting with zero/few-shot behaviour by training against a vision-language loss (e.g. CLIP on natural images, ConVIRT on medical images), and to future surgery-sensitive video-language systems (HecVL, SurgLaVi, GP-VLS)(Radford et al., 2021).

More recent vision-language models can be of either of two broad types, dual encoders, in which the visual and textual data are separately encoded and aligned at a common space, such as contrastive learning (e.g. CLIP); or encoder-decoder or instruction-tuned stacks, where the generation is conditioned by an input visual signal (e.g. BLIP/BLIP-2, LLaVA)(Radford et al., 2021). Other contrastively trained dual encoders, like CLIP (and its successor developed with ALIGN) are trained on a joint embedded space, and also learn or sample a temperature to maximize similarities, scales to web-scale data and can be queried with a text query to do zero-/few-shot recognition(Roy et al., 2024). Encoder-decoder models have likewise represented the retrieval and survived to grounded generation and thought when dealing with

images by vision encoder coupled with language model bound to light-weight adapters (BLIP-2), follow-the-prompt behavior through instruction tuning (LLaVA)(Hoque et al., 2024). ViLT shows another step closer step that includes the patches and tokens with the help of only one transformer at the expense of strong and heavy CNN backbones and expense of cross-modal attention between token representations(Roy et al., 2024). In these works, there are also sharing ideas: similar space (to access or condition), conflicting goals (or cross entropy between image-text pairs), temperature scaling to regulate the logits as well as prompting them in natural language to encode them(Radford et al., 2021). General-purpose surgical VLMs Video-language pretraining General-purpose video-based sensing and understanding tasks are a more recent area of research and recent studies have used task phase recognition by directly basing on textual anchors such as phase names, instruments or actions to pretrain VTL(Yuan et al., 2025a).

The benchmark datasets superimpose development of SPR. The frame-wise label of seven stages (Preparation, CalotTriangle Dissection, Clipping Cutting, Gallbladder Dissection, Gallbladder Packaging, Cleaning Coagulation, Gallbladder Retraction) and 80 videos of laparoscopic cholecystectomy make up Cholec80(Kostiuchik et al., 2024). M2CAI workflow benchmark (laparoscopic cholecystectomy) M2CAI workflow benchmark (laparoscopic cholecystectomy) is a frame-based phase annotation benchmark, and has been used most commonly in the challenge-based efforts at comparing across studies(Kostiuchik et al., 2024). Microscope-video Cataract-101collection relates to cataract surgery, and is phase/step (and, in most cases, generally) annotated at frame/segment granularity and generalizes between cholecystectomy in order to observe the differences in modality between laparoscopic and ophthalmic scope(Kostiuchik et al., 2024). Vision-only (unimodal) Older Systems Vision only systems consisted of per-frame C CNN classifiers, CNN + LSTMGRU sequence models and temporal convolutional (e.g. TCNMS-TCN) methods with HMM/CRF smoothing to add temporal coherence, but with no prior text or language knowledge(Park et al., 2023)(Funke et al., 2025). They are not only effective at within-dataset, they also have severe weaknesses, they lack (weak) semantics (they are not based on a language), they are expensive to run in real-time and can be scaled only to, and cannot be scaled to unseen procedures or queries previously(Funke et al., 2023)(Kostiuchik et al., 2024). Under these constraints, the emergent multimodal work was already in operation working in line with video with text e.g. hierarchical video-language pretraining to zero-shot phase recognition and early general-

purpose surgical VLMs/datasets, but the extent to which the data modalities could be stabilized under fair conditions was not known(Yuan et al., 2025a)(He et al., 2025)(Schmidgall, Cho, Zakka, & Hiesinger, 2024a).

2.2 CLIP and the Text Modality

Transformer encoder A CLIP-style dual encoder represents a text tower, which encodes a tokenized prompt linearly to the space of the image encoder to match it against(Radford et al., 2021)(Roy et al., 2024). The following design levers have been shown to influence this tower: the hidden size and depth (determining semantic capacity and long-range context), the vocabulary/tokenizer (determining the fragmentation of domain terms, such as instruments, phase names, etc.), and the positional encoding form (determining how orders of words are modelled) - the design levers directly influence both latency and memory footprint (activations KV caches) and hence, the quality of text-image alignment under InfoNCE can be learnt with a learnable temperature(Radford et al., 2021)(Hoque et al., 2024). In this paper, we substitute the CLIP-Text by light-weight transformers (MiniLM-L3/L6/12, DistilBERT), without modifying the standard CLIP image tower (512-D) or contrastive objective that is symmetric, which simply freezes the ViT encoder image tower and only trains text pathway, and a very narrow projection, thereby, making any improvement in alignment or performance merely owed to the text encoder(Radford et al., 2021)(Roy et al., 2024).

Visual encoder Pre-train on large-scale images-text data sets: visual backbone models RN50 and ViT-B /16, text tower models Transformer based, optimization InfoNCE learnable-temperature, image-text retrieval (Recall at K) and zero-shot top-1 image -text performance with prompt templates. Linear heads The heads are linear, the main inferred task is on frozen models with optional fine-tuning to the task; Efficiency Metrics Backbone scale and batch computation are efficiency metrics(Radford et al., 2021).

Recall@K and zero / few-shot performance on matched conditions and a large search of architecture tuning processes (prompt engineering, freezing vs. fine-tuning) and throughput / parameters trade-offs - they can be interpreted as text tower scale-performance trade-offs and domain transfer resilience can be discussed on a massive scale(Roy et al., 2024).

In order to achieve endoscopy based on CLIP-type contrastive learning, sequence of frames are broken into text labels/prompts (stage/instrument/procedure words). We have now a ViT-type CLIP image encoder, the text Transformer tower, which uses temperature-scaled InfoNCE, and in the form of retrieval recall at K, and no-shot accuracy with prompts. The model structure has thick links, as customarily believed in (stage/instrument strings) custom prompts, but the linear projection, and most methods used to freeze the image tower. Some measures of efficiency (parameters, latency) are also acquired, yet they are also readily more in common field than the CLIP model(He et al., 2025).

General Surgical Vision Language Model CLIP Style Dual Encoder of Diversity of Surgery Scenes: Introducing ViT-like Visual Backbone, Transformer Text Tower, InfoNCE+ Temperature Mechanism and Reporting Recall@ and Zero-Shot Phase/Tool Top-1 Metrics Under Prompt Variants. The paper examines domain shift and considers the methods of prompting and domain text augmentation methods to resolve the domain shift problem. The majority of the previous researches optimized the model by freezing a portion of the layers of the tower and lightweight projection heads. Problems of magnitude of parameters, qualitative latency at deployment are explained in some closely related works(Schmidgall et al., 2024a).

Since one holds a frozen visual tower, any decisions made in the CLIP central route (text encoder - scale/capacity, prompt engineering and projection/temperature) have a proportional influence on a zero-shot domain transfer retrieval/performance. Good prompts and stable training When the informative prompt is good enough and there is no catastrophic collapse / forgetting patterns during fine-tuning, then we can freeze ViT using a lightweight text path, and this is what we are attempting to accomplish in our frozen ViT with replaceable text experiments(Radford et al., 2021)(Schmidgall et al., 2024a)(He et al., 2025)(Roy et al., 2024).

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

Paper	Method	Approach	Text encoder used	Dataset
HecVL (video–language pretraining)(Yuan, Srivastav, Navab, & Padoy, 2025b)	CLIP-style dual encoders + hierarchical video–text pretrain	Frozen/partial-frozen ViT, InfoNCE + τ, prompt-based zero-shot phases	CLIP-style Transformer (text encoder)	Cholec80
GP-VLS(general-purpose surgical VLM)(Schmidgall, Cho, Zakka, & Hiesinger, 2024b)	CLIP-style dual encoders across mixed surgical corpora	ViT image + linear projection, temperature scaling, prompt variants	CLIP-style Transformer (text encoder)	Cholec80
Endo-CLIP (endoscopy adaptation)(He et al., 2025)	Contrastive image–text training (CLIP variant)	Frozen ViT, linear head, hand-crafted phase/tool prompts, InfoNCE + τ	CLIP-style Transformer (text encoder)	Cholec80
Adapting Generalist VLM for SPR(Faray De Paiva, Yuan, Srivastav, & Padoy, n.d.)	Transfer of CLIP-family model to phases	Frozen image tower; fine-tune text/head; retrieval + zero-shot via prompts	CLIP-style Transformer (text encoder)	Cholec80

2.3 Limitations & Motivation for Lightweight CLIP Text Towers

Though small and whatever-large variations of CLIP-Text have also been proposed to resolve such latency/memory problems, the overall forward pass by CLIP-Text of CC is not the best characterization to either the embedded/edge deployment, or real-time OR loop feedback(Radford et al., 2021). Moreover, CLIP is also trained on text on web-scale, which can lead to incompatibility of linguistic priors in CLIP with surgical jargon (e.g. the steps of a procedure, equipment, and movements), triggering zero-shot grounding without domain adaptation or prompting(Kostiuchik et al., 2024)(Hoque et al., 2024). Timeliness is also a key factor in performance, such prompts as phase of surgery is non-standard and the results are thus quite susceptible to phrasing decisions of the natural language, and the inter-study reproducibility is poor(Roy et al., 2024)(Funke et al., 2023). CLIP generates frame-level embeddings automatically which do not depend on any temporal signal, and thus cannot learn procedure dynamics without sequence modeling (i.e. MS-TCN or video-level pretraining)(Park et al., 2023)(Funke et al., 2025). Lastly, alignment and efficiency (parameters, throughput, wall clock) are not reported collectively in the majority of articles, and little evidence of real-time performance has been published, which added to the fact that almost no experiment of hypothesis can be undertaken in the OR environment until now makes arguments about deployability in the real world even more burdensome to develop(Schmidgall et al., 2024a)(Funke et al., 2023)(Roy et al., 2024)(He et al., 2025).

2.4 Research Focus & Evaluation Plan

Freeze vision tower CLIP ViT and replace CLIP-Text with MiniLM-L3/L6/L12 and DistilBERT and add a thin projection that projects text embedding into CLIP space 512-D and is trained on the known symmetric InfoNCE objective at an learnable temperature where any change of the alignment is unrealistically during text encoder representation(Radford et al., 2021)(Roy et al., 2024). we report Recall1510 Image-Text retrieval and Zero-shot phase prompt Top-1 and performance measures such as the number of parameters, throughput (images/s)(Funke et al., 2023)(Roy et al., 2024). We call upon other, more relevant medical contrastive adaptations, which can be used to humanize design decisions in the clinical space (e.g. contrastive training with bidirectional losses), neither do we draw our apples-to-apples comparisons on a fixed protocol(Zhang et al., 2022)(Kostiuchik et al., 2024).

2.5 Summery & Literature Review

Multimodal model used in this work is a CLIP-style dual-encoder model comprising of a frozen vision encoder and a learnable text pathway that shares an embedding space. To be more specific, the image encoder is a fixed CLIP ViT model that takes input frames and transforms them into 512-dimensional visual representations, whereas the text pathway uses small pretrained transformer encoders (MiniLM-L3, MiniLM-L6, MiniLM-L12, or DistilBERT). The text representations are projected to the same space of dimension 512 with a lightweight projection head and aligned with visual features with a temperature-scaled symmetric contrastive (InfoNCE) objective (Radford et al., 2021).

The information flow in this model is based on 1 FPS frame sampling and CLIP-stable image preprocessing, image visual feature extraction and parallel tokenization of fixed textual prompts, which can be seen as canonical surgical stages. Cosine similarity is calculated between image and text embeddings in a normalized form and training of these measurements motivates that image-prompt pairs are closer in the embedding space than mismatched pairs. Multimodal prediction during inference is done through a nearest-prompt decision rule whereby each frame is compared with all fixed prompts and a nearest prompt is picked. This design allows the multimodal alignment behavior to be evaluated in a controlled way without bringing with it any linguistic variability (Roy et al., 2024).

The contrastive vision-language models, such as CLIP, and their variants have shown effective zero-shot and few-shot performance in a wide range of domains due to their use of shared embedding space and temperature-scaled alignment. A number of limitations however are evident when such models are introduced to constrained multimodal settings. To begin with, the capacity and depth of text encoder is a significant factor that influences the latency, memory, and optimization dynamics. Second, CLIP has been trained on big web data and as such may experience semantic mismatch with domain-specific and fixed prompts common when dealing with medical or procedural data. Other recent researchers have established that the performance of a model can be very sensitive to the timeliness of its formulation, whereas reporting tends to exclude the analysis of accuracy, efficiency, and stability jointly, and hinders reproducibility and interpretability (Kostiuchik et al., 2024).

Late advances in medical vision-language learning have been on mass video-language pretraining and hierarchical multimodal representations to enhance zero-shot generalization. Though these methods are shown to achieve very high performance increases, they usually require concurrent modifications of vision backbones, temporal modeling, and scale of data and it is hard to separate the influence of the text encoder alone. Therefore, the behavioral value of compact transformer encoders in the conditions of fixed-prompt and low-data are not sufficiently investigated (Schmidgall et al., 2024a).

To fill this gap, the given study will have a controlled experimental design, that is, it will freeze the CLIP ViT image encoder and replace the text encoder with compact transformer variants in a systematic fashion. The experiments are all done on the Cholec80 dataset with the same preprocessing, video-wise partitions, constant phase prompts, and training programs. The temperature parameter, projection head, and the text encoder are optimized together on a symmetric InfoNCE objective, and performance metrics are measured in alignment metrics (Recall@K), fixed-prompt Top-1 classification accuracy, and efficiency-related measures, such as the number of parameters and throughput. Random seeds and hardware specifications are clearly reported to make it reproducible.

This study presents a narrow analysis of the effects of the encoder depth and the capacity on the multimodal alignment behavior with limited learning conditions by formulating the surgical phase recognition as a diagnostic standard and not an end use. This view builds upon literature with empirical evidence of stability and performance of fixed-prompt multimodal learning using compact transformer encoders, which augments previous studies that could focus on end-to-end performance and large-scale pretraining (Radford et al., 2021).

CHAPTER 3

METHODOLOGY

We start first by processing raw laparoscopic videos with time-stamped annotations of vision, before converting them to per-video phase files that are frame and phase-indexed i.e. Frame/Phase. The frames are then sampled with a crude rate of 1FPS and rescaled to cfg. Image size and stored into a clean per-video file format, to enable fast and portability of access (hexa shrink Rates/{8 or 32). All the generators are generated on top of the Preprocessed Frame Dataset which sort out the labelling of the frames in execution and at the load time normalizes CLIP-style to make fair comparisons. In order to avoid leakage we provide three train/validation/test splits (fixed random seed), indicate the actual video IDs in the appendix and include the per-splits video/ frame counts per class, datapoint statistics and phase-duration histograms showing the current class- imbalance.

Image processing Image preprocessing Image preprocessing Image preprocessing is performed using the canonical make clip tensor trans form that performs size-normalization CLIP normalization. The input data is in text form and it is motivated by a hard written dictionary, DEFAULTPHASEPROMPTS (seven Cholec80 phase prompts) that is available in the appendix in its raw form in order to be readable. They prepare the encoders by training them with their associated default tokenizers in the training loop, however, most importantly, in testing how similar or different two encoders are, the same prompt strings are presented to both encoders, and hence, only differences between encoders are compared and not wording effects.

Our two-encoder model and pre-trained CLIP ViT-B/16 image tower are not trainable and our text tower can be changed between MiniLM-L3/L6/L12 (and DistilBERT in both). An adjustable temperature ( ) magnifies the components of the correlations to regulate the acuity of the SoftMax in contrastive learning. The architecture section contains a description of projections, logits normalization, logits similarity and temperature-scaled logits.

The symmetric InfoNCE loss is an average of the image-text cross-entropy loss and the text-image cross-entropy loss, which is minimized, and we look at a (N) in-batch negatives that can be used to boost the retrieval signal in-batch without necessarily having to use external hard-negatives which are mined.

The optimization is performed with AdamW, a single constant learning rate (3x10-4), and weight decay (1x10-4) on all the trainable (text tower, projection head and temperature) is performed. We do not take 3 epochs, 3 epochs and batch size of 8 (clip-norm = 1.0). Python/NumPY/PyTorch is always concerned to ensure that splits and other mini-batches are chosen in a way that is reproducible. The loss (current (tau) and Recall@) on validation Log loss is saved at the end of each epoch and the checkpoint with the highest validation R1 is saved to be used in downstream experiments.

The used InfoNCE equation (appendix) reinforcement was also complete with 3-epoch/batch-8 schedule TBST Checkpointing with resume support also enabled. We test on a NVIDIA GeForce RTX 3050 Ti Laptop GPU that has the ability to provide the throughput needed to our frozen-vision swappable-text pipeline and also to provide realistic edge/OR HW impairments.

The measurement of alignment applied when the formal definition is applied to a ranked prompt list is recall at 1/5/10 (Top-k). In identifying imbalance-based classification detection report, we also give Precision, Recall also F1 macro-averaged and per-phase levels. A confusion matrix (representing systematic confusions between visually or semantically similar phases) is also graphically illustrated (through correct/incorrect assignment) to help in improvements of prompting, augmentation, or reweighting.

3.1 Data Collection

Purpose:

This section describes the Cholec80 data whose focal point is the study that is done in this thesis. The information provides the premises of exploring surgical phase recognition (SPR), that is, during the laparoscopic cholecystectomy operations. The fact that this experiment can be contextualized and that the methods introduced in this publication can be used to comprehend the nature of the dataset that was utilized, the difficulties, and the purpose of the data choice is significant.

Dataset Source:

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:

Videos Headcount: The dataset has 80 laparoscopic cholecystectomy videos all of which are complete procedure videos. The videos will also play an important role in training the SPR models since they will be fully informed of each step of the surgery.

Number of Phases: Cholec80 data set is identified with 7 phases of surgery:

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

These phases entail the most significant parts of a laparoscopic cholecystectomy and hence the data is ideal to the surgical phase recognition action.

Tool Annotations: The tool annotations are also generated at 1 frame per second (fps) which fact implies the availability of seven tools, that is, the grasper, bipolar, hook, scissors, clipper, irrigator, and the specimen bag. These annotations may also inform the phase recognition models, especially when there is an effort at mapping of the tools with specific surgical tasks.

Video Length: The videos of cholec80 dataset are of mean video length of 38 minutes with a standard deviation of 16 minutes.

Frames: Frame rate was taken to 25 fps which give approximately 1000 frames in a video and each frame is identified by a phase which it belongs to. This frame extraction can be properly analyzed and frame-level annotations can be used to train the model.

Phase Durations:

The table 1 below shows all 7 surgical phases of the Cholec80 dataset and their mean time in the 80 videos. The values are given in seconds alongside the 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 eras demonstrate the discrepancies of the different stages in which some phases like the Triangle Dissection by Calot and the Dissection of the Gallbladder are much longer than others like the Gallbladder Retraction. The period of time is to be mentioned in the case of the real time phase detection training models.

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 to be 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 phases 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 multi modals 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.

3.2 Data Preprocessing

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.

3.3 Model Architecture

We adopt a **CLIP-aligned dual-encoder** design with a **frozen image tower** and a **swappable lightweight text tower**, coupled in a shared embedding space and trained with a **temperature-scaled symmetric InfoNCE** objective.

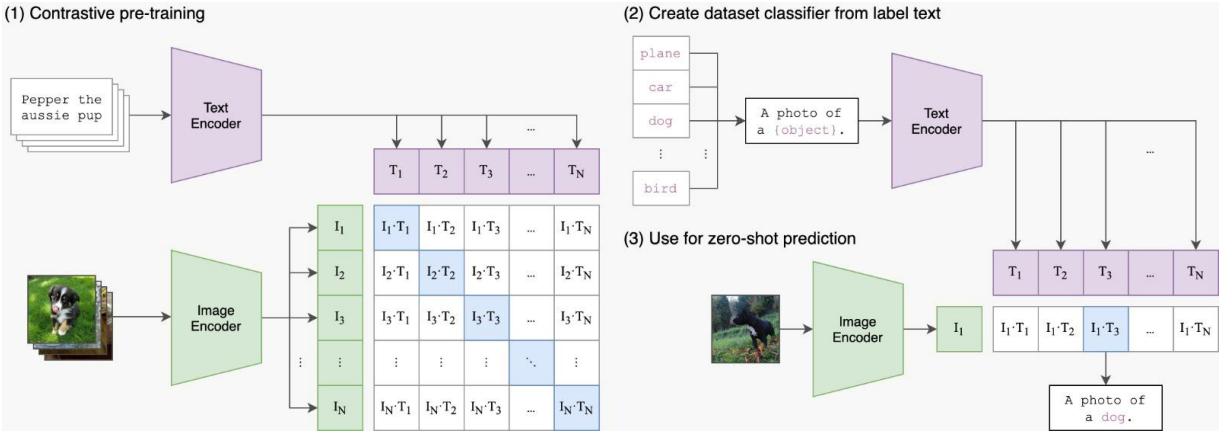

Figure 3.1 Clip Architecture

3.3.1 Encoders and projections

Image (frozen CLIP ViT)
Let, $x \in \mathbb{R}^{H\times W\times 3}$ be a preprocessed frame (CLIP normalization). The frozen vision tower produces a visual feature:

$$\mathbf{v}_0 \;=\; f_{\text{img}}(x) \in \mathbb{R}^{d_v}$$

If , $d_v \neq d$ (the target shared dimension, d=d = CFG.emb_dim = 512 we apply a (fixed or trainable) linear projection $\mathbf{P}_{\text{img}} \in \mathbb{R}^{d\times d_v}$:

$$\tilde{\mathbf{v}} \;=\; \mathbf{P}_{\text{img}}\,\mathbf{v}_0 \in \mathbb{R}^{d}$$

Finally, we L2-normalize (CLIP practice):

$$\mathbf{v} \;=\; \frac{\tilde{\mathbf{v}}}{\|\,\tilde{\mathbf{v}}\,\|_2} \in \mathbb{S}^{d-1}.$$

Text (swappable, lightweight)

Given with the input prompt string t , the text tower (MiniLM-L3/L6/L12 or DistilBERT) is tokenized to $\{w_i\}_{i=1}^{L}$ and encodes to token embeddings $\{\mathbf{h}_i\}_{i=1}^{L}$, $\mathbf{h}_i \in \mathbb{R}^{d_t}$. We construct a sentence embedding using either [CLS] pooling or mean pooling:

$$\mathbf{z}_{\text{text}} = \begin{cases} \mathbf{h}_{[\text{CLS}]} & \text{(CLS pooling)} \\ \frac{1}{L}\sum_{i=1}^{L} \mathbf{h}_i & \text{(mean pooling).} \end{cases}$$

A lightweight projection head (**TinyMLPHead**) maps to the shared space:

$$\tilde{\mathbf{u}} = \phi(\mathbf{W}_2\, \sigma(\mathbf{W}_1\, \mathbf{z}_{\text{text}} + \mathbf{b}_1) + \mathbf{b}_2) \in \mathbb{R}^d,$$

where $\sigma(\cdot)$ is a pointwise nonlinearity (e.g., GeLU) and $\phi(\cdot)$ can be identity (yielding a linear head) or a mild nonlinearity; in our implementation the head is intentionally tiny to minimize latency/params. We L2-normalize:

$$\mathbf{u} = \frac{\tilde{\mathbf{u}}}{\| \tilde{\mathbf{u}} \|_2} \in \mathbb{S}^{d-1}.$$

Trainable components: The text encoder, TinyMLPHead (text path), and a temperature parameter are trainable; the CLIP ViT image tower is frozen**.**

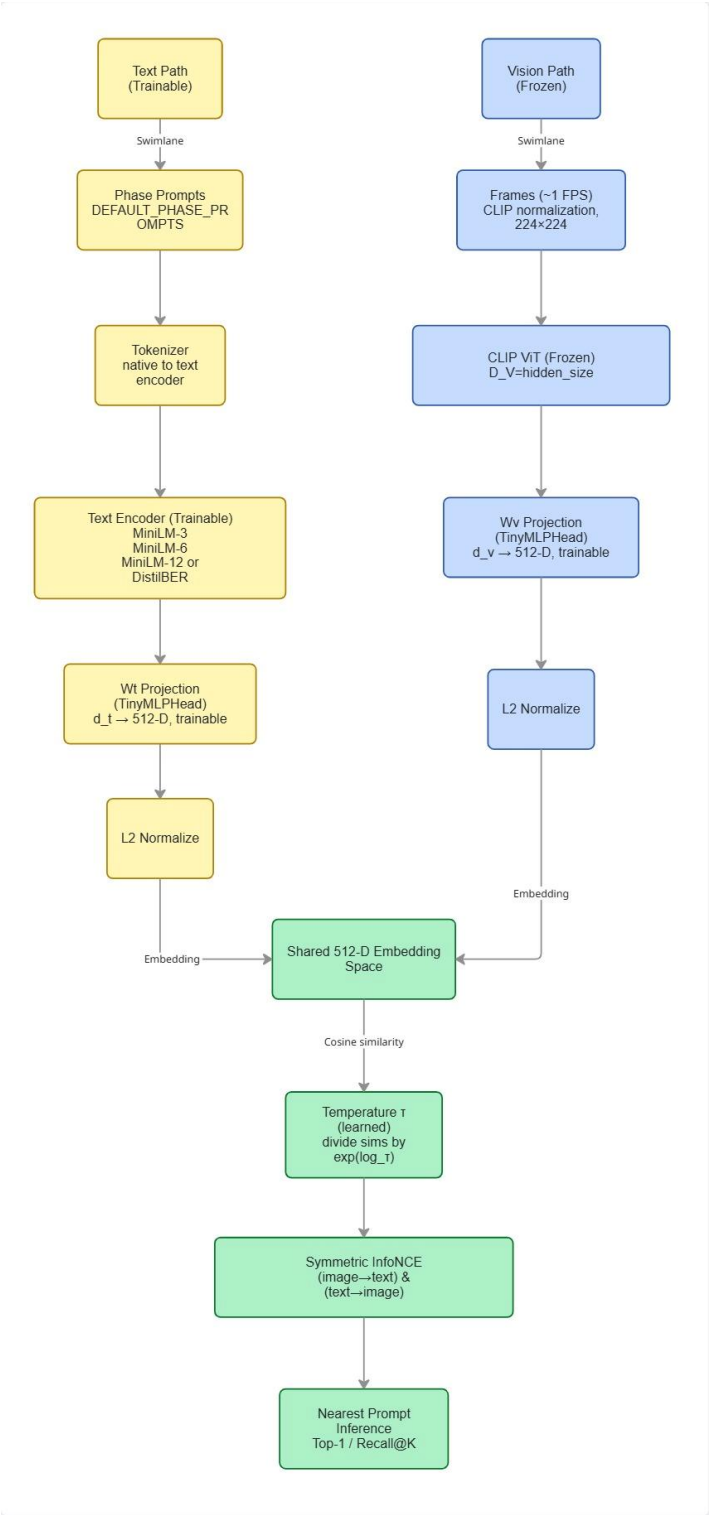

Figure 3.2 Architecture Pipeline

Shared embedding space & projection heads

- **Target dimensionality:** Each of the comparisons is done in a common-space of dimension CFG. embmm = 512 (CLIP default).
- **Text projection:** The projection of each text encoder hidden dimension (d n) to 512-D space, which uses a light projection head TinyMLPHead (din, dout=CFG. embdim). To reduce the overhead and provide enough capacity to be aligned to an enormous range of text spaces (MiniLM/DistilBERT vs CLIP-Text), small (e.g., linear or small MLP) head has been selected.
- **Visual projection (conditional):** When frozen vision encoder generates features whose dV not is not equal to 512 a visual projection in the composite model would be used to take it to CFG. embdim. As a matter of fact CLIP ViT-B/16 already generates 512-D and thus this branch usually does not matter; it is just to have a clean architecture and to be able to swap architectures in the future.
- **Normalization:** to normalize image and text on projection i.e. cosine similarity = dot product and stabilizing contrastive softmax.

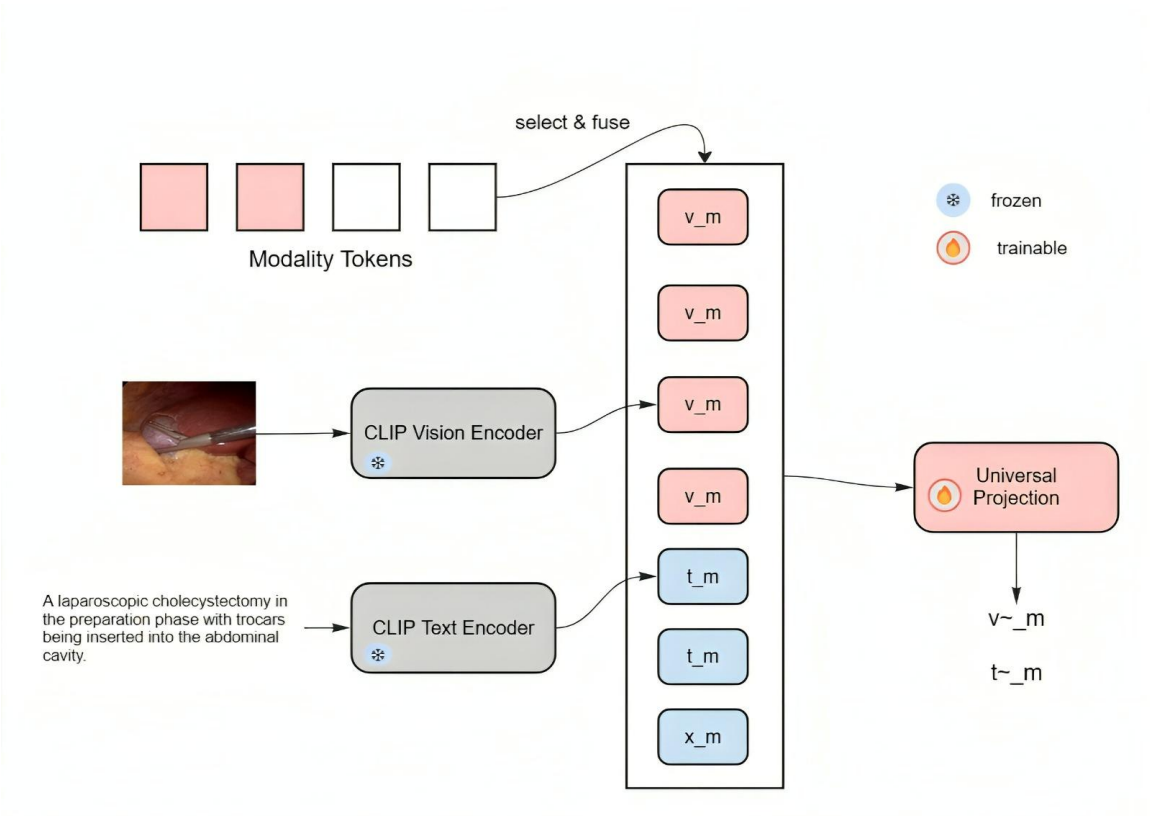

Figure 3.3 Linear Projection

3.3.2 Similarity and temperature scaling

For a minibatch of N frames and N prompts (paired by index), we compute the cosine-similarity matrix in the shared space:

$$\mathbf{S}_{ij} = \mathbf{v}_i^{\top}\mathbf{u}_j \in [-1,1].$$

We follow CLIP and learn **log-temperature** $\alpha = \log \tau$ (scalar), using $\tau = e^{\alpha}\>0$. The scaled logits are:

$$\mathbf{L} = \frac{\mathbf{S}}{\tau} = \frac{\mathbf{S}}{e^{\alpha}}.$$

Temperature modulates the **softmax sharpness**, a key factor in retrieval alignment and training stability.

3.3.3 Symmetric InfoNCE (bidirectional) objective

Image→Text loss (each image's positive is its paired prompt):

$$\mathcal{L}_{\text{i2t}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\mathbf{L}_{i,i})}{\sum_{j=1}^{N} \exp(\mathbf{L}_{i,j})}.$$

Text→Image loss (each prompt's positive is its paired image):

$$\mathcal{L}_{\text{t2i}} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(\mathbf{L}_{j,j})}{\sum_{i=1}^{N} \exp(\mathbf{L}_{i,j})}.$$

The **symmetric contrastive loss** is their average:

$$\mathcal{L} = \frac{1}{2}(\mathcal{L}_{\text{i2t}} + \mathcal{L}_{\text{t2i}}).$$

This bidirectional InfoNCE mirrors CLIP and empirically improves gradient balance across modalities.

3.3.4 Inference (nearest-prompt decision rule)

Given a frame x and the set of seven phase prompts $\{t_k\}_{k=1}^{7}$, we compute normalized embeddings $\mathbf{v}(x)$ and $\mathbf{u}(t_k)$, then select:

$$\hat{k}(x) \;=\; \arg\max_{k\in\{1,\dots,7\}} \; \mathbf{v}(x)^{\top}\mathbf{u}(t_k).$$

This yields **promptable zero-shot** phase classification consistent with the retrieval objective.

3.3.5 Interfaces

- **Vision tower (frozen):** Configured as CFG.IMAGE_ENCODER = "openai/clip-vit-base-patch16". Features exposed via an image-feature interface (conceptually, get_img_feats).
- **Text towers (swappable):** CFG.TEXT_ENCODER_LIST = {MiniLM-L3, MiniLM-L6, MiniLM-L12} (CLIP-Text for reference). Tokenization is **encoder-native**; features via a text-feature interface (conceptually, get_txt_feats).
- **Composite model:** A **Text Alignment Model** wraps the frozen image encoder and a chosen text encoder with projections to the **shared d=512d=512d=512 space**, and exposes **encode_image** / **encode_text** for clean forward paths.
- **Temperature:** Learnable $\alpha = \log \tau$ applied to the similarity logits.

3.3.6 Design levers and their effects

- **Pooling (CLS vs mean):** Affects the geometry and stability of $\mathbf{z}_{\text{text}}$; small encoders sometimes favor **mean pooling** for robustness, whereas larger encoders exploit **CLS** semantics.
- **Hidden size & depth (MiniLM-3/6/12):** Control representational capacity vs latency /memory. Better/broader more expensive; more semantics; more deep, more, our thin head would assist in compressing between different spaces of text and CLIP visual space.
- **Tokenizer/vocabulary:** Native tokenization encoding, which ensures that domain vocabulary (e.g. Calot Triangle Dissection) is also encodings) prompts are identical across encoders, and models are only allowed to vary in how they encode prompts (fairness rule).

- **Projection head size**: TinyMLPHead (not deep adapters) provides minimal flexibility just to accommodate text/vision space misalignment at maintaining efficiency - of critical interest to OR edge constraints.

3.4 Training Setup

Loss Function:

The contrastive loss function is used to train the visual-textual model that is in charge of aligning the multimodal features of the text and image encoders. To be more specific, our approach utilizes the InfoNCE loss as the means of calculating the similarity between the normalized image and text embedding. The following is the contrastive loss formula:

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

Where:

- $\text{sim}(z_v, z_t)$ is the cosine similarity between the image z_v and text z_t embeddings.
- τ is a learnable temperature parameter that scales the similarity score.
- N is the total number of samples in the batch, with the sum in the denominator accounting for all possible image-text pairs.

This loss causes the similarity of image-text pairs located close to each other in the common embedding space and the dissimilar pairs are separated.

Optimizer:

The optimization of the model is done with the AdamW optimizer that builds on Adam and adds a weight decay regularization. Such choice results in optimal optimization particularly in large models. The optimal parameters are determined with the help of the following configuration:

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

These parameters are useful to ensure that the learning process is controlled to prevent overfitting especially in large scale data.

Learning Rate and Weight Decay:

The learning rate is set to 3×10^{-4} which is a typical value of transformer based models (to ensure stable convergence). A decay of 1×10^{-4} is also applied to discourage large weights and discourages overfitting.

With this configuration, there is no explicit learning rate scheduling, although we have the advantage of the learning rate decay that was added to AdamW.

Epochs and Batch Size:

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

The batch size is chosen as large as possible that fits in memory on the GPU and provides a good rate of convergence and speed of training. We hope that the number of epochs that we take is also 3 to be able to learn models without overfitting on the costly to compute Cholec80 dataset.

Gradient Clipping:

The result is clipped to ensure that the gradient is not too large and the gradient will explode during back propagation. Specifically the magnitude of gradients was limited by gradient norm to a limit 1.0, and appears to have a stabilizing influence on training and a benefit us effect on model convergence

Checkpointing and Incremental Training:

We have a checkpointing scheme, which facilitates retraining in partial steps and model recovery in case of long training execution. The state of a model is stored at the completion of every epoch in such a way that the most appropriate (judged by the evaluation metrics) can be recovered to be utilized further. The check-pointing system is already installed:

- At the end of each epoch, the trained model will write out its weights and the best model will be written to best fusion model .pt.
- The continuation of the training process may proceed based on the last model checkpoint saved, instead of restarting, and the result will not be any strangled computations.

- As more information becomes available or more fine-tuning is needed, the model can be retrained in small steps using the existing checkpoints which facilitate long-term optimization as additional data is produced.

It is particularly helpful in deploying the model in the real-time surgical setting where it is anticipated that the model will be updated frequently as more and more videos will be made in the operating room and different phases / data outliers will be identified.

Hardware:

All the models are trained using a NVIDIA GeForce RTX 3050 Ti Laptop GPU due to the enormous size of CLIP models and that is where we are able to effectively train multimodal fusion. This GPU setup ensures that training proceeds with reasonable speed while maintaining the model's accuracy across multiple training iterations. This setup of GPU allows a good training rate and at the same time allows the model to maintain its accuracy by undergoing multiple training sessions.

3.5 Evaluation Metrics

Purpose: The primary evaluation measurements are stated and discussed to measure the work of the model in surgical phase recognition (SPR). These steps quantify the effectiveness of the model re-rise in placing phase, imbalance of classes and displaying errors patterns to guide better.

3.5.1 Top-k Accuracy / Recall@k (R@k)

Top-k accuracy The correct phase of the top k estimates of each sample. Top-1 (R@1), Top-5 (R@5) and Top-T0 (R@10).

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

Where:

- N is the number of test samples,
- y_i is the true label for sample iii,
- $\hat{y}_i^{(1:k)}$ is the set of top- k predicted labels for sample i
- $\mathbb{I}(\cdot)$is the indicator function (1 if the true label is within the top- k, else 0).

Interpretation

- **R@1 (Top-1):** measures if the top prediction is correct.
- **R@5 (Top-5):** measures if the correct phase appears within the top five.
- **R@10 (Top-10):** measures if the correct phase appears within the top ten.

This information can be used by SPR where the visual confusion may cause the near-misses that place the actual phase in the lead over the random one.

3.5.2 Precision, Recall, and F1-score

In order to evaluate the quality of classification, we compute precision, recall and F1-score, macro-averaged (i.e. comparing the phases with each other equally) and per-phase (to show class-specific results).

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

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

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

- **Macro-average:** Compute the metric, of each of the labels, and their unweighted average. (averaging does not take into account whether a label is imbalanced or not)
- **Per-phase:** report measures phase by phase, which phases (e.g. rare or temporary) are more difficult to find.

3.5.3 Per-Phase Metrics

Phase wise precision/recall/F1 provides an analytical view to the performance. They can be used in the discovery of these underrepresented stages that are either over-missed (poor recall) or over-predicted (poor precision) and direct selective data augmentation, reweighting, or timely refinements.

3.5.4 Confusion Matrix

The confusion matrix $\mathbf{C}$ summarizes correct and incorrect predictions across phases:

$$C_{ij} = \text{number of samples of true class } i \text{ predicted as class } j.$$

Here, C_{ij} counts how often ground-truth phase i is classified as j. From $\mathbf{C}$ we can derive accuracy, precision, recall, and F1 per phase, and inspect which phases are most frequently confused (e.g., semantically or visually similar steps). This analysis informs further improvements in prompting, augmentation, or handling of class imbalance.

3.6 Summery of Methodology

- **Problem framing:** We study text-path efficiency in CLIP-style surgical phase recognition by holding the **vision tower fixed** and swapping in lightweight text encoders to test whether alignment quality can be preserved at lower compute.
- **Data & preprocessing:** We use Cholec80 only. Videos are sampled at ~1 FPS, resized to CFG.img_size, and normalized with CLIP-style transforms**.** Human timestamps are converted to frame-indexed phase files (Frame, Phase). Frames and labels are aligned at load time. Class imbalance is documented and, where noted, mitigated with simple strategies (e.g., oversampling/reweighting). Splits are video-wise **with a** fixed seed; exact video IDs are listed in the appendix. We report per-split counts (videos/frames), per-phase counts, and phase-duration histograms.
- **Prompts & tokenization:** A fixed dictionary DEFAULT_PHASE_PROMPTS (seven canonical phases) is used unchanged across all encoders (fairness rule). Each encoder uses its native tokenizer; tokenization occurs inside the training loop.

- **Model architecture**:
    - Vision tower: CLIP ViT-B/16 **(**frozen**).**
    - Text towers (swappable): MiniLM-L3/L6/**L12** (and optionally DistilBERT/CLIP-Text for reference).
    - Projection: a TinyMLPHead maps text features to the shared 512-D space; an image-side projection is available if needed to match dimensions (typically a no-op for ViT-B/16).
    - Normalization & similarity: L2-normalized embeddings; cosine similarity in the shared space.
    - **Temperature:** a learnable scalar τ (stored as log-τ) scales similarities.
- **Training objective:** Symmetric InfoNCE (CLIP loss) over mini-batches: we average image→text and text→image cross-entropy terms using in-batch negatives. Batch size **N** yields **N−1** negatives per anchor. Stability uses global-norm gradient clipping and mixed precision (AMP)**.**
- **Optimization & hyperparameters (used settings):** AdamW (β_1=0.9, β_2=0.999, ε=1e−8), constant LR = 3×10^{-4} for all trainables, weight decay = 1×10^{-4}**,** no LR scheduler**,** 3 epochs**,** batch size = 8**,** clip-norm = 1.0. Seeds fixed across Python/NumPy/PyTorch; reproducible splits and batches.
- **Evaluation protocol:**
    - **Alignment**: Recall@1/5/10 (image↔text retrieval) and zero-shot phase Top-1 via nearest-prompt decision rule.
    - **Efficiency:** parameters (M)**,** throughput (images/s), and wall-clock time (s) to best-val.
    - **Diagnostics:** per-phase precision/recall/F1 and confusion matrix to analyze error patterns under class imbalance.
- **Reporting & reproducibility:** We publish CSV logs, best-validation checkpoints, hardware specs (GPU/VRAM, CPU/RAM), library/tokenizer versions, seeds, and **exact split IDs**. Figures/tables relate **accuracy to parameters** and accuracy to throughput, enabling apples-to-apples comparisons across text encoders.

- **Scope control:** The vision tower is always frozen; only the **text encoder**, text projection head**,** and **τ** are trained. This isolates the role of the text pathway in maintaining CLIP-level alignment for surgical phase prompts under realistic, edge-oriented constraints.

CHAPTER 4

RESULTS

4.1 Overview of Results

This chapter evaluates lightweight text encoders paired with a frozen CLIP ViT on Cholec80 using frame-level prompts. Headline accuracy shows MiniLM-L3 with the highest Top-1 (44%)**,** DistilBERT and MiniLM-L6 near ~39–40%**,** and MiniLM-L12 lowest **(**24–25%**)** (see Fig. 4.1**).** Despite this spread, models converge at broader neighborhoods**—**Top-5 94–96% and Top-10 100%**—**indicating that the correct phase prompt is usually among the top candidates even when the first choice differs.

Per-class analyses reveal a common tendency to over-predict Gallbladder Dissection, especially for uncertain frames, while the hardest boundary is between Clipping Cutting and Gallbladder Dissection due to overlapping tools and tissue context. MiniLM-L3 attains high Top-1 largely via frequent assignments to Gallbladder Dissection**;** MiniLM-L6 reduces this bias and improves recognition of Preparation and Clipping Cutting; MiniLM-L12 provides the fairest per-class behavior with a cleaner separation between Clipping Cutting and Gallbladder Dissection but at a Top-1 cost; DistilBERT maintains competitive Top-1 primarily through over-assignment to Gallbladder Dissection.

Overall, the results point to strong neighborhood alignment across encoders and an accuracy–balance trade-off governed by final selection: models differ less in retrieving the right prompt neighborhood and more in how decisively they choose among adjacent surgical stages.

4.2 Overall Top-k accuracy

Top-1: MiniLM-L3 attains the highest Top-1 **(**44%), followed by DistilBERT (39–40%**)** and MiniLM-L6 (**39%**); MiniLM-L12 is lowest **(**24–25%).

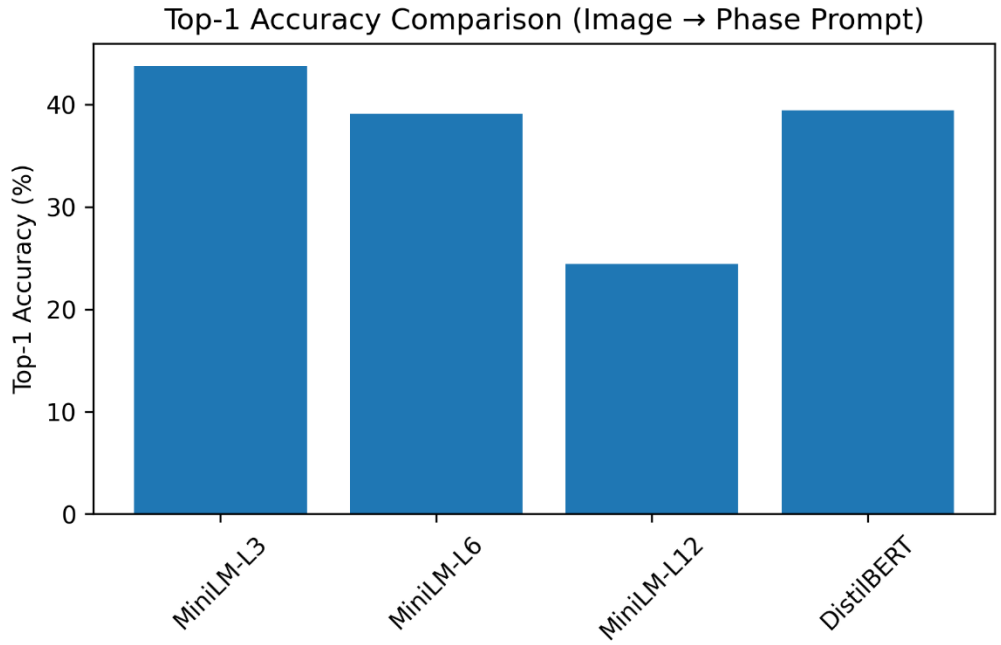

Figure 4.1 Top-1 Accuracy Comparison

Top-5 / Top-10: All models converge at Top-5 94–96% and Top-10 100%, indicating the correct phase prompt is usually in the immediate neighborhood even when Top-1 differs.

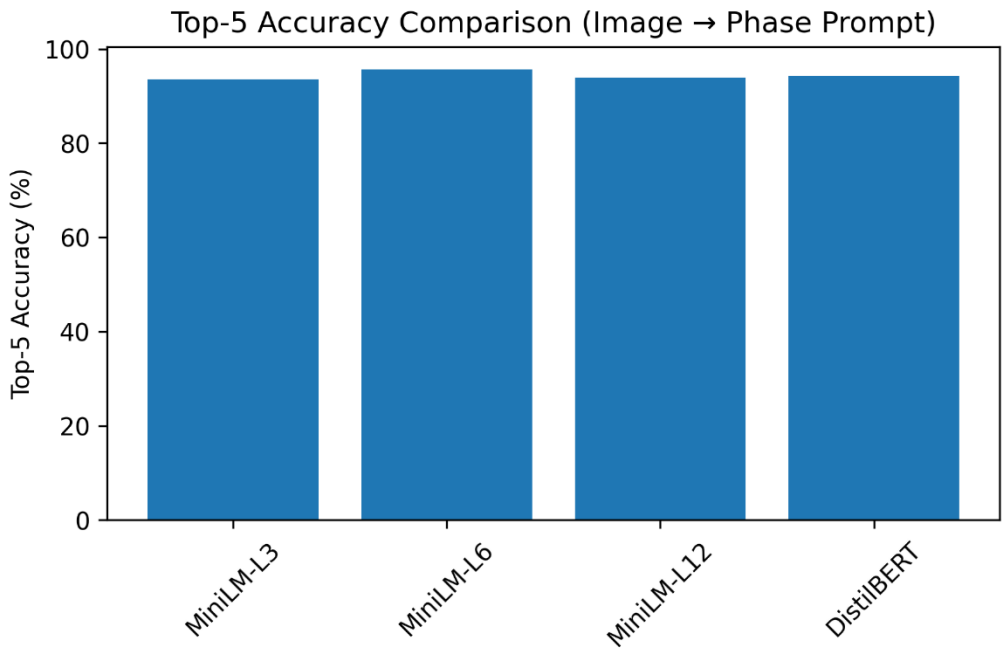

Figure 4.2 Top 5 Accuracy Comparison

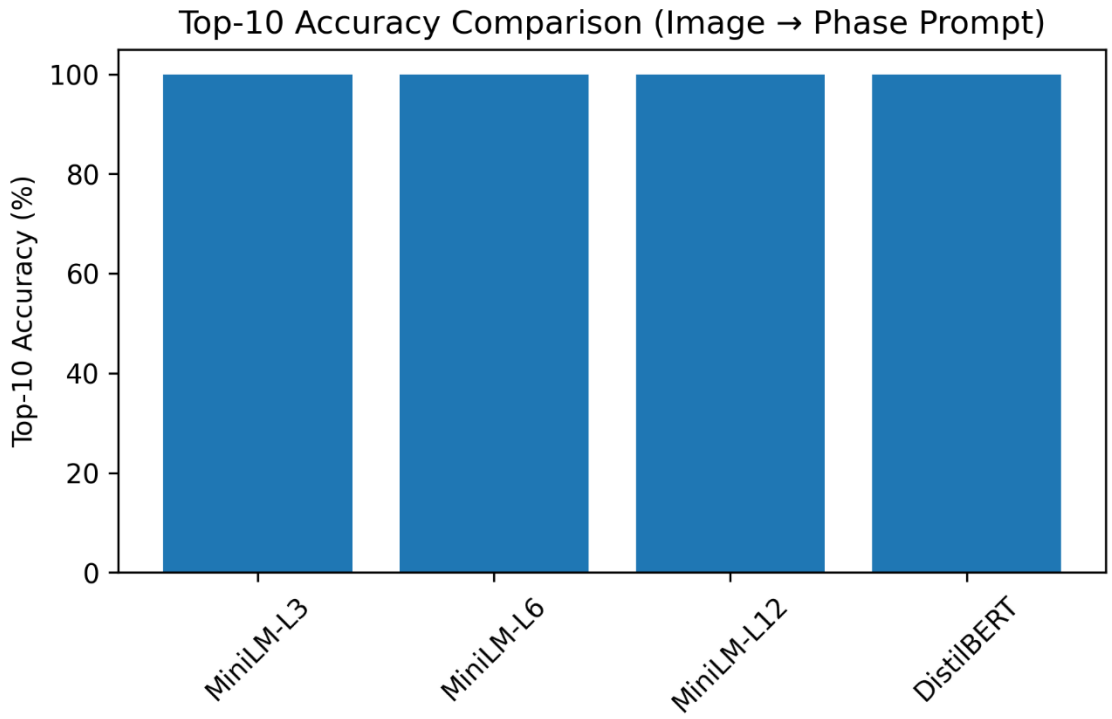

Figure 4.3 Top 10 Accuracy Comparison

Tight Top-5/10 spread indicates the true phase lies near the correct prompt for all models; Top-1 differences reflect collapse to a dominant class vs fine separation between adjacent phases.

4.3 Per-class behavior (confusion matrices & error-flow)

4.3.1 MiniLM-L3

Observation: The confusion matrix is dominated by predictions for Gallbladder Dissection. Correct recognitions for Preparation and Clipping Cutting are present but relatively weak, while Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, and Gallbladder Retraction are seldom identified. This pattern indicates a strong bias toward the long, visually varied Gallbladder Dissection stage.

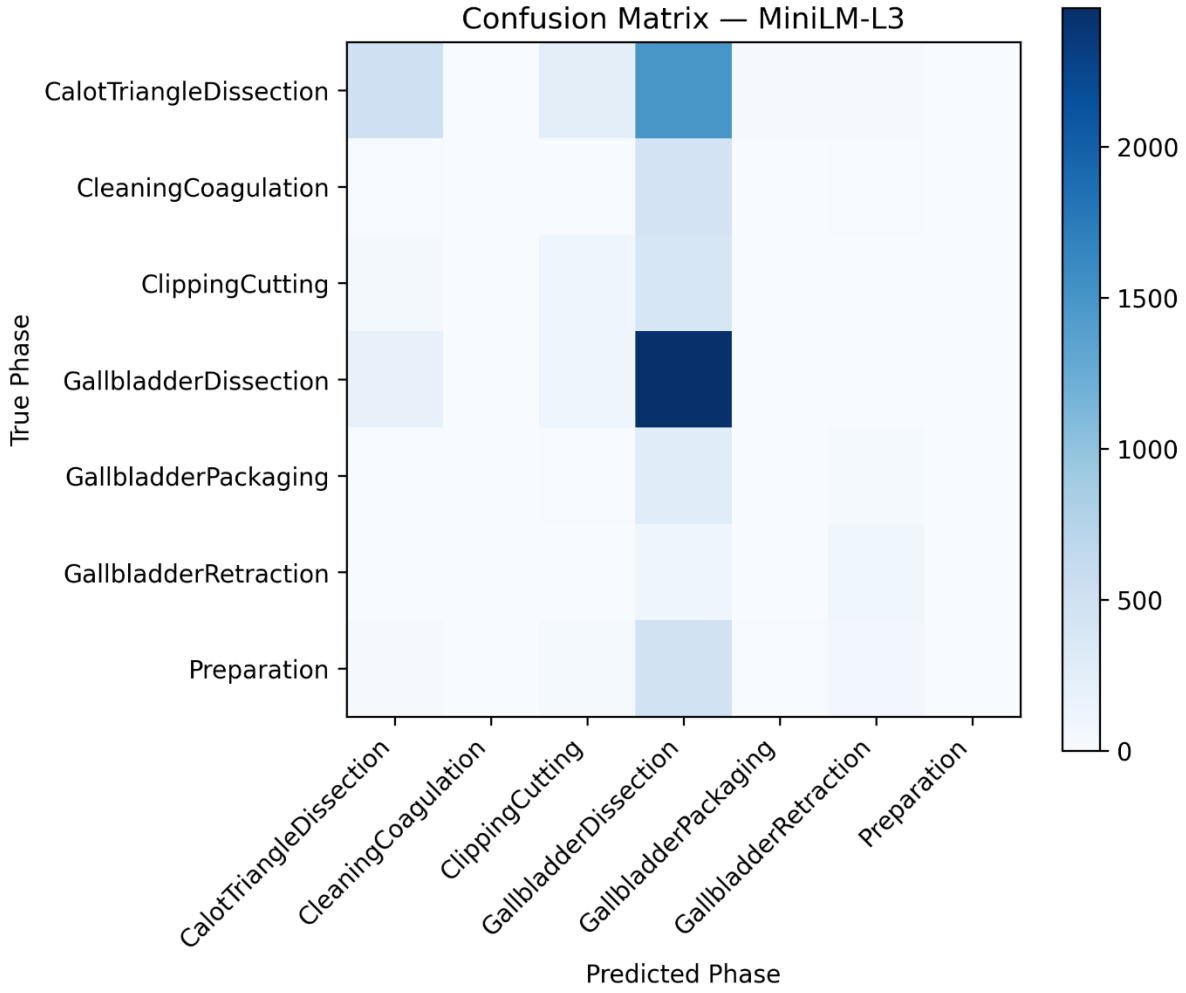

Figure 4.4 Confusion Metrix – MiniLM-L3

Error-flow: The MiniLM-L3 error-flow matrix reveals a strong bias toward predicting Gallbladder Dissection, regardless of the true phase. High misclassification rates from Calot Triangle Dissection (0.68), Cleaning Coagulation (0.95), Clipping Cutting (0.74), Gallbladder Packaging (0.84), and Preparation (0.77) indicate that the lightweight model overgeneralizes common dissection cues and treats this phase as a default when uncertain. This occurs because dissection visually dominates the dataset, while MiniLM-L3 lacks fine semantic discrimination for tool-specific actions such as clip application, coagulation, or bag insertion. As a result, subtle phases or shorter tasks (e.g., Packaging, Retraction) are overshadowed, and even thermally focused Cleaning Coagulation is nearly absorbed into dissection. Although the model achieves high Top-k performance, its accuracy arises from generic visual matching rather than balanced class recognition. Therefore, deployment would require enhancements such as tool-action-aware prompts, balanced loss strategies, and temporal smoothing to prevent critical decision phases from being masked by dissection dominance.

Figure 4.5 Error Flow(Misclassification) – MiniLM-L3

4.3.2 MiniLM-L6

Observation: MiniLM-L6 strengthens the diagonals for Preparation and Clipping Cutting compared with MiniLM-L3, while reducing indiscriminate assignments to Gallbladder Dissection. Minority phases—Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, Gallbladder Retraction—still show low recall, but overall class balance is improved.

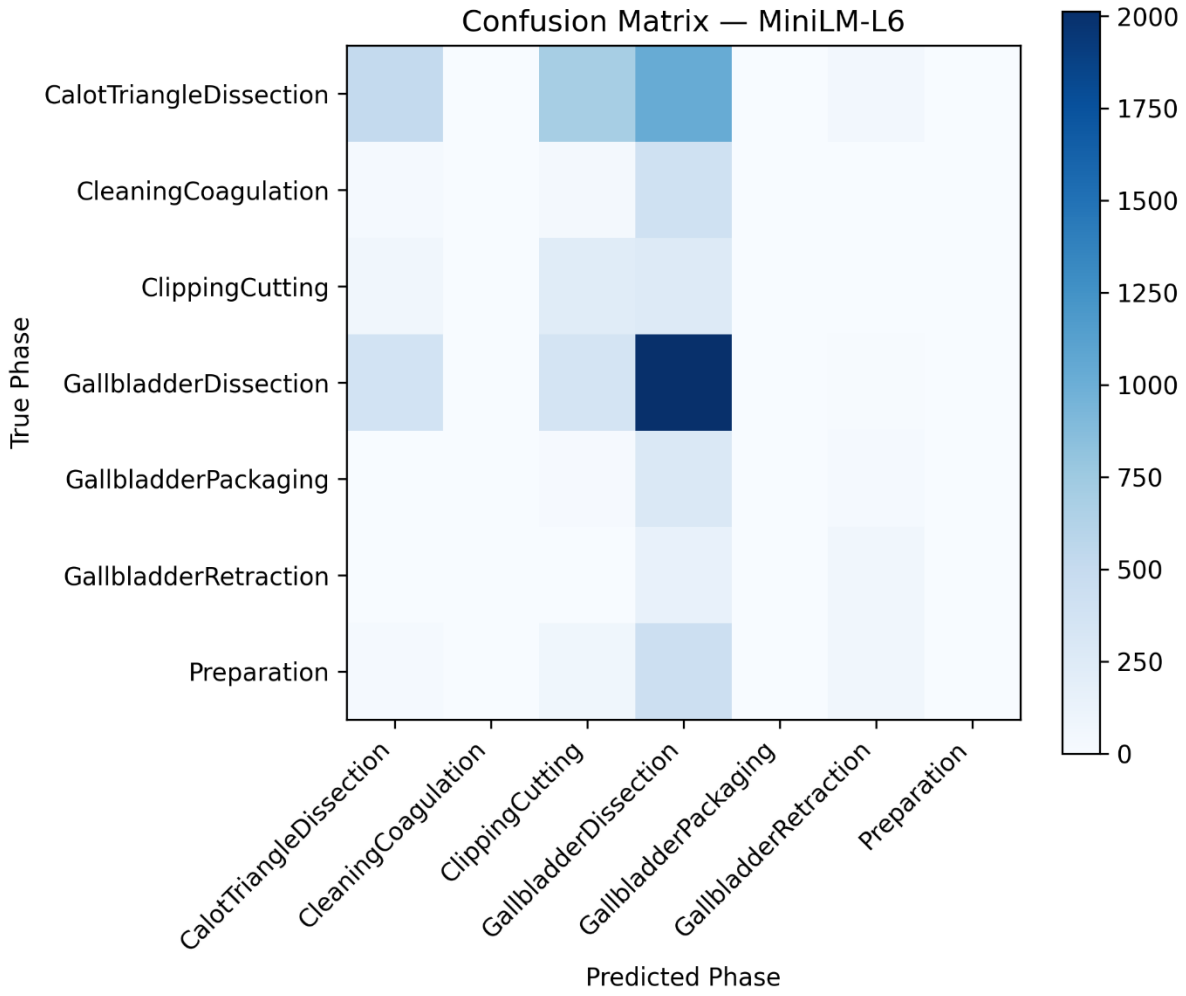

Figure 4.6 Confusion Metrix – MiniLM-L6

Error-flow: The error-flow distribution of MiniLM-L6 continues to show a strong bias toward Gallbladder Dissection, although the bias is less extreme compared to MiniLM-L3. Misclassifications from key phases such as Calot Triangle Dissection (0.44), Cleaning Coagulation (0.84), Clipping Cutting (0.45), Gallbladder Packaging (0.85), and Preparation (0.72) indicate that the model still overgeneralizes common dissection cues but captures more variation in early actions, especially clip placement. Higher confusion between Calot Triangle Dissection and Clipping Cutting (0.32) also reflects improved recognition of preparatory anatomical exposure, even if it remains visually ambiguous. Meanwhile, Packaging and Retraction retain notable uncertainty due to limited visual distinctiveness and tool visibility. Thus, MiniLM-L6 demonstrates more balanced semantic separation than MiniLM-L3 but still relies heavily on generic mid-phase cues, suggesting that enhanced tool-action prompting and temporal reasoning remain essential for reliable phase identification.

Figure 4.7 Error Flow (Misclassification) – MiniLM-L6

4.3.3 MiniLM-L12

Observation: MiniLM-L12 provides the clearest separation between Clipping Cutting and Gallbladder Dissection and also shows better recognition of Preparation. The overall tendency to predict Gallbladder Dissection for many inputs is visibly reduced, yielding fairer per-class behavior even if headline Top-1 is lower.

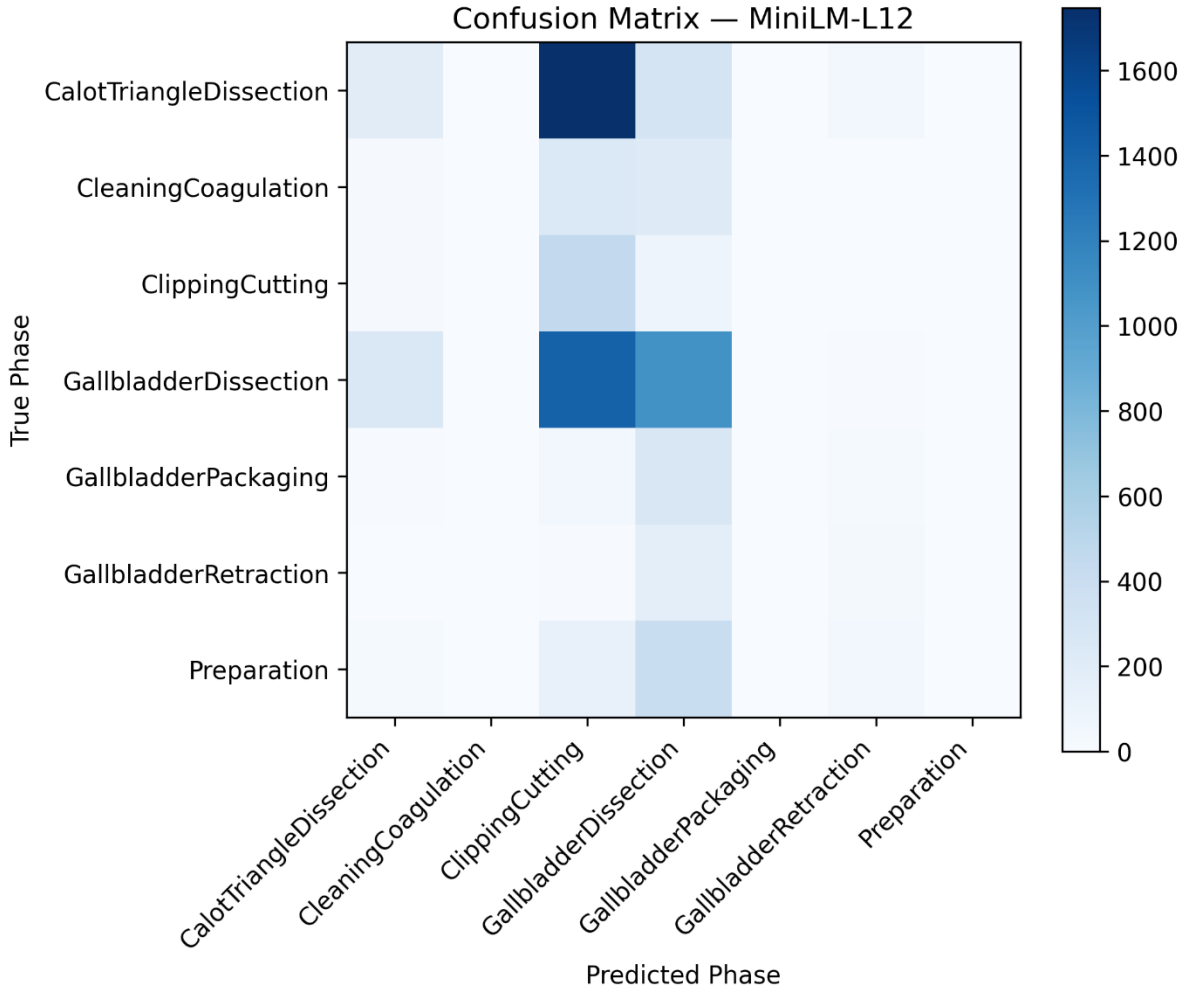

Figure 4.8 Confusion Metrix – MiniLM-L12

Error-flow: MiniLM-L12 exhibits a noticeably more balanced error structure compared to MiniLM-L3 and L6, reducing over-prediction of Gallbladder Dissection and instead redistributing errors toward Clipping Cutting in several phases. Misclassification of Calot Triangle Dissection (0.74) and Cleaning Coagulation (0.48) into Clipping Cutting reflects stronger sensitivity to anatomical exposure and tool-based clipping cues, even though the model still struggles to differentiate specific actions that occur close in the surgical timeline. Meanwhile, misrouting into Gallbladder Dissection persists but at lower intensity (e.g., Packaging: 0.78; Retraction: 0.81) compared to earlier models, indicating improved distinction of mid-phase semantics. MiniLM-L12 therefore learns finer tool-action relationships but remains limited when phases share similar motion patterns, showing that lightweight encoders need explicit instrument-level textual prompts to avoid collapsing neighboring dissections into semantically adjacent clipping tasks.

Figure 4.9 Error Flow (Misclassification) – MiniLM-L12

4.3.4 DistilBERT

Observation: DistilBERT shows the strongest bias toward Gallbladder Dissection: correct hits outside this phase are scarce, and minority phases—Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, Gallbladder Retraction—are rarely detected. Headline Top-1 is maintained largely by frequent predictions of Gallbladder Dissection.

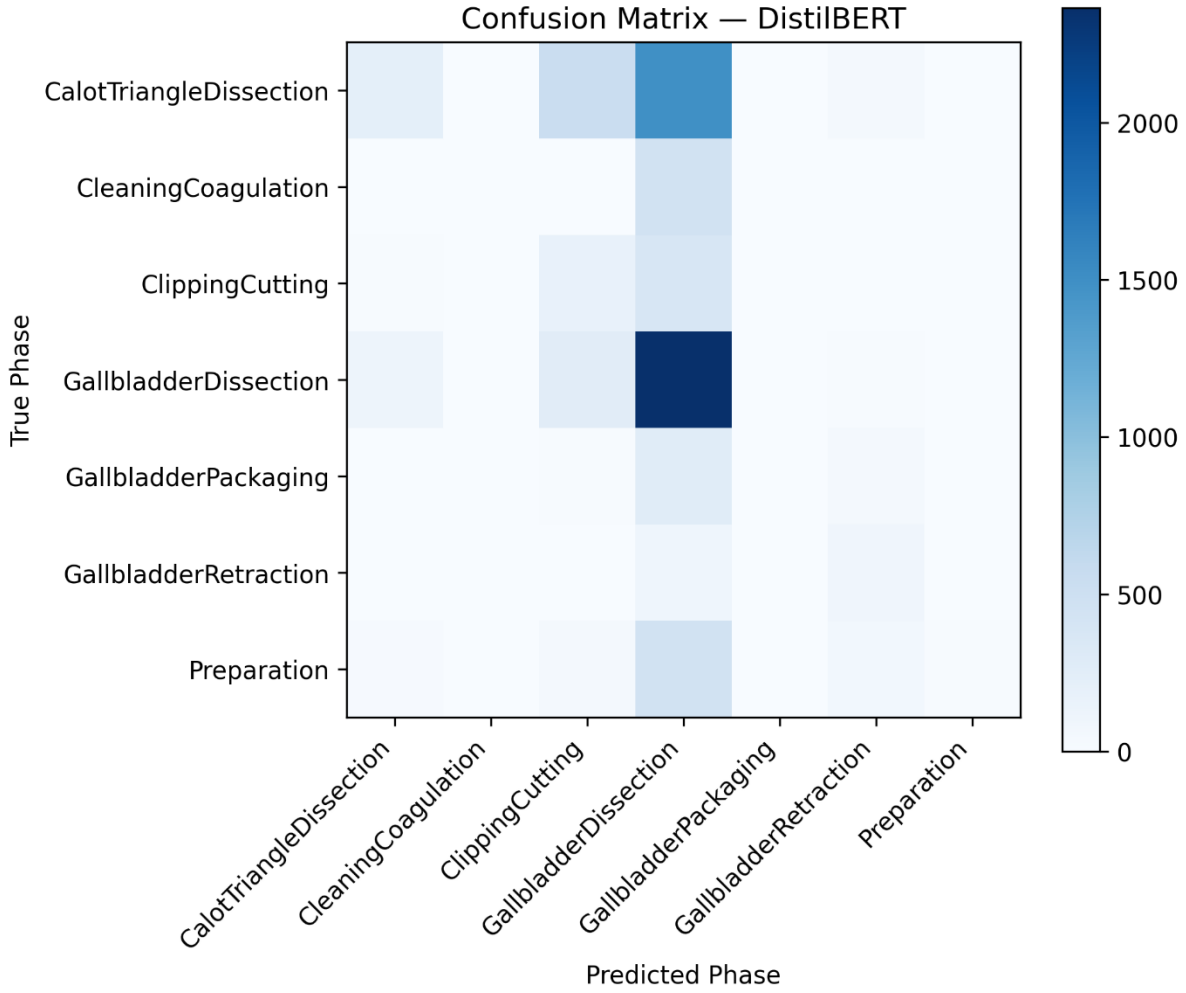

Figure 4.10 Confusion Metrix – DistilBERT

Error-flow: DistilBERT exhibits the strongest phase bias among all evaluated encoders, heavily funneling errors into Gallbladder Dissection, particularly from Cleaning Coagulation (0.96), Clipping Cutting (0.68), Packaging (0.82), and Preparation (0.75). This reflects an extreme reliance on dominant dissection cues and a failure to separate fine-grained surgical actions, especially thermal coagulation, which is nearly indistinguishable to the model. Although Calot Triangle Dissection shows a secondary spillover into Clipping Cutting (0.23), this does not reflect meaningful semantic learning; rather, the encoder collapses nearly all mid and late-stage actions into the longest and most visually generic dissection phase. DistilBERT therefore achieves competitive Top-k accuracy by exploiting phase duration imbalance rather than learning meaningful instrument-action relations. This highlights the need for explicit tool-action textual encoding and class balancing before DistilBERT-based alignment can be considered reliable for surgical decision support.

Figure 4.11 Error Flow (Misclassification) – DistilBERT

4.4 Cross-model comparison & trade-offs

Accuracy vs. balance:

As shown in Fig. 4.3.1, MiniLM-L3 achieves the highest Top-1 largely by predicting Gallbladder Dissection on ambiguous frames, which helps on videos dominated by that stage but depresses minority-phase recall. In Fig. 4.3.3, MiniLM-L6 softens this bias and improves recognition of Preparation and Clipping Cutting, yielding a more balanced profile with only a small Top-1 trade-off. Fig. 4.3.5 shows MiniLM-L12 with the fairest per-class behavior—especially a cleaner boundary between Clipping Cutting and Gallbladder Dissection—but the lowest Top-1 because it resists collapsing uncertain cases into the long dissection stage. Finally, Fig. 4.3.7 indicates DistilBERT maintains a competitive Top-1 mainly by over-assigning Gallbladder Dissection, producing the weakest overall balance.

Shared failure modes:

All models struggle to keep Clipping Cutting distinct from Gallbladder Dissection due to temporal adjacency and overlapping tools/tissue appearances. Shorter phases—Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, (Gallbladder Retraction—remain under-recognized across models.

Neighborhood agreement:

Despite these differences, Top-5 and Top-10 results converge across encoders (see Fig. 4.2.2 and Fig. 4.2.3) indicating that the correct prompt is usually retrieved among the top candidates; most of the gap is in the final argmax choice, not in coarse alignment.

Side-by-side confusion matrices make the trade-off obvious: heavier reliance on Gallbladder Dissection raises Top-1 but degrades fairness, whereas stronger separation—particularly between Clipping Cutting and Gallbladder Dissection—improves balance at some Top-1 cost (Fig 4.3.1, Fig 4.3.3, Fig 4.3.5 and Fig 4.3.7)

4.5 Error analysis & practical remedies

Primary error sources

- The long and visually diverse Gallbladder Dissection dominates training signals, so ambiguous frames from other stages are often assigned there.
- Clipping Cutting and early Gallbladder Dissection share tools, tissue exposure, and verbs ("dissect," "separate"), blurring their boundary.
- Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, and Gallbladder Retraction are briefer or visually understated, leading to low support and weak recall.
- Without temporal smoothing, predictions flicker at stage transitions.

Targeted remedies (drop-in and lightweight)

1. **Make prompts phase-specific and instrument-aware**

    - Refine each prompt to include verb + instrument + object cues, e.g.:

        - **Clipping Cutting:** "apply clip applier to cystic duct and cut with scissors,"
        - **Gallbladder Dissection:** "dissect gallbladder from liver bed using cautery,"
        - **Cleaning Coagulation:** "coagulate oozing bed and rinse field with suction–irrigation."

    - Optionally ensemble multiple prompts per phase (average their text embeddings).
    - Quick check: confusion between Clipping Cutting and Gallbladder Dissection should shrink, with precision for each improving.

2. **Stabilize over time**
    - Post-process frame scores with majority/median filtering over a short window; or use a light HMM/TCN on top of embeddings at inference.
    - Fewer rapid label flips near transitions, higher per-video consistency without retraining the vision tower.

3. **Guard against overfit to background cues.**
    - Apply modest augmentation (color jitter, small rotations) and randomize non-diagnostic borders so the model relies more on tool–tissue interaction than background.

What to report after applying fixes

- Add a small ablation table showing macro-precision/recall/F1 and Top-1 before vs after:
    - rebalancing,
    - prompt assembling,
    - pooling/τ selection,
    - temporal smoothing.

- Emphasize improvements for Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, Gallbladder Retraction, and reduced confusion between Clipping Cutting and Gallbladder Dissection—while keeping overall efficiency unchanged (frozen vision).

4.6 Summary of findings

MiniLM-L3 delivers the highest Top-1 (44%) but does so by frequently predicting Gallbladder Dissection; MiniLM-L6 moderates this bias and improves recognition of Preparation and Clipping Cutting; MiniLM-L12 shows the fairest per-class behavior—especially a cleaner boundary between Clipping Cutting and Gallbladder Dissection—at the cost of lower Top-1; DistilBERT attains a competitive Top-1 largely through over-assignment to Gallbladder Dissection, resulting in the weakest balance.

Across encoders, Top-5 is consistently 95% and Top-10 100%, indicating that the correct phase prompt almost always appears among the top candidates. Thus, differences across models stem mainly from the final selection step rather than from a failure to retrieve the correct neighborhood.

The most persistent confusion occurs between Clipping Cutting and Gallbladder Dissection due to temporal adjacency and overlapping tool–tissue cues. Shorter or subtler phases—Calot Triangle Dissection, Gallbladder Packaging, Cleaning Coagulation, Gallbladder Retraction—remain under-recognized.

With the image tower frozen, lightweight text encoders can maintain strong neighborhood alignment, but class balance depends on prompt specificity and supervision balance. Simple, deployment-friendly remedies—class-aware sampling, verb + instrument prompt design (with optional prompt assembling), careful pooling/temperature choices, and light temporal smoothing—are expected to raise macro-level performance and reduce confusion between Clipping Cutting and Gallbladder Dissection while preserving efficiency.

CHAPTER 5

CONCLUSION

5.1 Summary of Findings

This experiment examined the dynamics of small trained transformer text encoders in a fixed prompt multi-modes learning model, where surgical phase recognition is the controlled trial. The only variation was the encoder depth and capacity of the text pathway with all other training factors held constant to identify the effect they had in influencing multimodal alignment behavior in low-data conditions.

The results show that small encoders of text can provide high rates of neighborhood scale compatibility with a frozen vision backbone. High results were observed in all the assessed models on Top-5 accuracy and Top-10 accuracy was very close to saturation indicating that the correct phase suggest was almost accurately retrieved in local semantic environment. This fact suggests that there exists rough multimodal alignment which is largely preserved regardless of the size of encoders.

However there were great differences in final selection behavior. Minimally shallow encoders in particular MiniLM-L3 performed more accurately at Top-1, though with a higher amount of bias in classes, most evidently to Gallbladder Dissection. This bias was reduced in MiniLM-L6 with an encoder of medium depth, and it improved the recognition of less intense phases. The less unbalanced per-class behaviour and the more pronounced dissimilar phases of MiniLM-L12 encoder are accompanied by the lower Top-1 accuracy of the judgements in fixed-prompt and low-data scenarios. DistilBERT managed to achieve competitive Top-1 scores yet a low class balance as it was over-assigned to dominant stages.

Further trans-location was observed between a temporal related and a visual related phase, namely, Clipping and Cutting and Gallbladder Dissection. Even the small and less prominent intermittences were not detected in all the models, which were the combined effects of the imbalance in classes and the lack of oversight.

Overall, the findings can show that under fixed-prompt conditions compact text encoders would be able to sustain multimodal alignment, though the depth of the encoder is an

essential consideration establishing which encoder selection is stable and the behavior of classes. The higher the capacity, the higher the difference in performance, and can lead to instability in case of minimal supervision. The results are noteworthy because they justify the importance of studying behavior in comparison to restricted multimodal learning situations when comparing text encoders.

5.2 Contributions to the Field

The paper is something significant to the area of multimodal learning and vision-language representation analysis, especially on constrained and fixed-prompt scenarios.

To begin with, this paper gives a behavioral systematic analysis of compact transformer text encoders under a fixed-prompt multimodal learning design. In contrast to the previous works with the focus on end-to-end performance or large-scale pretraining, this thesis separates the text pathway by simply freezing the vision encoder and standardizing all the other experimental factors. The design allows the direct analysis of the effect of the encoder depth and capacity on multimodal alignment behavior in low-data conditions.

Second, the research presents the empirical findings that the greater the text encoder depth, the higher the multimodal performance is not always achieved in case of the limitations of linguistic supervision. These findings show that less deep encoders are able to be more stable and exhibit more successful alignment behaviour than their deep counterparts, and that larger language models are not necessarily superior in multimodal tasks.

Third, this study presents a controlled and repeatable assessment guide of the behavior of text encoders in multimodal systems. With fixed prompts, fixed preprocessing, consistent data splits, and identical optimization parameters, the study will solve long-term comparability problems in earlier literature of multimodal and surgical phase recognition. This is a protocol that could be utilized again to form a diagnostic platform in subsequent studies on encoder design decisions.

Fourth, it is also observed that the class-wise behavior and selection stability of fixed-prompt multimodal learning are significantly dependent on coarse retrieval capability alone. The fact that all encoders are highly aligned at neighborhood-level and at least partially different in the

choice of final classes offers novel information on how compact transformers are able to tackle semantic ambiguity under limited supervision.

Last but not the least, this is practically oriented and empirically based advice to the selection of text encoders in constrained multimodal learning context. The study can guide model design choices in applications where size, performance, and stability trade-offs are significant factors by understanding the size-performance-stability trade-offs across compact encoders.

5.3 Future Work

Although this research offers controlled analysis of compact transformer text encoders in a fixed-prompt multimodal learning, there are still a number of areas that can be explored in the future.

To begin with, adaptive or learned prompt-representations can be investigated in the future work subject to a controlled state of evaluation. Though this makes it possible to isolate the behavior of the text encoder by use of fixed prompts, prompt optimization or prompt ensembles may indicate how the depth of the encoder can interact with linguistic flexibility and how shallow encoders can be made, without collapsing in the presence of higher prompt variety.

Second, it would be beneficial to apply the analysis to larger and more diverse datasets to establish how far the observed patterns of behavior are applicable to the conditions outside the low-data regime. Assessment of compact text encoders on multi-procedure or cross-institutional surgical datasets can help to understand whether more powerful encoders are more active with increased supervision or a broader domain of variation.

Third, the literature might explore the topic of lightweight temporal modeling as an addition to fixed-prompt alignment in the future. More basic temporal smoothing or phase-transition constraints can eliminate uncertainty between temporally neighboring phases without the need to model the sequence of temporal behavior explicitly, which allows one to further examine the role of text encoders in facilitating temporal consistency.

Fourth, further studies can be conducted on different projection and pooling techniques in the text pathway such as token-level pooling or attention-based summarization to gain a better insight into the impact of representational compression on alignment stability. On the same note, systematic temperature calibration plans might be researched to determine their effects on selection confidence and proportion of classes.

Lastly, this framework could be extended to other multimodal fields in the future besides surgical video analysis (radiology, pathology, procedural robotics). The identical controlled, fixed prompt evaluation protocol would permit a comparative evaluation of compact text encoder behavior across tasks to add to the influence of representing learning in constrained multimodal systems.

5.4 Conclusion

The thesis discussed the behavior of small pretrained transformer text encoders in a fixed-prompt multimodal learning system where a controlled evaluation test is surgical phase recognition. In the study, decoupling of the effects of encoder depth and capacity on multimodal alignment behavior under low-data conditions was also achieved by freezing the vision backbone and manipulating the text encoder systematically under all the same training and evaluation settings.

It has been discovered that small text encoders can maintain good multimodal consistency at a neighborhood scale with small model capacity. The findings do show though that increase in encoder depth does not always prove to be better when one is working with fixed-prompt constraints. The encoders were more stable, and the behaviour of the deeper encoders was more balanced, but the less robust to the issues of optimization and worsening behaviour. This means that representational capacity comes at a heavy trade-off with the stability of alignment of constrained multimodal learning.

It is worth noting that the Top-5 and Top-10 retrieval performance of all the tested encoders remains high and this indicates that most of the models can retrieve the correct semantic neighborhood. The discrepancies in performances are therefore primarily due to the final

selection of prompts that are in close relation with each other rather than attributing it to the failure to fine tune the multimodal performance. The observation adds to the reason why the study of encoder behavior is relevant beyond the aggregate measures of accuracy in the assessment of multimodal systems.

Its capability to provide text pathway behavior and a controlled and repeatable protocol of evaluation is an empirical clarity of the design and analysis of multimodal learning system that is helpful in informing the selection of the compact text encoders in terms of the stability and behavioral attributes in comparison to the model size. In general terms, the contemporary study demonstrates the reductionist analysis can be applicable in describing the multimodal representation learning and forms the platform on which future investigations can be carried out on constrained and interpretable multimodal systems.

CHAPTER 6

REFERENCES

1. 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*.

2. Funke, I., Rivoir, D., Krell, S., & Speidel, S. (2025). TUNeS: A Temporal U-Net With Self-Attention for Video-Based Surgical Phase Recognition. *IEEE Transactions on Biomedical Engineering*, *72*(7), 2105–2119. IEEE Computer Society.

3. Funke, I., Rivoir, D., & Speidel, S. (2023). Metrics Matter in Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2305.13961

4. He, Y., Zhu, Y., Fu, P., Yang, R., Chen, T., Wang, Z., Li, Q., et al. (2025). Endo-CLIP: Progressive Self-Supervised Pre-training on Raw Colonoscopy Records. Retrieved from http://arxiv.org/abs/2505.09435

5. Hoque, M., Hasan, R., Emon, S., Khalifa, F., & Rahman, M. M. (2024). *Medical Image Interpretation with Large Multimodal Models Notebook for the CS_Morgan Lab at CLEF 2024*. Retrieved from https://github.com/Hasan-MdRakibul

6. Kostiuchik, G., Sharan, L., Mayer, B., Wolf, I., Preim, B., & Engelhardt, S. (2024). Surgical phase and instrument recognition: how to identify appropriate dataset splits. *International Journal of Computer Assisted Radiology and Surgery*, *19*(4), 699–711. Springer Science and Business Media Deutschland GmbH.

7. Park, M., Oh, S., Jeong, T., & Yu, S. (2023). Multi-Stage Temporal Convolutional Network with Moment Loss and Positional Encoding for Surgical Phase Recognition. *Diagnostics*, *13*(1). Multidisciplinary Digital Publishing Institute (MDPI).

8. Perez, A., Nwoye, C., Kermani, R. R., Mohareri, O., & Jamal, M. A. (2025). SurgLaVi: Large-Scale Hierarchical Dataset for Surgical Vision-Language Representation Learning. Retrieved from http://arxiv.org/abs/2509.10555

9. 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.

10. Roy, S., Parhizkar, Y., Ogidi, F., Khazaie, V. R., Colacci, M., Etemad, A., Dolatabadi, E., et al. (2024). Benchmarking Vision-Language Contrastive Methods for Medical Representation Learning. Retrieved from http://arxiv.org/abs/2406.07450

11. Schmidgall, S., Cho, J., Zakka, C., & Hiesinger, W. (2024a). GP-VLS: A general-purpose vision language model for surgery. Retrieved from http://arxiv.org/abs/2407.19305

12. Schmidgall, S., Cho, J., Zakka, C., & Hiesinger, W. (2024b). GP-VLS: A general-purpose vision language model for surgery. Retrieved from http://arxiv.org/abs/2407.19305

13. Yuan, K., Srivastav, V., Navab, N., & Padoy, N. (2025a). HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2405.10075

14. Yuan, K., Srivastav, V., Navab, N., & Padoy, N. (2025b). HecVL: Hierarchical Video-Language Pretraining for Zero-shot Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2405.10075

15. 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

CHAPTER 7

APPENDICES

221-35-900

ORIGINALITY REPORT

12%	10%	5%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
2	Submitted to Midlands State University Student Paper	2%
3	Submitted to Daffodil International University Student Paper	1%
4	arxiv.org Internet Source	1%
5	stanford.edu Internet Source	<1%
6	issuu.com Internet Source	<1%
7	Kohei Yamamoto, Tomohiro Kikuchi. "Feasibility Study of CLIP-Based Key Slice Selection in CT Images and Performance Enhancement via Lesion- and Organ-Aware Fine-Tuning", Bioengineering, 2025 Publication	<1%
8	Zheyuan Zhang, Muhammad Ibtsaam Qadir, Matthias Carstens, Evan Hongyang Zhang et al. "Prompt injection attacks on vision-language models for surgical decision support", Cold Spring Harbor Laboratory, 2025 Publication	<1%
9	ouci.dntb.gov.ua Internet Source	<1%

10	Submitted to Edith Cowan University Student Paper	<1 %
11	umpir.ump.edu.my Internet Source	<1 %
12	www2.mdpi.com Internet Source	<1 %
13	www.nature.com Internet Source	<1 %
14	Submitted to University of Nottingham Student Paper	<1 %
15	uhra.herts.ac.uk Internet Source	<1 %
16	pure.mpg.de Internet Source	<1 %
17	rucore.libraries.rutgers.edu Internet Source	<1 %
18	Zhenzhong Liu, Kelong Chen, Shuai Wang, Yijun Xiao, Guobin Zhang. "Deep learning in surgical process Modeling: A systematic review of workflow recognition", Journal of Biomedical Informatics, 2025 Publication	<1 %
19	Hernández-Cámara, Pablo. "Perceptual Alignment in Artificial Vision: Bio-Inspired Design and Psychophysical Evaluation.", Universitat de Valencia (Spain) Publication	<1 %
20	Submitted to University of Surrey Student Paper	<1 %
21	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
	Submitted to The University of Manchester	

22	Student Paper	<1 %
23	uir.unisa.ac.za Internet Source	<1 %
24	www.mdpi.com Internet Source	<1 %
25	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2023", Springer Science and Business Media LLC, 2023 Publication	<1 %
26	etda.libraries.psu.edu Internet Source	<1 %
27	www.math.utep.edu Internet Source	<1 %
28	Submitted to Liverpool John Moores University Student Paper	<1 %
29	Submitted to De LaSalle - College of Saint Benilde Student Paper	<1 %
30	Submitted to Fakultet elektrotehnike i računarstva / Faculty of Electrical Engineering and Computing Student Paper	<1 %
31	Shi, Xueying. "Towards Cost-Effective Medical Image Analysis: From Active Learning, Semi-supervised Learning to Cross-Domain Learning", The Chinese University of Hong Kong (Hong Kong) Publication	<1 %
32	Submitted to Universiti Malaysia Pahang Student Paper	<1 %

33	Submitted to University of Adelaide Student Paper	<1 %
34	Zihao Zhao, Yuxiao Liu, Han Wu, Mei Wang, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Zhiming Cui, Qian Wang, Dinggang Shen. "CLIP in medical imaging: A survey", Medical Image Analysis, 2025 Publication	<1 %
35	jyx.jyu.fi Internet Source	<1 %
36	lutpub.lut.fi Internet Source	<1 %
37	Nasseh Hashemi, Matias Mose, Lasse R. Østergaard, Flemming Bjerrum et al. "Closing the data gap: leveraging pretrained neural networks for robotic surgical assessment on limited clinical data", Journal of Robotic Surgery, 2025 Publication	<1 %
38	Sandy Engelhardt. "Why Thorough Open Data Descriptions Matters More Than Ever in the Age of AI: Opportunities for Cardiovascular Research", European Heart Journal - Digital Health, 2024 Publication	<1 %
39	Submitted to Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) Student Paper	<1 %
40	www.southlewis.org Internet Source	<1 %
41	inass.org Internet Source	<1 %
42	link.springer.com Internet Source	

		<1 %
43	openscholar.dut.ac.za Internet Source	<1 %
44	www.arxiv.org Internet Source	<1 %
45	"Medical Image Learning with Limited and Noisy Data", Springer Science and Business Media LLC, 2022 Publication	<1 %
46	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 %
47	opus.hs-furtwangen.de Internet Source	<1 %
48	raw.githubusercontent.com Internet Source	<1 %
49	repository.nwu.ac.za Internet Source	<1 %
50	shura.shu.ac.uk Internet Source	<1 %
51	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 %
52	Zhe Min, Jiewen Lai, Hongliang Ren. "Innovating robot-assisted surgery through	<1 %

	large vision models", Nature Reviews Electrical Engineering, 2025 Publication	
53	eprints.leedsbeckett.ac.uk Internet Source	<1 %
54	mdpi-res.com Internet Source	<1 %
55	oulurepo.oulu.fi Internet Source	<1 %
56	www.researchgate.net Internet Source	<1 %
57	www.ufs.ac.za Internet Source	<1 %
58	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2020", Springer Science and Business Media LLC, 2020 Publication	<1 %
59	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2024", Springer Science and Business Media LLC, 2024 Publication	<1 %
60	Ngigi, William K.. "Open-Set Recognition in Computer Vision.", Indiana University of Pennsylvania Publication	<1 %
61	Peng Jun Xu, Shuang Xiang Kan, Jing Jin, Zhou Jing Zhang, Ya Xin Gu, Bo Zhang, You Lang Zhou. "Multimodal large language models in medical research and clinical practice: Development, applications, challenges and future", Neurocomputing, 2026 Publication	<1 %

62	dr.ur.ac.rw Internet Source	<1 %
63	dspace.bracu.ac.bd:8080 Internet Source	<1 %
64	eprints.usm.my Internet Source	<1 %
65	files.core.ac.uk Internet Source	<1 %
66	ir.mu.ac.ke:8080 Internet Source	<1 %
67	scholar.sun.ac.za Internet Source	<1 %
68	theses.hal.science Internet Source	<1 %
69	www.preprints.org Internet Source	<1 %
70	www.researchsquare.com Internet Source	<1 %
71	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2021", Springer Science and Business Media LLC, 2021 Publication	<1 %
72	Gavrikov, Paul. "Decoding Robust Generalization in Object Recognition Models", Universitaet Mannheim (Germany), 2025 Publication	<1 %
73	Kubilay Can Demir, Hannah Schieber, Daniel Roth, Andreas Maier, Seung Hee Yang. "Surgical Phase Recognition: A Review and Evaluation of Current Approaches", Institute	<1 %

of Electrical and Electronics Engineers (IEEE), 2022
Publication

74	Taghain Dinani, Soudabeh. "Leveraging Advanced Deep Learning Models for Disaster Response", Kansas State University, 2024 Publication	<1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		