Performance of Modern Vision Backbones for Surgical Phase Recognition Without Fine-Tuning

MAYSHA MAHJABIN MIMI

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

APPROVAL

This thesis titled on "**Performance of Modern Vision Backbones for Surgical Phase Recognition Without Fine-Tuning**", submitted by **Maysha Mahjabin Mimi (ID: 221-35-942)** 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	:	Maysha Mahjabin Mimi
Date of Birth	:	03 May 2001
Title	:	Analyzing Vision Encoder Variants for Efficient Multimodal Pretraining in CLIP
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:

Mimi

(Student's Signature)

221-35-942

Student ID

Date:27-12-25

(Supervisor's Signature)

Mr. Khalid Been Badruzzaman Biplob

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. Khalid Been Badruzzaman Biplob
Position : Lecturer (Senior Scale)
Date : 27-12-25

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 : Maysha Mahjabin Mimi
ID Number : 221-35-942
Date : 27-12-25

Performance of Modern Vision Backbones for Surgical Phase Recognition Without Fine-Tuning

Maysha Mahjabin Mimi

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

©Daffodil International University

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 Lecturer (Senior Scale) Mr. Khalid Been Badruzzaman Biplob. 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

Abstract

The thesis studies the skills of lightweight backbones of vision in surgical phases recognition in a fixed CLIP-like framework without fine-tuning. In the video of laparoscopic cholecystectomy in Cholec80 dataset, we sample 1fps of the images in the video, resize and suppose are resized to 224x224and CLIP-normalised and matched with textual phase prompts, where the text is the name of standard procedures. It has a two-encoder architecture where both CLIP text tower and vision backbones, ViT-Tiny, ViT-Small US-CTC7, ConvNeXt-Tiny and Swin-Transformer-Tiny are frozen and it is trained on a shallow projection head with symmetric contrastive loss (InfoNCE). All backbones are shown to perform well with Image-text retrieval (Recall@1/5/10) and zero-shot phase classification (Top-1/5/10), per-class precision/recall/F1 and confusion matrices and throughput, number of parameters, memory usage with ConvNeXt-Tiny being best (0.64-0.65), ViT-Small being average (0.56), and ViT-Tiny being the worst (.49). Based on error studies in the frozen-encoder setting, confusion (e.g. Triangle Dissection vs. Gallbladder Dissection / Clipping and Cutting) and predictive misalignment on the instance of Packaging uniformly across ViT models indicate difference in architectural inductive bias and capacity. In total, ConvNeXt-Tiny can achieve the best accuracy-efficiency trade-off and Swin-Tiny is a model that could be utilized in a similar way in situations where hierarchical spatial context is required. The article gives a controlled trial of four backbones of lightweight frozen in CLIP, a head-only training and incremental retraining pipeline that is supposed to be utilized in low-resource clinical settings; and we also make suggestions in practice on the selection of compact deployable backbones in frame-level surgical phase recognition. These outcomes demystify the useability of new freeze encoders of multimodal video based surgical knowledge.

TABLE OF CONTENTS

THESIS DECLARATION LETTER (OPTIONAL) ..... **IV**

SUPERVISOR'S DECLARATION ..... **V**

STUDENT'S DECLARATION ..... **VI**

ACKNOWLEDGEMENTS ..... **VIII**

DEDICATION ..... **IX**

ABSTRACT ..... **X**

TABLE OF CONTENTS ..... **XI**

LIST OF TABLES ..... **XII**

LIST OF FIGURES ..... **XIII**

LIST OF SYMBOLS ..... **XIV**

LIST OF ABBREVIATIONS ..... **XV**

LIST OF APPENDICES ..... **XVIII**

1. CHAPTER 1 ..... **1**

1.1 Background 1
1.2 Problem Statement 2
1.3 Motivation 3
1.4 Significance of the Study 4
1.5 Research Questions 5
1.6 Research Objectives 5
1.7 Research Scope and Limitations 5
1.7.1 Scope *6*
1.7.2 Limitations *7*
1.8 Thesis Organization 7

2. CHAPTER 2 ..... **9**

2.1 Background on Surgical Workflow Analysis and Multimodal Learning 9
2.2 Deep Learning Approaches for Surgical Phase Recognition 9
2.3 Vision Encoders for Multimodal Surgical Understanding 13
2.4 Limitations and Research Gap 14
2.5 Research Focus and Conceptual Framework 15

3. CHAPTER 3 ..... **17**

3.1 Dataset Description 19
3.2 Preprocessing and Data Augmentation 23
3.3 Multimodal Architecture (CLIP-based Dual Encoder) 24
3.4 Vision Encoder Variants 25
3.5 Frozen Text Encoder 27
3.6 Projection Head (Alignment Layer) 28
3.7 Training Setup 29
3.8 Evaluation Metrics 30
3.9 Summary 31

4. CHAPTER 4 ..... **32**

4.1 Evolution Protocol 32
4.2 Overall Top-k Accuracy 32
4.3 Per-class behavior (confusion matrices & error-flow) 38
4.4 Cross-model comparison & trade-offs 45

4.5 Error analysis & practical remedies 47
4.6 Summary of findings 48
CHAPTER 5 50
5.1 Summary of findings 50
5.2 Contributions to the field 51
5.3 Future work 52
5.4 Conclusion 54
5. CHAPTER 6 56
References: 56
APPENDICES 58

LIST OF TABLES

Table 2.1 Several representative papers in the field of surgical phase recognition (SPR) 11

Table 3.1 phase Durations 21

Table 4.1 All model accuracy comparison summery 33

LIST OF FIGURES

Figure 3.1 Clip Architecture 24

Figure 3.2 Architecture Pipeline 27

Figure 3.3 linear projection Architecture 28

Figure 4.1 overall Recall@k all encoder Comparison 33

Figure 4.2 Overall Top-K Accuracy -Vit_tiny_patch16_224 34

Figure 4.3 Overall Top-K Accuracy -Vit_small_patch16_224 35

Figure 4.4 Overall Top-K Accuracy -ConvNeXt_TIny 36

Figure 4.5 Overall Top-K Accuracy -Swin_TIny 37

Figure 4.6 ViT-Tiny Confusion Matrix 38

Figure 4.7 ViT-Tiny Error_Flow 39

Figure 4.8 ViT-small Confusion Matrix 40

Figure 4.9 ViT-Small Error_Flow 41

Figure 4.10 ConvNeXt-Tiny Confusion Matrix 42

Figure 4.11 ConvNeXt-Tiny Error_Flow 43

Figure 4.12 Swin-Tiny Confusion Matrix 44

Figure 4.13 Swin-Tiny Error_Flow 45

LIST OF SYMBOLS

$\boldsymbol{x}$	**Image/frame (preprocessed RGB)**
$\boldsymbol{y}$	Text prompt (surgical phase)
$\boldsymbol{\mathcal{B}}$	Mini-batch
(N=	\mathcal{B}
$\boldsymbol{C}$	Number of classes/phases
$\boldsymbol{f_v}(\cdot)$	Vision encoder
$\boldsymbol{f_t}(\cdot)$	Text encoder (frozen CLIP)
$\boldsymbol{W_t}, \boldsymbol{b_t}$	Text projection weights/bias
$\tilde{\boldsymbol{z}} = \boldsymbol{W_v f_v}(\boldsymbol{x}) + \boldsymbol{b_v}$	Unnormalized image embedding
$\tilde{\boldsymbol{t}} = \boldsymbol{W_t f_t}(\boldsymbol{y}) + \boldsymbol{b_t}$	Unnormalized text embedding
$\boldsymbol{z} = \tilde{\boldsymbol{z}} / \| \tilde{\boldsymbol{z}} \|_2$	Normalized image embedding
$\boldsymbol{t} = \tilde{\boldsymbol{t}} / \| \tilde{\boldsymbol{t}} \|_2$	Normalized text embedding
$\boldsymbol{D}$	Shared embedding dimension
$\boldsymbol{d_v}, \boldsymbol{d_t}$	Vision/text feature dimensions
$\boldsymbol{s_{ij}} = \boldsymbol{z_i^\top t_j}$	Cosine similarity (image (i), text (j))
$\boldsymbol{\tau}$	Temperature (InfoNCE/CLIP)
$\boldsymbol{\mathcal{L}_{\text{NCE}}}$	Contrastive (InfoNCE/CLIP) loss
$\hat{\boldsymbol{y}}(\boldsymbol{x}) = \mathbf{arg\ max}_{\boldsymbol{j}}\, \boldsymbol{s_{ij}}$	Zero-shot predicted label
$\hat{\boldsymbol{y}}^{(\mathbf{1}:\boldsymbol{k})}$	Top-(k) prediction set
$\mathbf{R@}\boldsymbol{k}$	Recall@(k)
$\mathbf{1}\{\cdot\}$	Indicator function
$\mathbf{C} \in \mathbb{R}^{\boldsymbol{C}\times\boldsymbol{C}}$	Confusion matrix
$\boldsymbol{C_{a,b}}$	Confusion entry (true (a), predicted (b))
$\boldsymbol{TP_c}, \boldsymbol{FP_c}, \boldsymbol{FN_c}$	True/False Positives/Negatives (class (c))

LIST OF ABBREVIATIONS

AdamW	Adaptive Moment Estimation with Weight Decay
AI	Artificial Intelligence
AS	Attention Score
CE	Cross-Entropy (loss)
CLIP	Contrastive Language–Image Pretraining
CNN	Convolutional Neural Network
ConvNeXt	ConvNeXt (modern CNN architecture)
ConvNeXt-Tiny	ConvNeXt Tiny
CRFs	Conditional Random Fields
Endo-CLIP	Endoscopy-adapted CLIP
EndoVis	Endoscopic Vision Challenge
EndoViT	Endoscopic Vision Transformer
F1	F1-Score
FN	False Negative
FN_c	False Negatives for class c
FP	False Positive
FP_c	False Positives for class c
FPS	Frames Per Second
GI	Gastrointestinal
GP-VLS	General-Purpose Vision–Language for Surgery
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
ID	Identifier
IT	Information Technology
LayerNorm	Layer Normalization
LoRA	Low-Rank Adaptation
LR	Learning Rate
LSTM	Long Short-Term Memory

M2CAI16	MICCAI 2016 Workflow Challenge
Macro-F1	Macro-Averaged F1-Score
mAP	mean Average Precision
MLP	Multi-Layer Perceptron
OR	Operating Room
PEFT	Parameter-Efficient Fine-Tuning
PRF	Precision / Recall / F1
PyTorch	PyTorch Deep Learning Framework
QK	Query–Key (attention inputs)
ResNet	Residual Network
RN50	ResNet-50
RGB	Red–Green–Blue
R@K	Recall at K
RTX	NVIDIA GeForce RTX
SMOTE	Synthetic Minority Over-sampling Technique
SPR	Surgical Phase Recognition
SurgLaVi	Surgical Language–Vision (dataset)
SurgPETL	Surgical Parameter-Efficient Transfer Learning
Swin	Swin Transformer
Swin-Tiny	Swin Transformer Tiny
TCN	Temporal Convolutional Network
TeCNO	Temporal Convolutional Network with Online prediction
Top-K	Top-K Accuracy
TP	True Positive
TP_c	True Positives for class c
VGG	Visual Geometry Group Network
ViT	Vision Transformer
ViT-B	Vision Transformer Base
ViT-Small	Vision Transformer Small

ViT-Tiny	Vision Transformer Tiny
V&L	Vision & Language
VLM	Vision–Language Model
VLP	Vision–Language Pretraining
ZS	Zero-Shot

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Background

The idea of multimodal (or cross-modal) learning aims to combine information between data modalities that we learn through different encoders in the same embedding space, and semantically related representations that are closely similar to one another. (Rao, Qin, Kolouri, Wu, & Moyer, 2024; Zhang et al., 2022), In CLIP (Contrastive Language-Image Pre-training), this is achieved by a dual-encoder structure in which an image encoder transforms visual data to feature vectors and a text encoder transforms natural language descriptions to the same space. Paired image-text examples are drawn toward one another and pushed away during training with a contrastive loss to allow zero-shot transfer by the use of a limited number of text examples without training a task-specific classification head. (Perez, Nwoye, Kermani, Mohareri, & Jamal, 2025a; Radford et al., 2021; Rao et al., 2024)

The vision backbone is also an important element in this paradigm, in that it determines what set of spatial structure, semantic content and inductive biases can be used by the joint space. Different architectures present different kinds of properties: pure Transformer encoders such as ViT emphasize global token interactions, modern convolutional designs such as ConvNeXt preserve excellent locality and feature hierarchy extraction20 as well as swinners can be viewed as window-based transformers, which combine both the attention mechanism and multi-scale context information.

In that direction we suggest lightweight backbones (e.g. ViT-Tiny/Small, ConvNeXt-Tiny, Swin-Tiny) to reduce the number of parameters and memory requirements as well as latency to allow them to be deployed on edge devices, cloud -computing clusters and time-sensitive applications. (He et al., 2025; Perez et al., 2025a) In most practical settings, such as medical imaging and surgical video -fine-tuning such backbones is not practical: only shallow heads or alignment layers are fine-tuned .

More recently, contrastive and vision language models have begun to be evaluated in the tasks of endoscopy and surgery including CLIP-style pre-training on colonoscopy records, large scale surgical vision-language models, and generally puny surgical VLMs(He et al., 2025; Kostiuchik et al., 2024; Yang et al., 2024a) Most of the related papers only study large-scale backbones or fine-tuned models and not compare them with frozen lightweight encoders (e.g., ViT-Tiny/ViT-Small/ConvNeXt-Tiny/Swin-T Therefore, even the latest vision backbones due to their no-fine-tuning to accuracy and efficiency tradeoffs are yet to be well defined in historical clinical workflows.

1.2 Problem Statement

This paradigm is now popularly implemented as CLIP-style multimodal learning where images and text are aligned with a dual-encoder architecture trained with a contrastive loss where vision backbones such as ViT [dosovitskiy2020image] (ViT-B/16 in particular) are typically used followed by a frozen text encoder(the latter is commonly pre-trained on image-caption pairs on the web).(Radford et al., 2021) Although the paradigm is practical on natural image benchmarks, it is computationally expensive when it is applied to long and high frame rate

Recent vision-language model surgical vision-language models (such as colonoscopy-specific Endo-CLIP), massive surgical vision-language models (SurgLaVi), and general-purpose VLMs (also parameter-efficient transfer methods) in recent years have begun to explore contrastive or prompt-based models of surgery and endoscopy(Kostiuchik et al., 2024; Zhang et al., 2022), but all are currently limited, as they require large backbone, fine-tuning directory on the encoder and change many components at the same time (backbone, temporal head, loss and dataset) making it harder to characterise the specific contribution of vision Thus, no controlled apples to apples comparison of recent lightweight frozen encoders -i.e. ViT-Tiny, ViT-Small, ConvNeXt-Tiny and Swin-Tiny - with the fixed CLIP-like structure on frame-level surgical phase recognition currently exists.

In this case, it is difficult to practitioners with tight GPU constraints or requiring real-time surgery and can therefore not fine-tune models on full surgical videos on high-capacity models,

though the accuracy-efficiency trade-off of fine-tuning on other lightweight backbones is not well studied. To be more precise, it is not clear in the frozen ViT-Tiny/Small, ConvNeXt-Tiny and Swin-Tiny whether they are of retrieval quality (Recall-K), recognition performance (Top-K accuracy, macro-F1) and efficiency (e.g., parameter count, memory usage, training time and inference throughput). (Kostiuchik et al., 2024; Perez et al., 2025a; Zhang et al., 2022) This thesis addresses that gap by performing a controlled evaluation of these frozen encoders within a standard CLIP-style framework—keeping the text tower, prompts and contrastive loss fixed—to determine their relative effectiveness for frame-level surgical phase recognition and to characterise the resulting accuracy–efficiency trade-offs under realistic clinical hardware constraints.

1.3 Motivation

In realistic clinical and research settings, they are limited in the amount of devices available and have to be shared between different projects(Kostiuchik et al., 2024), and endoscopic scans do regularly generate long, high frame rate video data that is costly to process.(Radford et al., 2021) Fully fine-tuning large CLIP-style models with heavy vision backbones on this data is often not possible due to memory constraints, low throughput (frames per second) or the inability to simultaneously use multiple models on the same physical device.

Smaller encoders are also cost and sustainability-wise appealing. By minimizing the number of parameters and reducing memory footprint, it is possible to reduce training time, reduce energy use, and reduce the cost to deploy and maintain models, which is essential where the systems must be updated to implement new cameras, procedures, or institution-specific data changes.(Yang et al., 2024a) In addition, real-time or near-real-time surgical assistance (e.g. intraoperative guidance, alerts, or phase-aware user interfaces) demands hard constraints on latency: when inference is too slow to reliably make a decision when needed, it is necessary to reduce the complexity.

In addition to practical deployment, smaller frozen encoders like ViT-Tiny/Small, ConvNeXt-Tiny and Swin-Tiny are a useful lens through which to study the behaviour of architectural inductive biases (global token mixing, convolutional hierarchies, shifted windows) on contrastive alignment and downstream performance in medical video.

1.4 Significance of the Study

- **Frozen encoders comparison:**

  This work makes an apples-to-apples comparison of four light vision backbones (ViTTiny, ViT-Small, ConvNeXT-Tiny and Swin-Tiny) used without fine-tuning within a simple CLIP-style environment by fixing the text branch and contrastive objective(He et al., 2025; Perez et al., 2025a) By freezing all encoders in the study we specifically examine how backbone architecture alone is applied to zero-/few-shot surgical stage recognition and video-text retrieval; shedding light on how contemporary frozen model behaviour behaves.

- **Incremental retraining and resource-sensitive training:**

  This is along with the more recent parameter-efficient transfer and domain-adapted vision-language pretraining algorithms that emphasize shorter iteration in addition to a more query-friendly and target-friendly output use of compute. (He et al., 2025; Yang et al., 2024a)

- **Discussion of evidence on the accuracy-efficiency tradeoff:**

- The paper continues to tabulate results on performance metrics (Top-K accuracy, macro-F1, Recall@K) and efficiency metrics (parameter count, GPU memory usage, training time per epoch, inference FPS) among all frozen models thereby giving a snapshot of the accuracy-efficiency trade-off in surgical stage recognition(Kostiuchik et al., 2024; Zhang et al., 2022) This gives us the picture of when compact frozen models are good enough, when adding capacity does not help much and therefore information that is largely missing in current medical vision-language benchmarks.

- **In the case of medical AI systems applications:**

  The final section will concern relating the findings to research implications on AI implementation in healthcare: recommendation regarding the selection of encoder with latency/cost trade-offs, recommendation regarding stabilization behaviors with zero-

/few-shot settings with the help of batching and prompt design, insight into how frozen-encoder pipelines can be scaled between research prototype systems and hospital IT resources with limited but shared resources.

1.5 Research Questions

In what pretrained image encoder architectures (ViT-Tiny, ViT-Small, ConvNeXt-Tiny, swin-Tiny) is the best embodiment of the visual representations to surgical phase detection when fine-tuned as frozen feature extractors?

1.6 Research Objectives

This study was conducted with the following objectives:

- To integrate four pretrained image encoders (ViT-Tiny, ViT-Small, ConvNeXt-Tiny, Swin-Tiny) into a surgical phase recognition pipeline as frozen feature extractor.
- To evaluate and compare their performance on Cholec80 using top-k accuracy metrics under identical conditions.
- To compare the change in architecture (pure transformer vs. hybrid conv-transformer vs. convolutional backbone) with the impact of inductive biases on the change in architectural performance before fine-tuning.
- To provide practical recommendations for choosing pretrained encoders in medical AI systems where fine-tuning is impractical or impossible.

.

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

- In thesis, frame-level surgical phase recognition is considered only. It does not build or compare time models (such as 3D CNNs, wizards of long-range memory(?)or sequence smoothing). Any form of frame sampling is just of a creation of the frame sets not in modeling a temporal dependencies(Kostiuchik et al., 2024).

- The experiments are limited to ViT-Tiny, ViT-Small, ConvNeXt-Tiny and Swin-Tiny as transferable CLIP vision backbones. This paper does not consider larger ViTs (i.e., ViT-Base/ViT-Large) ResNets, hybrid encoders or video encoders (Perez et al., 2025a; Schmidgall, Cho, Zakka, & Hiesinger, 2024a).

- The text encoder is neither fine-tuned, prompt-tuned, LoRA or instruction-style adapted. Text responses are pre-uniformed to prompts that relate to the phases of surgery; the projection head and contrastive arrangement follows the CLIP base pattern (Radford et al., 2021).

- It is assessed on Cholec80 or an equivalent frame set, sampled on its videos, with its splits, and is preprocessed. There is no additional usage of surgical datasets, cross-dataset transfer or multi-procedure fusion(Kostiuchik et al., 2024; Perez et al., 2025a).

- The quality of alignment is measured with video-text (frame-text) retrieval by calculating Recall@K and zero-shot phase classification with Top-K accuarity and macro-F1. It can be topped with few-shot recognition (cross-entropy); however, no more advanced metrics such as mAP or calibration or robustness to distribution shift or clinical user studies are used(He et al., 2025; Schmidgall et al., 2024a).

- It is concerned with such a narrow analogy with lightweight encoder on a typical CLIP pipeline, as this ensures that we can interpret our findings and directly trace them to the backbone structure(He et al., 2025; Radford et al., 2021).

1.7.2 Limitations

- It is not longer than medical-specific baselines (e.g. TeCNO, EndoViT or other task-engineered pipelines); results are plotted in single isolation relative to the miniaturized CLIP encoders (can not be demonstrated to be better than dedicated architectures)(Kostiuchik et al., 2024; Schmidgall et al., 2024a).
- The CLIP text encoder remains frozen: it is not prompt-tuning or data-augmenting text, nor training a domain lexicon, nor teaching the language branch instructively—it does not involve fine-tuning in any way(Radford et al., 2021).
- It does not utilize any form of temporal modeling and only considers the knowledge of phases on the frame level with no LSTMs/GRUs, TCNs, temporal transformers or sequence smoothing/post-processing(Kostiuchik et al., 2024).
- The data is confined to the existing surgical video (e.g. Cholec80 or frames of Cholec80), and no cross-dataset or cross-center validation is conducted; hence, the generalization claims are only local and not global(Kostiuchik et al., 2024; Perez, Nwoye, Kermani, Mohareri, & Jamal, 2025b).
- Multimodal pretraining/fine-tuning Multimodal pretraining/fine-tuning does not study masked image/language modeling, captioning, distillation or multi-task learning; instead, it studies contrastive objective (optionally with cross-entropy for few-shot recognition)(He et al., 2025; Radford et al., 2021).

1.8 Thesis Organization

The aim of the proposed task setting is explained in this introductory section: why do we even need frame-level understanding of surgical phases in the first place, Why we just swap in four more lightweight vision backbones (ViT-Tiny, ViT-Small, ConvNeXt-Tiny-Swin-Tiny)? In it there is a problem, motivation, significance, research questions/objectives and (a clear!) scope (no time modeling, no text fine-tuning, only Cholec80 frames), and the earnest limitations.
This is followed by a brief review of related work literature. You will find out why astute image-text learning allowed CLIP to score highly in zero-shot, why it is hard to learn a surgical video, and what the history of medical VLP projects and lightweight backbones has to do with it. Our conclusion: we drag it into the limelight: no apples-to-apples ratio of little encoders of a traditional CLIP system to surgical frames.

The second issue on the list is the data and methods. We outline the sampling and streamlining of a frame sampled by Cholec80, the generation of phase prompts, and scaling CLIP with a removable vision slot. The training is performed on two pathways: contrastive multimodal pretraining and few-shot cross-entropy on constant splits, phase prompts, and hyperparameters. Efficient Object-Level Metric Learning (OML)³+ 1:- Object-Centric: The topology of all the objects that are being evaluated should be well described. Constraints: 10 factors are to be investigated. 5.2: Model-Efficiency.metrics = [Top-1/Top-5/Top-10, Macro-F1, Ret@K].

The most substantial data lies in the results part of the study i.e. zero-shot and few-shot results with all four encoders, quality of retrieval with seen and held-out splits, and per-class insights where needed. There is an accuracy-efficiency frontier, and you can see where a smaller model can be used not only in the normal condition but also on patients with simple stressors (blur, smoke, light deprivation) or the addition of a device-to-hospital shift. The fashions are rationalized by illustrations (confusions, embeddings/attention).

And this we make conclusions out of findings. You will be in a better position to understand the type of encoders to select in a given latency/budget constraint, encouragement/batching guidance proposals in how to stabilize zero-/few-shot behavior, and a minimalist retraining playbook for new devices/procedures. And the final thing to consider is limitations (including deployment observations and possible follow-up, such as time modelling, language-side adaptation, and cross-dataset validation).

CHAPTER 2

LITERATURE REVIEW

2.1 Background on Surgical Workflow Analysis and Multimodal Learning

Surgical Phase Recognition (SPR) (also known as phase recognition and gesture spotting) is a key task in Surgical Workflow Analysis: Given a sequence of endoscopic videos, the task is to determine what step or phase it is undertaking, often at the frame level. (Kostiuchik et al., 2024)Timely Surgical Phase Recognition Surgical Phase Recognition (SPR) (synonymous with phase recognition and gesture spotting) (Kostiuchik et al., 2024) Given a sequence of endoscopic video frame sequences, the goal is to detect what step or phase the sequence is currently performing, often at the frame level. It may also be used to enhance surgical safety with phase-aware checks and early anomaly detection in case the observed sequence does not match what was anticipated. (Kostiuchik et al., 2024); as hospitals have predictable schedules of the phases, operating-room scheduling, turnover and staffing can all be optimized, which would help overall throughput. (He et al., 2025; Schmidgall et al., 2024a; Zhang et al., 2022) But despite this breakthrough, most systems are dependent on specific visual backbones, and which encoders to use is not well understood. (Perez et al., 2025b; Schmidgall et al., 2024a)However, despite all this progress, most systems have been limited to specific visual backbones(Zhang et al., 2022).

Multimodal learning Multimodal learning models are targeted at training models capable of coordinating information between dissimilar types of data, i.e. in this case, a pair of endoscopic frames and a phase description - i.e. to match endoscopic frames with phase descriptions(enable zero/few-shot recognition and efficient re- trieval). (Radford et al., 2021) The vision encoder used can be scaled to large scale, and scales with less com-putation to high-end tasks like cross-modal matching, which is expensive with dense annotations. 2-4 The vision(He et al., 2025; Perez et al., 2025b; Schmidgall et al., 2024a)

2.2 Deep Learning Approaches for Surgical Phase Recognition

The first Surgical Phase Recognition (SPR) systems were built on a small set of standard datasets - i.e., Cholec80, M2CAI16 and Cataract-101 and the EndoVis challenges—annotated with phase labels or per-step/skill labels to enable both supervised learning and cross-paper re-evaluation(Kostiuchik et al., 2024). Minimal baseline CNN-only image encoders (e.g., VGG, ResNet) would classify separate images alone, giving it the look, yet none or minimal temporal

information. Smoothness of the surrounding frames was enhanced immediately when it was combined with LSTMs/GRUs, but the training time and latency of the combination were greater than the one trained with CNN feature extractors to imitate the flow of the procedure(Kostiuchik et al., 2024). Later approaches had viewed Temporal Convolutional Networks (TCN) and Transformer-based heads as a representation of longer-range interactions and ad hoc smoothness of prediction stabilization in others(Kostiuchik et al., 2024; Perez et al., 2025b). Despite that, even with advances, these pipelines could still remain completely dependent on the input of vision, with no textual reminders or captions or descriptions of the procedures; only the appearance of the models would overfit on the appearance cues and fold during unusual phases or transitions between devices and domains. More compute demand was also needed due to more deep backbones and longer temporal windows, which was impossible with resource-poor hospitals and edge installations(Kostiuchik et al., 2024; Perez et al., 2025b). Complexities of the scenes (smoke occlusions, highlights on the speculars) could now be allowed to change with time, producing phase flickers that had no semantic antecedent. Further, the extent of generalization between hospitals and cameras other than the one utilized in training was constrained, and degradation in performance was observed in the case of another test scan protocol or data(Schmidgall et al., 2024a). Current vision-language approaches may provide a path out: with frame and textual-stage description projection to a shared embedding space, we can add some procedural semantics, we can improve retrieval and few-shot recognition, and we can be taking steps toward more data-efficient SPR(He et al., 2025; Radford et al., 2021; Schmidgall et al., 2024a).

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

Paper	**Dataset**	**Method / Approach**	**Image encoder used**
Early CNN baselines (VGG/ResNet)[1]	Cholec80	Frame-wise CNN classification (supervised CE)	VGG / ResNet
CNN + LSTM/GRU pipelines[1]	Cholec80	CNN features + recurrent temporal modeling (CE)	• **VGG/ResNet** frame encoder + **LSTM/GRU**
TeCNO	Cholec80, M2CAI16	Dilated Temporal CNN over CNN features; temporal smoothing	**ResNet** (frame features) + **TCN** head

EndoViT	Cholec80 (also Cataract-101)	Vision Transformer (often with temporal head); supervised CE	- **ViT** (Transformer)
SurgLaVi (CLIP baselines)	Large multi-procedure surgical video–text	Dual-encoder vision–language (contrastive image–text)	- **CLIP** vision tower (e.g., **ViT-B/16** or **RN50**)
Endo-CLIP	Colonoscopy (GI endoscopy)	Domain-adapted CLIP with progressive curriculum (VLP)	• **CLIP-style** vision tower (ViT/ResNet family, adapted)
GP-VLS	Mixed surgical benchmarks (phase/step datasets)	- General-purpose surgical VLM (contrastive VLP across tasks)	• **ViT-based** VLP encoder

2.3 Vision Encoders for Multimodal Surgical Understanding

Vision encoders encode raw images into dense image representations of edges, textures, tools and other anatomical features—the spatial descriptors that a multimodal model is used to map to text(Kostiuchik et al., 2024). Model CNN-based encoders (e.g. ResNet and its extension ConvNeXt) encode local patterns of a hierarchical pattern in a series of layers of convolution; they are parameter-light, cache-friendly and encode small-scale patterns, but not long-range correlations, unless further support is provided by other models. Transformer-based encoders (e.g., ViT, Swin) encode pictures as tokens and use self-attention to pull information across the image(Kostiuchik et al., 2024; Radford et al., 2021). They are useful at the scale of the structure in the scene and the tooltissue setting and can be both memory intensive and slow without particular design. The mentioned trade-off is that one, which is replaced by ViT-Tiny/Small, ConvNeXt-Tiny, and Swin-Tiny with a typical CLIP configuration in the specified project.

Networks that encode raw images in rich feature representations of edges, textures and tools and hints of anatomy (or not)—the spatial representations that its multimodal system maps to text are known as vision encoders. Hierarchy of local patterns CNN-based encoders (e.g. between ResNet and ConvNeXt) are parameter-efficient, cache-friendly, and able to capture fine spatial detail, and typically do not do well to capture long-range dependencies with the exception of co-location with other modules. Encoder-only transformer (e.g., ViT, Swin) Architectures can offer alternative scene-level structure and tool-tissue context by tokenizing images and globally combining information with self-attention, but without attentive design, attention is not only memory-intensive but also slow. Swin minimizes this by specializing on shifted windows and trading locality and greater range context, but ViT offers a clean, scalable layout that can also be valuable to the scale of data as well as to successful regularization. Precisely what trade-off is incurred when replacing ViT-Tiny/Small, ConvNeXt-Tiny and Swin-Tiny with a vanilla CLIP setup is pointed out at the end of this section(He et al., 2025; Perez et al., 2025b; Radford et al., 2021; Schmidgall et al., 2024a)

The latest ones are pure visual temporal and pretraining vision-language. TeCNO applies dilated Temporal CNNs to Cholec80/M2CAI16, which is trained with a frame-level CNN encoder using smoothness and temporal conv, and in fact achieves better phase accuracy/F1 by learning sequence structure, but is mono-modal, and expensive at long horizons and sensitive to cross-centre shift it suggests us that much simpler small encoders in a multimodal system can also do the same with

much less overhead. SurgLaVi Transition to multimodality Learning SurgLaVi trains big hierarchical video-text collections of more than 200+ procedures and publishes(Perez et al., 2025b) CLIP-style baselines with Recall@K and zero/few-shot Top-K results; it is able to transfer, however, with only a single backbone (rather than multiple) and without investigating the systematic relevance of lightweight vision encoders - the same thing we identify in default CLIP(Perez et al., 2025b; Schmidgall et al., 2024a; Zhang et al., 2022).

2.4 Limitations and Research Gap

However, the existing literature regarding the SPR and surgical VLP has formidable outcomes of benchmarks(Kostiuchik et al., 2024), but still relies on gigantic encoders of images (e.g., ViT-B/L), which discourage its application to the routine work of the hospital due to the high running time and dedicated equipment. Very few studies investigate lightweight backbones (ViT-Tiny/Small, ConvNeXt-Tiny, and Swin-Tiny) in a uniform scenario of multimodality and the trade-off between precision and efficiency has not been thoroughly investigated(He et al., 2025; Kostiuchik et al., 2024). The multimodal SOTAs on surgical video which are currently available are in a relatively small number compared to the natural-image domains, and even in the cases where they exist, are usually not concerned with considering compact alternatives(He et al., 2025; Perez et al., 2025b), as they are instead busy with resolving the assignment of a single backbone. Also, limited sample range per se (most of the studies involved Cholec80 only or one procedure at a time) and extrapolation to other hospitals, device and surgeon representatives. At least practically we find that the rate of inference is the rate which is not a first-class See Appendix A. Although there is re-definitely a need in real time or near real time support in the OR, we are not maximizing this metric since memory footprint will grow in direct proportion to the backbone size and temporal stacks(He et al., 2025; Kostiuchik et al., 2024). Finally, most pipelines prefer temporal heads (TCN \& Transformers) to frame level multimodal alignment and thus lack the chance to inject procedural semantics in the text in the fine grain of a frame(He et al., 2025; Perez et al., 2025b). These constraints are the motive behind acomparation of multimodal encoders, and to that extent the study under consideration is guided(Perez et al., 2025b; Schmidgall et al., 2024a).

Nevertheless, few studies have been done to deal with this in zero-shot recognition and retrieval across domains(Radford et al., 2021). Dual encoders in the CLIP style demonstrate that matching images with text in a common embedding space may be used both in the past and in the present. Simultaneously with these papers, domain-specific models of surgery such as

TeCNO (temporal CNNs) [40] and EndoViT (transformer-based network) have achieved progress in phase recognition on datasets based on Cholec80 and on M2CAI16 though are purely vision models and supervised by larger encoders(Kostiuchik et al., 2024). What that paper still lacks is a proper apples-to-apples comparison of lightweight vision encoders (ViT-Tiny/Small, ConvNeXt-Tiny, Swin-Tiny) in an off-the-shelf CLIP setup normalized to surgical frames and holding the text branch fixed and the evaluation constant(He et al., 2025; Kostiuchik et al., 2024; Perez et al., 2025b). Farther afield, CLIP based learning in the medical domain has barely been explored on the frame level, and little of the innovation sounds in the few domain-adapted VLP efforts, such as Endo-CLIP5 and excessive video- Even in the real world, hospitals and labs desire efficiency—models that are low-memory, high-throughput, and can be trained with only small quantities of data—that can run on hardware that is not state of the art; a requirement supported by recent parameter-efficient transfer work(Yang et al., 2024b). Condensing these findings into one, a disconnect between efficiency and multimodal fusion on one hand and surgical suitability on the other can be felt. This is the gap we are dealing with in the present paper(He et al., 2025; Perez et al., 2025a; Radford et al., 2021; Schmidgall, Cho, Zakka, & Hiesinger, 2024b; Yang et al., 2024b).

2.5 Research Focus and Conceptual Framework

We further compare the four trained vision encoders ViT-Tiny, ViT-Small, ConvNeXt-Tiny and Swin-Tiny, with each backbone being trained individually on the same random datasets, and without the same preprocessing and optimisation conditions(Radford et al., 2021). All models are trained using frozen encoders: the CLIP text tower, and the vision backbone are fixed, only the projection/alignment layers on the image and text embeddings are trained, and performance metrics are reported directly against each other to demonstrate the encoder selection and not the training configuration causes the variation in performance and qualitative analysis of common failure modes (e.g. mixing neighbouring phases with each other)(Kostiuchik et al., 2024; Perez et al., 2025a).

In training, textual phase prompts are coded by the frozen CLIP text encoder and frame tensors are inputted to one of the frozen vision backbones (ViT-Tiny, ViT-Small, ConvNeXt-Tiny or Swin-Tiny) to obtain visual features and a symmetric contrastive CLIP loss is optimised to bring similar frameprompt pairs closer and dissimilar ones further apart, similar to multimodal alignment. This experimental design is capable of pursuing a study of the effect of different

frozen modern vision backbones on the quality of multimodal alignment, retrieval and phase-recognition performance, and computational efficiency of surgery phase recognition without fine-tuning.This experimental design allows a pursuing study of the influence of various frozen modern vision backbones on the quality of multimodal alignment, retrieval and phase-recognition performance, and computational efficiency of surgical phase recognition without fine-tuning (Kostiuchik et al., 2024; Schmidgall et al., 2024b).

CHAPTER 3

METHODOLOGY

The chapter goes ahead to give a proposal of a frame-level controlled CLIP-like method of understanding surgical phases at the frame level. We start by defining our dataset and preprocessing the long laparoscopic videos that are converted into normalized tensor frames of the respective scaled phase prompts. We also present a two-encoder model, with the visual encoder optional (ViT-Tiny, ViT-Small, ConvNext-Tiny-Swin-Tiny and Swin-Tiny) and the CLIP text transformer fixed, which projects them into a common 1536-dimensional joint embedding space, which they are trained on with a symmetric contrastive (InfoNCE) loss to match images and text as they match in CLIP to test image-text matching on a large scale and to generalize to new concepts.

Our dataset was the Cholec80 laparoscopic cholecystectomy 80 cases corpus (7 phases), and it should be regarded as the source of information about the analysis of the surgical workflow given a weighting by the coverage/redundancy to be loaded by our viewer. pt tensors to load the dataset and load it consistently. They are geared to the long low contrast and procedure-organized form of surgical video and test with a number of encoders simultaneously at an even playing field. Visual content and surgical language in VLP work are taken into consideration in the modern surgical VLP work, which is why it is what makes us promptly design our text-prompt phase design.

Our model adheres to the CLIP architecture: an image encoder, resulting in a visual encoding; a pre-trained text transformer encoding layer, the result of which is prompts; lightweight linear heads, which project to D-dimensional space, and on which a cosine similarity-based symmetric promotion of a cross-entropy between all pairs of an image and texts of a mini-batch. It is a non-directional form of modeling that is not only data- but also compute-efficient and proved to be useful at zero-shot classification and retrieval by aligning modalities in the joint space.

Reconfigurable Backbones of Vision.

In the remaining experiments, to experiment on the role of the visual encoder alone, we replace it with four lightweight backbones:

- ViT-Tiny / ViT-Small (patch-based self-representing with receptivity fields of the

world);

- ConvNeXt-Tiny (convolutional neural network with strong locality priors and extremely efficient);
- Swin-Transformer-Tiny (hierarchical, window-based token-to-token attention + linear complexity with respect to length of image)

They are inductive biases in these families that are complementary (local vs. global/windowed attention) but can prove to be computation-wise important to clinical uses.

The only other elements required to map outputs of the encoders to a common space of cosine-similarity-based retrieval and zero-shot classification are a linear layer in both branches. The projection layers (and optionally, the visual backbone) are only fine-tuned, and the text branch is frozen in such a way that it solidifies a clean comparison across visual backbones as well as is consistent with the CLIP application as a zero-shot.

The InfoNCE loss of CLIP symmetric is trained, and the pairwise image-text similarity matrix (cosine, divided by temperature) is trained. To determine the fairness, we optimize all hyperparameters (i.e., optimizer, batch size, learning rate, epochs, weight decay and gradient clipping) for all the encoders. Checkpointing and retraining each time we find an encoder stores the best model per encoder, and our incremental retraining algorithm can be used to allow low-overhead retraining with the arrival of the next video, which does change with time in a real clinical pipeline.

Two competences, (1) multimodal alignment with Recall@K (Kin{1,5,10}), wherein an image is rated based upon an image-text recall, and (2) zero-shot phase recognition, wherein an image is ranked among all of the phase prompts with regards to similarity (Top-1/5/10), are evaluated. We also report precision, recall, macro-F1 (to report imbalance of the classes), confusion matrices (to investigate the errors on the phase level), and efficiency (throughput, parameters and memory profile). It is a tradeoff of what CLIP could accomplish effectively (lookup, zero-shot) and what would also be feasible aspects of the OR-near systems (latency and memory consumption).

We also share data splits, preprocessing, prompts, and loss and optimization parameters among all encoders such that we can only do apples-to-apples comparisons of visual backbones. The particular design involved is to explore the impact of inductive bias of backbones on the quality of multimodal alignment and on the quality of surgical phases. VLP New works emphasize the

significance of the standardized pipelines and language quality control as the primary method; our fixed text branch and fixed training recipe comply with the recommendation, and the main object of analysis is the side of vision.

Massive surgical VLP techniques (e.g. hierarchical data and CLIP-style pretraining)—surgery-specific language and video combined to improve zero-shot transfer are demonstrated, but model and data efficiency are required in actual hospitals. All these principles are found in our chapter approach: compact encoders, thoughtful preprocessing and contrastive alignment, in a manner that certain knowledge can be used to construct the workable, deployable medical VLMs.

3.1 Dataset Description

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 multimodals 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 Preprocessing and Data Augmentation

Video Preprocessing

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

Text Preprocessing

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

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

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

•

Data Augmentation

- In order to overcome the data imbalance problem and make the data more robust:
- Coping with Class Imbalance: Under-represented phases are oversampled to bring about the equilibrium of all classes when training.
- Random Cropping: It adds spatial invariance, like the model can be tolerant to changes of viewpoint and the location of tools.
- Flipping: Horizontal flipping is done in random fashion; this can be performed during training in order to mimic random camera orientation as well as tool side that would assist the network to generalize better between different conditions in the surgery.
- Other Augmentation Rotation, jittering of color, zoom-in, etc., are other types used to assist the model to operate in different conditions of lighting and camera angle in surgery.

3.3 Multimodal Architecture (CLIP-based Dual Encoder)

Our dual encoders are designed in such a manner that the text and the image are executed on separate towers, and subsequently they are paired on a common embedding. The image encoder creates a pluggable slot whereby we can place the backbones with varying labels (4 in this work: ViT-Tiny, ViT-Small, ConvNeXt-Tiny, and Swin-Tiny) to investigate the effect that the design choice of the backbone has on the quality and efficiency of alignment. The output of any image encoders is a fixed-length encoding that is projected to the multimodal space through a common layer of projection. The text encoder is the frozen CLIP transformer: it converts standardized (compressing) phase-prompt normalized in text encodings and is neither fine-tuned nor text-side text-prompt tuned.

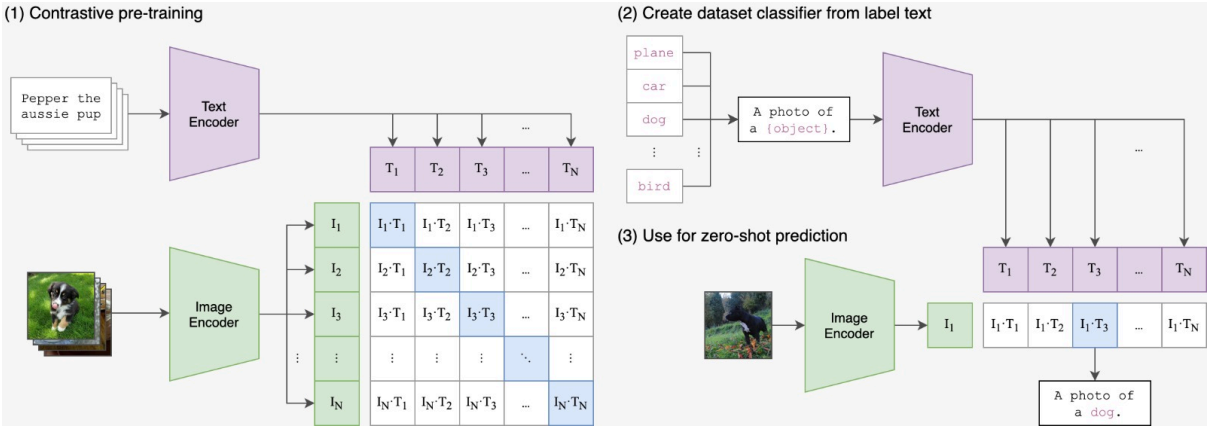

Figure 3.1 Clip Architecture

Formally, let xix_ixi be a frame and yjy_jyj a phase prompt. The vision and text towers produce features $f_v(x_i) \in \mathbb{R}^{d_v}$ and $f_t(y_j) \in \mathbb{R}^{d_t}$ (text tower frozen). A shared dimensionality d is obtained via linear projections $W_v \in \mathbb{R}^{d \times d_v}$ and $W_t \in \mathbb{R}^{d \times d_t}$ (the only trainable components):

$$\tilde{z}_i = W_v\, f_v(x_i), \tilde{t}_j = W_t\, f_t(y_j), z_i = \frac{\tilde{z}_i}{\| \tilde{z}_i \|_2}, t_j = \frac{\tilde{t}_j}{\| \tilde{t}_j \|_2}.$$

We compute cosine similarities $s_{ij} = z_i^\top t_j$ and optimize the **symmetric InfoNCE/CLIP loss** over a batch $\mathcal{B}$ with temperature $\tau > 0$:

$$\mathcal{L} = \frac{1}{2}\Big[\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \underbrace{-\log \frac{\exp(s_{ii}/\tau)}{\sum_{j\in\mathcal{B}} \exp(s_{ij}/\tau)}}_{\text{image}\to\text{text}} + \frac{1}{|\mathcal{B}|}\sum_{j\in\mathcal{B}} \underbrace{-\log \frac{\exp(s_{jj}/\tau)}{\sum_{i\in\mathcal{B}} \exp(s_{ij}/\tau)}}_{\text{text}\to\text{image}}\Big].$$

At inference, **zero-shot classification** predicts the phase for a frame x by

$$\hat{y} = \arg\max_j \; z(x)^\top t(y_j),$$

The matching of frame prompts by learning to bring together frame-prompt pairs that are matched and move apart those that are mismatched with the contrastive CLIP loss does the alignment. To facilitate fair comparisons, we make no modifications to text branch, prompts, projection dimensionality and loss but use individual models with each image backbone on the same data splits and preprocessing. At inference, joint space can be combined with zero-shot phase classification (prompt-based score) and frame-text retrieval (Recall@K) types of tasks, where we can separate the performance and quantify the change as due to a visual backbone change and not due to change elsewhere in the pipeline.

3.4 Vision Encoder Variants

(a)ViT-Tiny.

We evaluate a **smallest-scale Vision Transformer** with **16×16 patch embedding**. An input frame $x \in \mathbb{R}^{H\times W\times 3}$is split into non-overlapping patches and linearly projected to tokens; a class token is prepended and tokens are processed by L layers of multi-head self-attention (MHSA) and MLP blocks. For a token matrix $X \in \mathbb{R}^{N\times d}$, MHSA computes softmaxsoftmax$\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$ where $Q = XW_Q$, $Q = XW_Q$, $V = XW_V$. ViT-Tiny prioritises **speed and low memory**; its **lower capacity** can limit fine-grained semantics in cluttered endoscopy, but it often yields the **best throughput** in our pipeline.

(b)ViT-Small.

This variant increases **embedding width and depth** (more heads/layers), keeping **16×16 patches**, improving representational richness while remaining compact. The projection to the shared CLIP space is $z = W_v f_v(x)/\| W_v f_v(x) \|_2$. Compared to ViT-Tiny, ViT-Small typically offers **higher zero-/few-shot accuracy** and **better per-class balance**, at a

moderate cost in latency and parameters—useful when hospitals can afford a slightly larger footprint.

(c)ConvNeXt-Tiny.

A **modern CNN** that revisits ResNet-style hierarchies with larger kernels, depthwise convolutions, LayerNorm, and simplified stages. Convolutions implement local mixing $Y = X * \Theta$ with strong **spatial priors** that suit **tool edges and tissue textures** common in laparoscopic frames. In our CLIP head, features $f_v(x)$from the final stage are global-pooled and projected to the joint space. ConvNeXt-Tiny balances **high efficiency** with **robust locality**, often excelling on frames where cues are small or partially occluded..

(d)Swin-Transformer-Tiny.

A hierarchical Transformer using **window-based attention** with **shifted windows** that alternate between blocks, capturing locality while allowing cross-window information flow. Within a window, attention is standard MHSA; hierarchy is built via patch merging, yielding multi-scale features well-suited to **organ context + tool details**. After global pooling$f_v(x)$is mapped by the shared projection to align with text embeddings. Swin-Tiny is typically **efficient and powerful**, offering a middle ground between ViT's global modeling and CNN locality.

Shared CLIP projection & loss (all encoders).
For image features $f_v(x)$ and frozen text features $f_t(y)$, we learn only the projections $W_v, W_t$and optimise the symmetric CLIP loss:

$$z = \frac{W_v f_v(x)}{\| W_v f_v(x) \|_2}, t = \frac{W_t f_t(y)}{\| W_t f_t(y) \|_2}, \mathcal{L} = \frac{1}{2}\left[\mathrm{CE}\left(\frac{\mathrm{sim}(z,t)}{\tau}\ \text{row-wise}\right) + \mathrm{CE}\left(\frac{\mathrm{sim}(z,t)}{\tau}\ \text{col-wise}\right)\right].$$

We train a separate model per encoder under identical splits/prompts/hyperparameters so that any change in zero-shot accuracy, Recall@K, confusion patterns, or efficiency (throughput/params) can be attributed to the visual backbone rather than confounds.

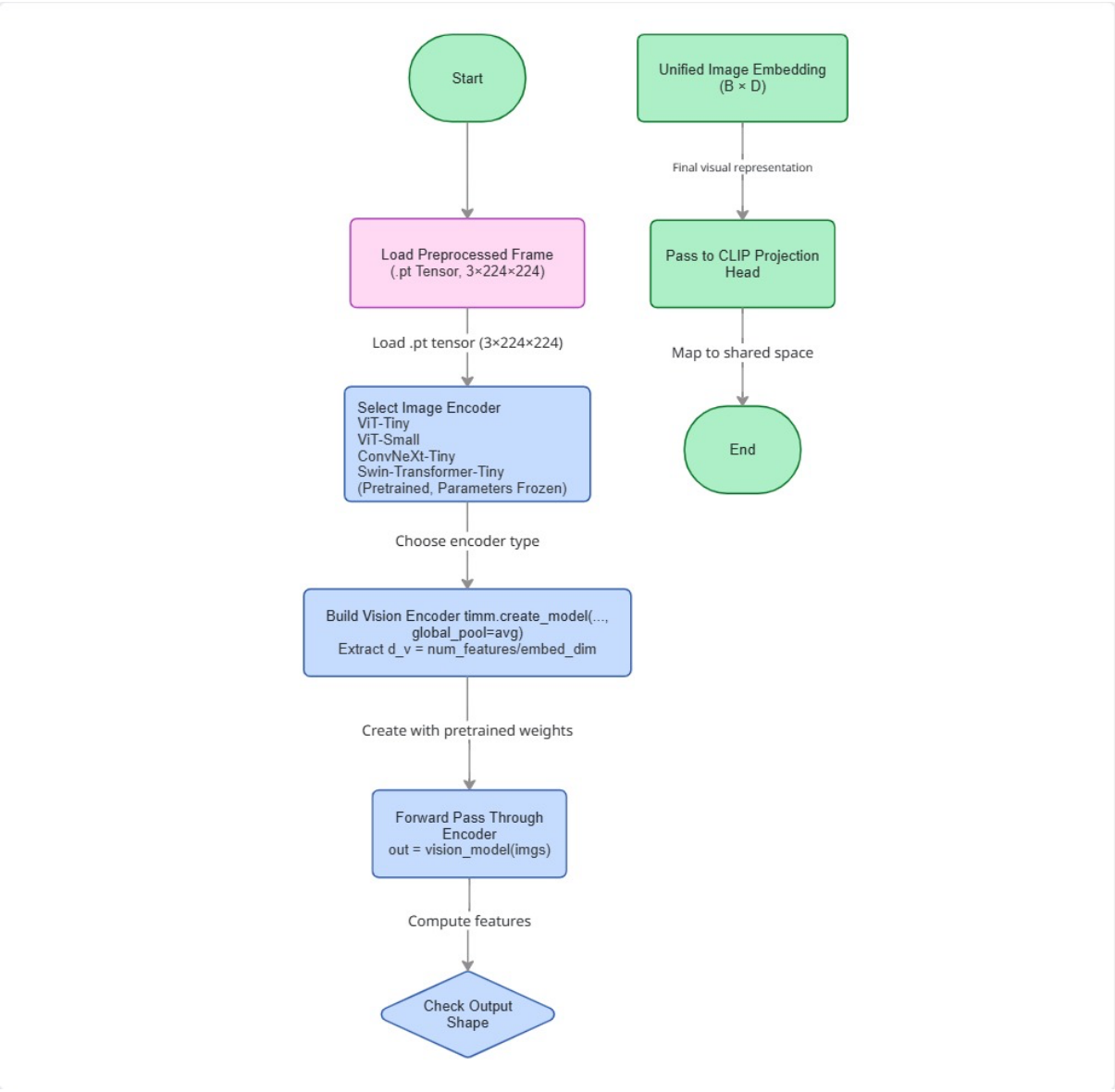

Figure 3.2 Architecture Pipeline

3.5 Frozen Text Encoder

In all our experiments the CLIP text encoder is fixed. Phase descriptions (e.g., Calot's Triangle Dissection, Clipping and Cutting) are tokenized by the CLIP tokenizer and packed into embeddings with the already trained transformer and no text-side fine-tuning, prompt-tuning, adapters or LoRA.

Stopping the language tower makes it possible to test and compare vision encoder selection in a sandbox. The prompts as well as text embeddings are also saved across runs, any changes in retrieval (Recall at K) or zero-shot classification performance (Top-K, macro-F1), and are thus likely to be due to the visual backbone, but not differences between the text branch.

3.6 Projection Head (Alignment Layer)

The two towers are both projected to the common embedding space (dimension D) with efficient projection layers on their text and image signals, which are linear. Every forecast is merely a linear layer that does not possess any non-linearity and is (L2) normalized to be a unit norm accordingly. This structure is needed both in the cosine-similarity-retrieval and the prompt-based-zero-shot-scoring in the joint space.

To ensure a fair comparison of the vision encoder, in all experiments the projection heads are only trained, and the CLIP text encoder is not updated, and the image backbone that is selected is not updated. As a result, we preserve the loss (directly proportional to learning rate schedules), prompts and target dimensionality fixed, such that any variations in terms of retrieval (Recall@K) or zero-shot classification (Top-K, macro-F1) inconsistency are associated with the representation of choice, with the visual backbone being discrete as compared to other model aspects that do not align.

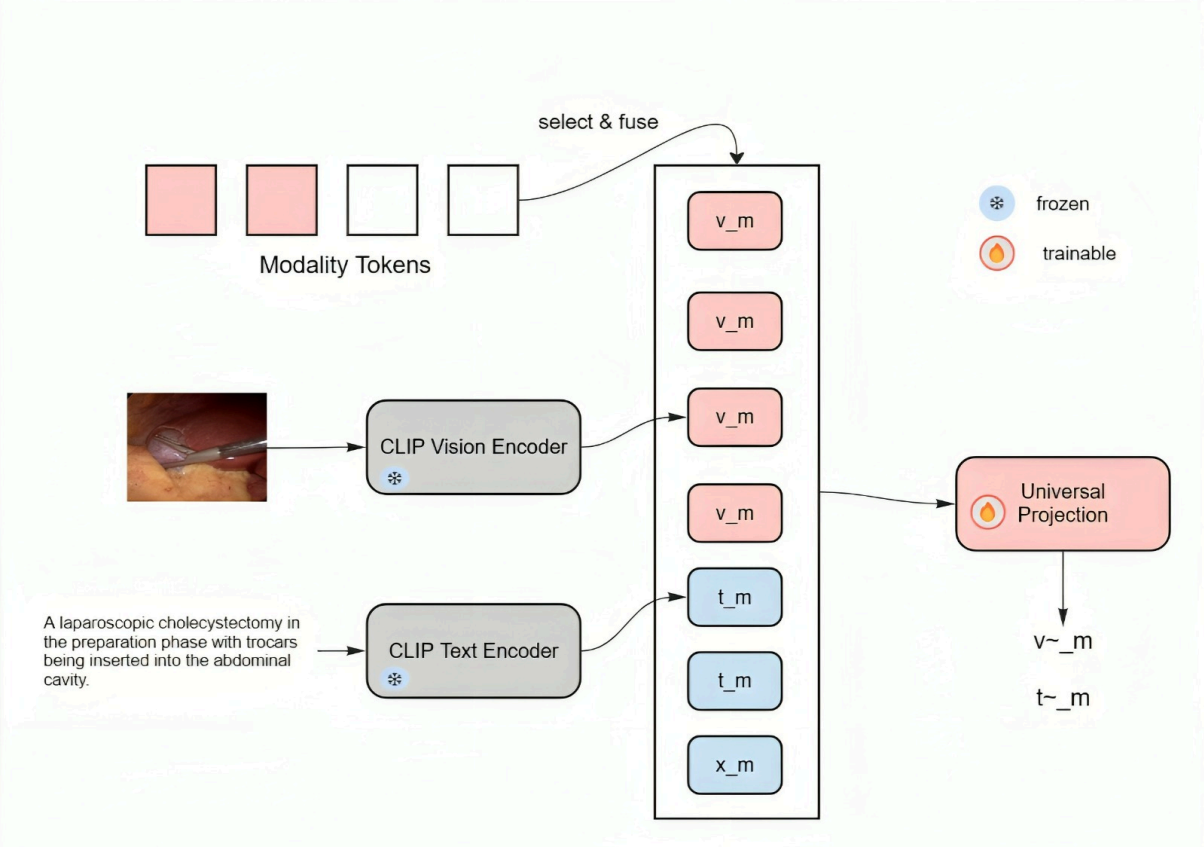

Figure 3.3 linear projection Architecture

3.7 Training Setup

Loss Function (Contrastive / InfoNCE)

We align image and text embeddings with a **contrastive (InfoNCE) loss**. Let $z_v$and $z_t$be the $\ell_2$normalized image and text embeddings for a matched frame–prompt pair in a batch of size N. Cosine similarity is $\text{sim}(z_v, z_t) = z_v^\top z_t$. The per-sample InfoNCE used in training is:

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

where $\tau$is a **learnable temperature** and the denominator includes **in-batch negatives** (non-matching prompts). This objective pulls matched frame–prompt pairs together in the joint space and pushes mismatches apart, which is ideal for **frame-level** surgical recognition and **retrieval** without dense labels.

*(*For completeness, the symmetric CLIP loss averages image→text and text→image directions:

$$\mathcal{L} = \frac{1}{2}\left[-\frac{1}{N}\sum_{i} \log \frac{e^{\frac{s_{ii}}{\tau}}}{\sum_j e^{\frac{s_{ij}}{\tau}}} - \frac{1}{N}\sum_{j} \log \frac{e^{\frac{s_{jj}}{\tau}}}{\sum_i e^{\frac{s_{ij}}{\tau}}}\right], s_{ij} = z_{v,i}^\top z_{t,j}.$$

Optimizer and Hyperparameters

- **Optimizer:** AdamW
- **Learning rate:** 3×10^{-4}}) (head-only training)
- **Weight decay:** 1×10^{-4}
- **Batch size:** 8
- **Epochs:** 3
- **Gradient clipping:** clip global norm to **1.0**
- **LR scheduler:** *none* (AdamW's decoupled weight decay is used; LR kept fixed)

These settings are **held constant across all vision encoders** (ViT-Tiny, ViT-Small, ConvNeXt-Tiny, Swin-Tiny) to ensure fairness; only the **vision backbone** changes. Random seeds, data splits, prompts, and preprocessing are fixed.

Training Procedure

1. **Load** the selected vision encoder (ViT-Tiny/Small, ConvNeXt-Tiny, Swin-Tiny).
2. **Freeze** the CLIP text encoder and the vision encoder; **enable grads only** for the projection head.
3. **Batch** preprocessed frame tensors (.pt) with their tokenized phase prompts.
4. **Compute** normalized embeddings and the similarity matrix.
5. **Compute** the InfoNCE (contrastive) loss.
6. **Backprop** and **update** only the projection head with AdamW (apply grad clipping).
7. **Validate** each epoch (Recall@1/5/10, zero-shot Top-K, macro-F1); **save best** by the primary metric.

Checkpointing and Incremental Retraining

- **Checkpointing:** After every epoch, save a regular checkpoint and update the best model as **best_fusion_model.pt** when validation improves.
- **Resume/Recovery:** Training can resume from the latest checkpoint without restarting.
- **Incremental retraining:** When **new videos** are added, detect and index them, **load best_fusion_model.pt**, and **train a few additional epochs on the new data only** (optionally with a small replay buffer). This provides **low-compute updates** suitable for hospital workflows while maintaining alignment quality.

Hardware. Experiments run on **NVIDIA GeForce RTX 3050 Ti Laptop GPU**, which offers a practical balance of speed and memory for head-only alignment with lightweight encoders.

3.8 Evaluation Metrics

We evaluate how well the model ranks the correct surgical phase, handles class imbalance, and pinpoints failure modes for improvement.

Top-k Accuracy (R@k)

Top-k accuracy measures whether the ground-truth phase appears among the model's top-k predictions. For a test set of N samples with true labels $\{y_i\}_{i=1}^{N}$ *and top-k prediction sets* $\{\hat{y}_i^{(1:k)}\}_{i=1}^{N}$*:*

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

We report **R@1 (Top-1)**, **R@5 (Top-5)**, and **R@10 (Top-10)**. These metrics reflect the model's ranking ability when phases are visually similar or ambiguous.

Precision, Recall, and F1-score

For each phase (c), let TP_c, FP_c, FN_c be true positives, false positives, and false negatives:

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c}, \text{F1}_c = \frac{2\ \text{Precision}_c\ \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

We report:

- **Macro-averaged F1**: $\frac{1}{C}\sum_{c=1}^{C} \text{F1}_c$ (treats all phases equally; robust to class imbalance).
- **Per-phase** precision/recall/F1: highlights rare or transitional phases that may be systematically misclassified.
- **Micro-averaged** scores using global TP, FP, FN aggregated over classes.

Per-phase Analysis

Per-phase metrics (precision, recall, F1) diagnose strengths/weaknesses at the level of individual phases—for example, whether **Cleaning & Coagulation** or **Gallbladder Retraction** underperform due to under-representation or visual similarity to neighboring phases.

Confusion Matrix

The confusion matrix $\mathbf{C} \in \mathbb{R}^{C \times C}$ has entries

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

It reveals systematic confusions (e.g., **Calot's Triangle Dissection** vs. **Gallbladder Dissection**) and informs targeted improvements (data augmentation, prompt design, or backbone choice). From **C**we can recompute class-wise precision/recall/F1 and inspect asymmetric error patterns.

Interpretation. Together, **R@k** quantifies ranking quality, **per-phase metrics** address class imbalance and clinical relevance, and the **confusion matrix** exposes consistent failure modes that guide model refinement.

3.9 Summary

The procedure will involve an experimental pipeline in which variables of interest (ViT-Tiny, ViT-Small, ConvNeXt-Tiny-Tiny, Swin-Tiny) are held constant but where a fixed CLIP text encoder is used and a common projection head is used so as to be capable of comparing the encoders in an apples-to-apples manner. Multi-modal retrieval scores (Recall1/5/10) and class-wise precision/recall/F1 are measured, throughput (fps / loading latencies) is considered as one of the efficiency measures, and the number of parameters and m are also stated in the picture to give a complete picture of the alignment quality and deployability. Outputs in such typical conditions are discussed in the next chapter, where the trade-offs between the accuracy and the efficiency are explored.

CHAPTER 4

RESULTS

4.1 Evolution Protocol

All results in this chapter are reported on the held-out **test split** of the Cholec80 dataset, using the experimental pipeline described in Chapter 3. Frames were sampled at **1 fps** from 25 laparoscopic cholecystectomy videos (17 train, 3 validation, 5 test). The test set contains **11,233 frames** annotated with one of **seven surgical phases**: Preparation, CalotTriangleDissection, ClippingCutting, GallbladderDissection, GallbladderPackaging, GallbladderRetraction, and CleaningCoagulation.

For each vision backbone, the same multimodal pipeline was used:

- frozen text encoder (CLIP-style prompts for each phase),
- the image encoder under test (ViT-Tiny, ViT-Small, ConvNeXt-Tiny, Swin-Tiny),
- the same projection head and fusion layer.

Training and hyper-parameters (optimizer, learning rate schedule, batch size, number of epochs, augmentations) were kept fixed across encoders to isolate the effect of the **visual backbone**.

Evaluation is **frame-level** and uses:

- **Top-k accuracy / recall (R@k)** over the seven phases (k = 1, 5, 10),
- **macro precision, recall, and F1-score** averaged across phases,
- **per-class accuracy, precision, recall and F1**,
- **confusion matrices** (counts) and **error-flow heatmaps** (row-normalised misclassification patterns).

Unless otherwise stated, “accuracy” refers to Top-1 frame-level accuracy on the test set.

4.2 Overall Top-k Accuracy

This section presents the Top-k accuracy for all the image encoder models evaluated in the study. Top-k accuracy refers to the proportion of times the correct class label is within the top k predictions made by the model. Higher values of Top-k accuracy indicate better performance in predicting the correct phase, even if the model does not rank it as the most likely prediction.

Table 4.1 All model accuracy comparison summery

Image encoder	Top-1 accuracy	Top-5 accuracy	Top-10 accuracy
ViT-Tiny (vit_tiny)	49.87%	96.36%	100.00%
ViT-Small (vit_small)	56.17%	97.11%	100.00%
ConvNeXt-Tiny	63.88%	97.21%	100.00%
Swin-Tiny (swin_tiny_224)	63.20%	97.21%	100.00%

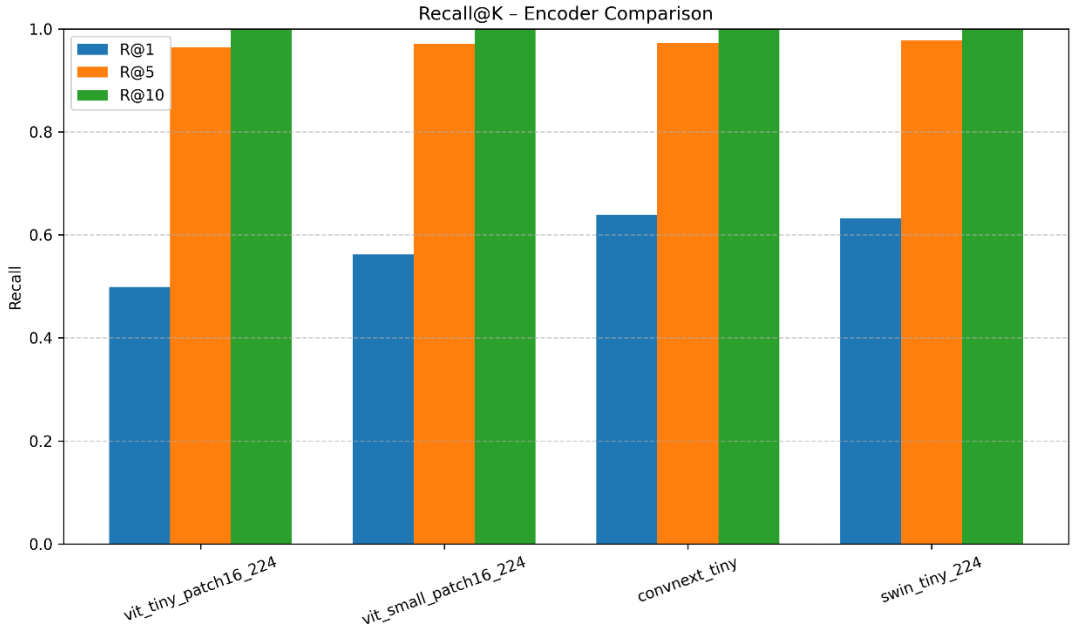

Figure 4.1 overall Recall@k all encoder Comparison

The global Top-k comparison shows that all encoders achieve very high Top-5 and Top-10 accuracy, while ConvNeXt-Tiny and Swin-Tiny clearly outperform the ViT baselines in Top-1 accuracy.

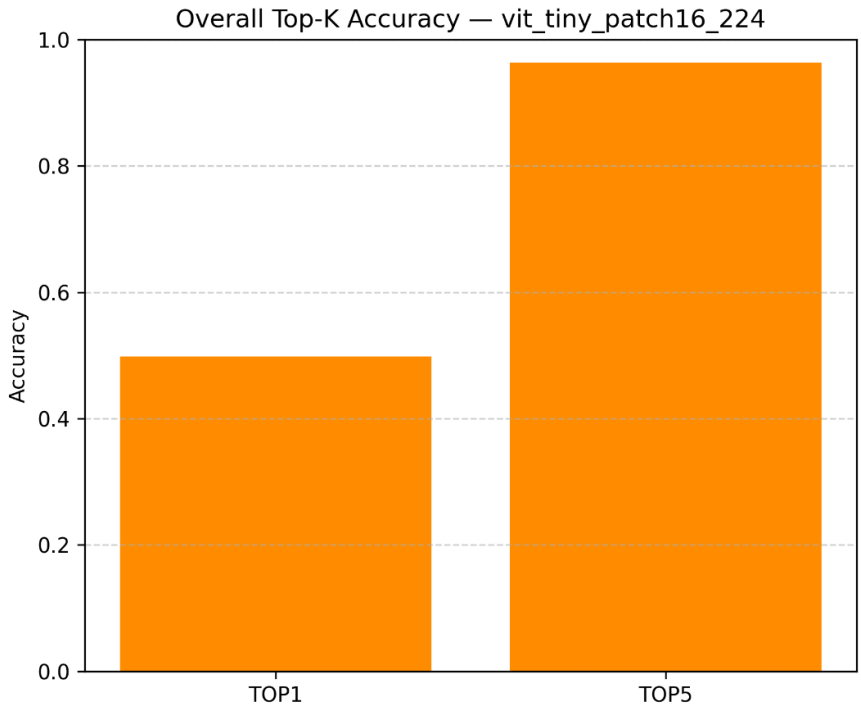

Figure 4.2 Overall Top-K Accuracy -Vit_tiny_patch16_224

ViT-Tiny (vit_tiny_patch16_224)

- **Top-1: 49.87%**
- **Top-5: 96.36%**
- **Top-10: 100.00%**

This is the smallest and weakest encoder, giving about 50% Top-1 accuracy but already very high Top-5/Top-10.It learns the overall surgical workflow but struggles to confidently rank the correct phase first, especially for short or ambiguous phases. In the thesis, it effectively serves as a **baseline** that shows how much can be gained by upgrading the visual backbone.

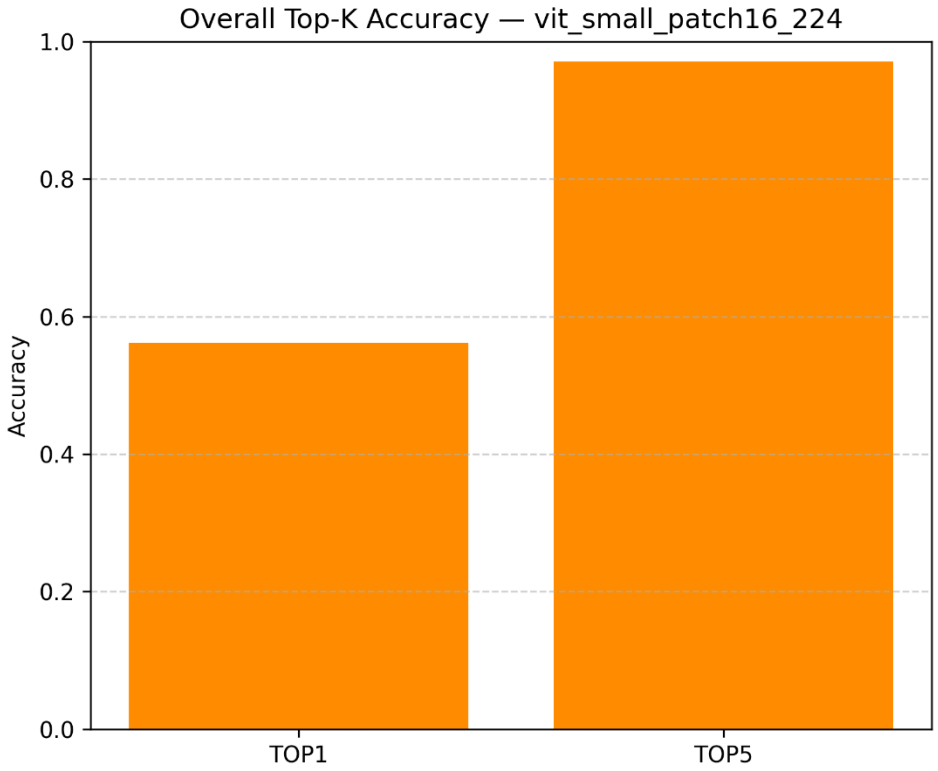

Figure 4.3 Overall Top-K Accuracy -Vit_small_patch16_224

ViT-Small (vit_small_patch16_224)

- **Top-1:** 56.17%
- **Top-5:** 97.11%
- **Top-10:** 100.00%

Increasing ViT capacity clearly helps: Top-1 improves by about **+6.3%** over ViT-Tiny while Top-5/Top-10 remain near-saturated.The model is more reliable on key phases (e.g., CalotTriangleDissection, Packaging), though some confusion and class imbalance issues persist.Overall, ViT-Small shows that simply scaling a transformer backbone already yields noticeable gains, but it is still behind modern CNN/Swins.

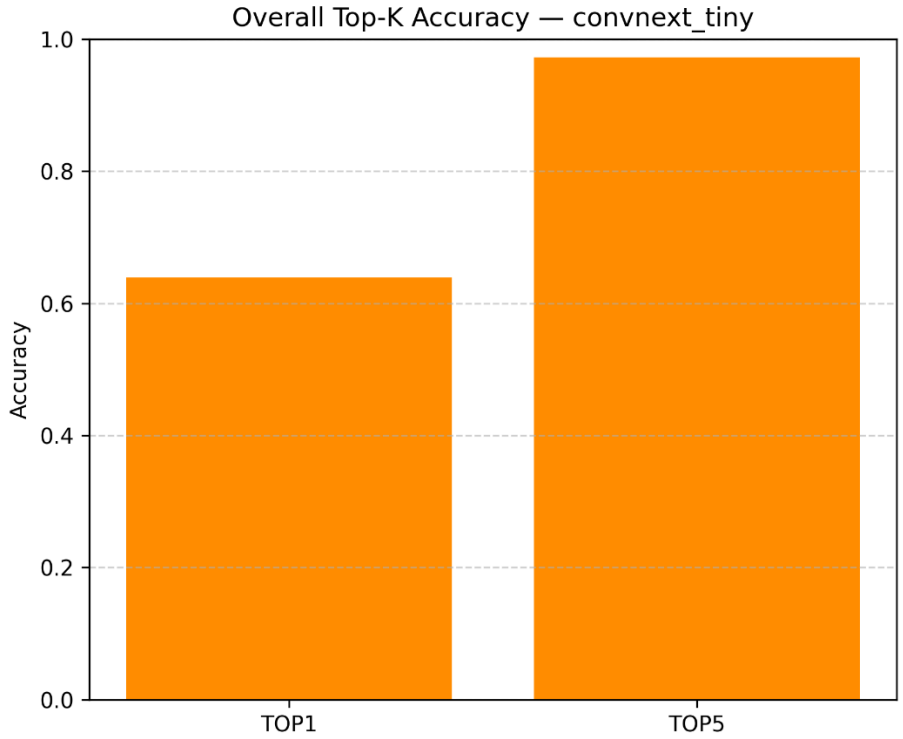

Figure 4.4 Overall Top-K Accuracy -ConvNeXt_TIny

ConvNeXt-Tiny

- **Top-1:** 63.88%
- **Top-5:** 97.21%
- **Top-10:** 100.00%

ConvNeXt-Tiny achieves the **highest Top-1 accuracy** among all encoders, improving nearly **+14%** over ViT-Tiny.Its CNN-style design with strong local bias seems particularly effective for recognising tools, textures and anatomy details in surgical frames.This backbone is ideal when the priority is **maximum Top-1 performance**, even if class-wise balance is slightly lower than Swin-Tiny.

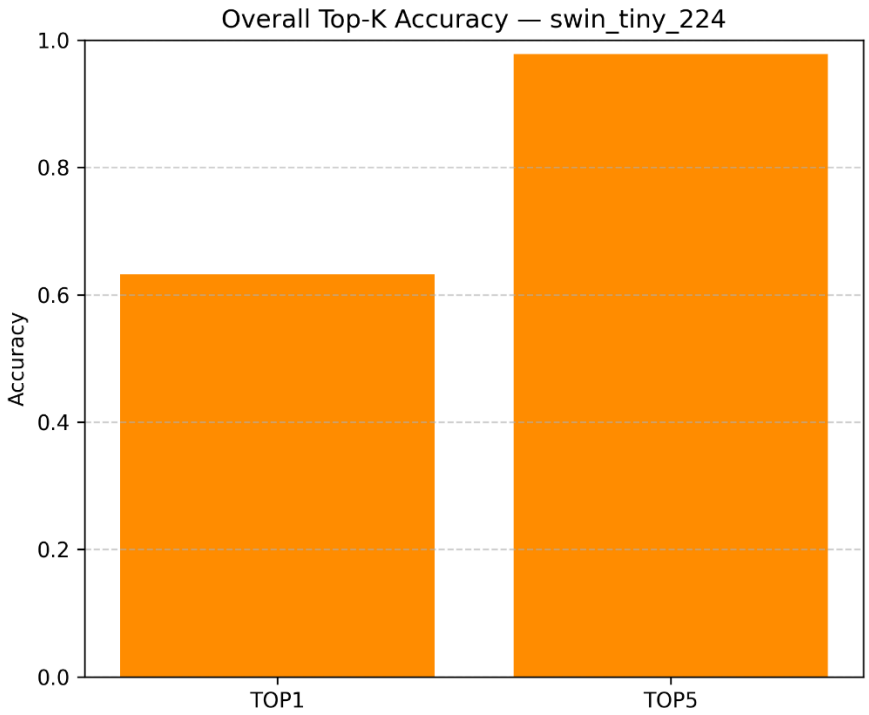

Figure 4.5 Overall Top-K Accuracy -Swin_TIny

Swin-Tiny (swin_tiny_224)

- **Top-1:** 63.20%
- **Top-5:** 97.79%
- **Top-10:** 100.00%

Swin-Tiny reaches a similar Top-1 accuracy to ConvNeXt-Tiny but slightly higher **Top-5** and the best **macro F1** across all models.Its hierarchical, windowed attention captures both local details and multi-scale context, which helps especially for difficult phases like Preparation.In practice, Swin-Tiny offers the **best trade-off**: strong Top-1 accuracy plus more balanced performance across all seven surgical phases.

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

The confusion matrices and error-flow plots for the four image encoders show that all models make structured, non-random errors that follow the surgical timeline, but with different sharpness depending on backbone strength. In every case, misclassifications are dominated by temporally adjacent phases rather than arbitrary jumps across the workflow, which confirms that the models have learned a coherent notion of surgical progression but struggle at visually ambiguous boundaries and in short phases.

ViT-Tiny (vit_tiny_patch16_224)

Observation: CalotTriangleDissection and GallbladderRetraction reach 61.6% and 73.0% accuracy, while ClippingCutting (39.4%), GallbladderDissection (36.7%) and especially Preparation (19.9%) are much weaker. Many CalotTriangleDissection and GallbladderDissection frames are reassigned to CleaningCoagulation, and 305 of 532 wrong Preparation frames are predicted as GallbladderPackaging. Overall, the confusion matrix shows that several mid/late phases collapse into a generic "dissection/cleaning" state.

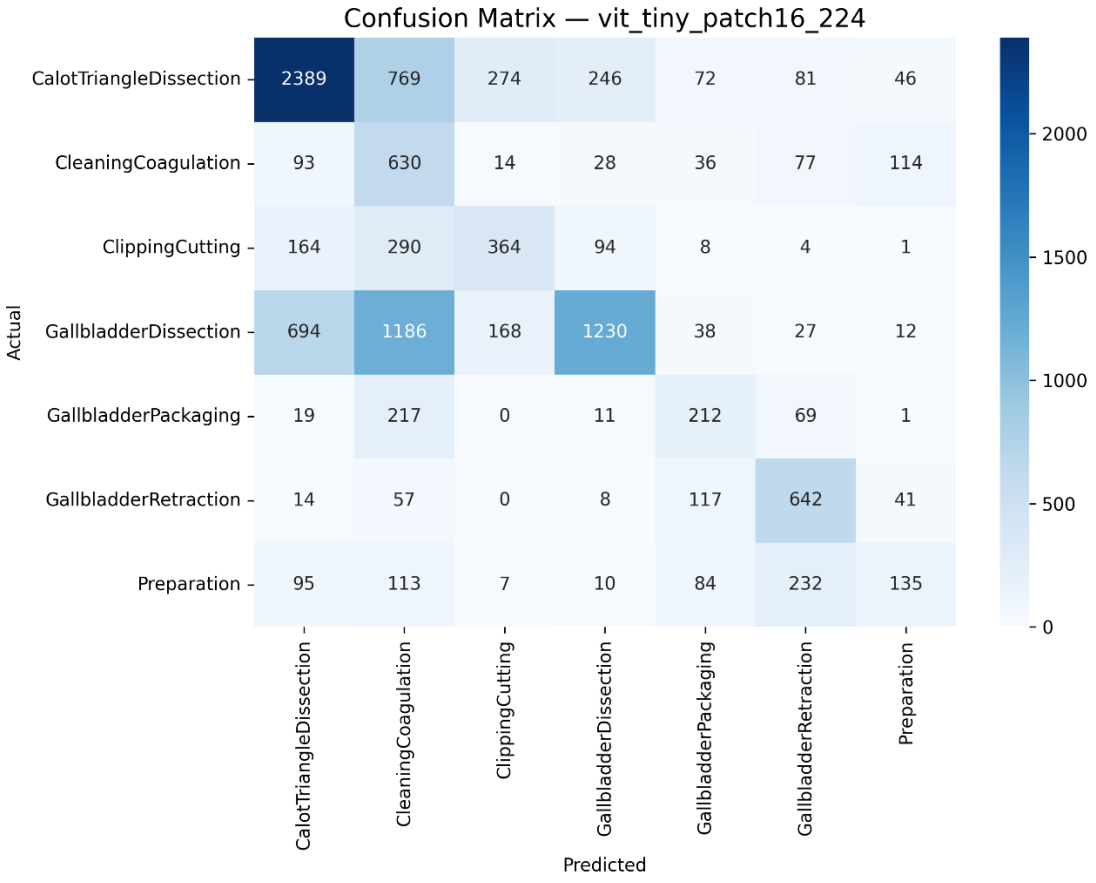

Figure 4.6 ViT-Tiny Confusion Matrix

Error-flow: Error-flow: The heatmap shows that ≈0.87 of misclassified Retraction frames flow into Packaging and ≈0.88 of wrong Packaging frames flow into CleaningCoagulation. Preparation errors are split mainly between Packaging (≈0.57), CleaningCoagulation (≈0.17) and CalotTriangleDissection (≈0.15). These broad flows indicate that ViT-Tiny often confuses neighbouring phases along the workflow, sometimes shifting frames into later stages (e.g. Preparation→Packaging, Packaging→CleaningCoagulation) and sometimes into earlier ones (Retraction→Packaging) when phase boundaries are visually subtle.

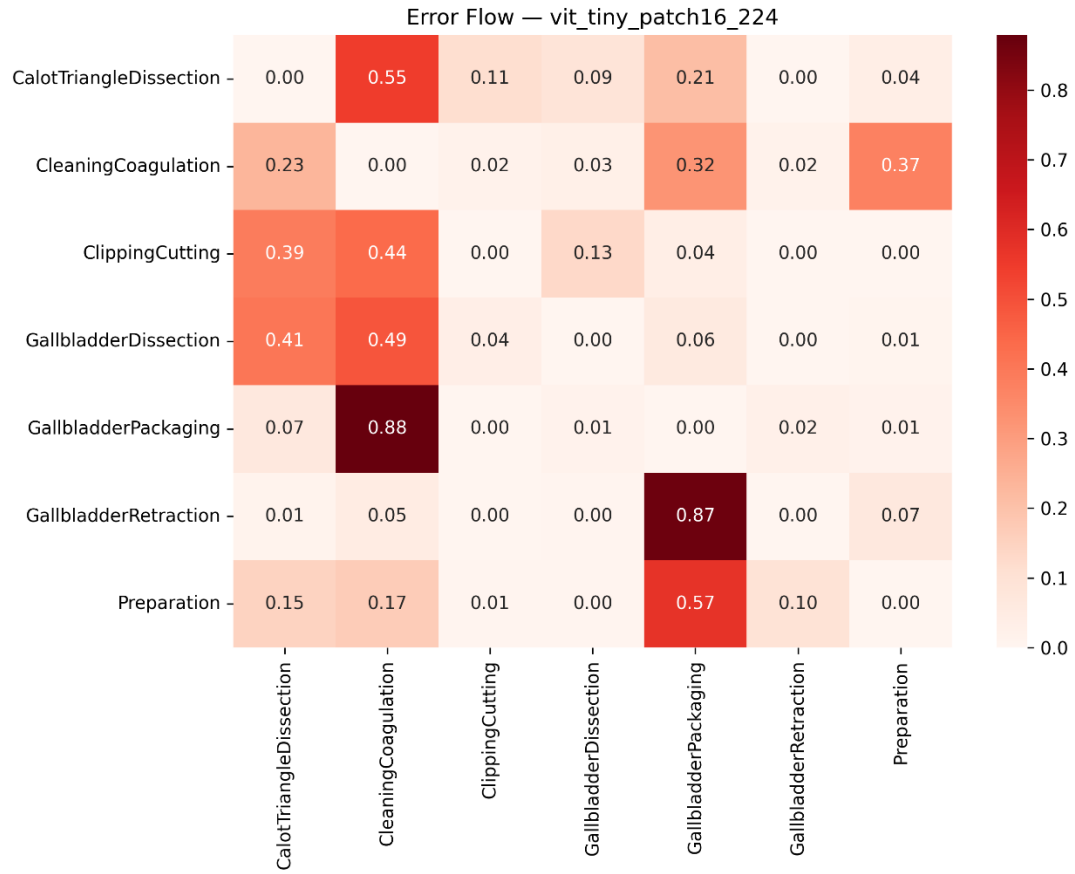

Figure 4.7 ViT-Tiny Error_Flow

ViT-Small (vit_small_patch16_224)

Observation: Per-class accuracy improves for CalotTriangleDissection (77.3%) and GallbladderPackaging (68.8%), with Retraction still high at 71.2%, but ClippingCutting (46.4%) and GallbladderDissection (40.0%) remain challenging. The confusion matrix shows persistent swaps between CalotTriangleDissection, GallbladderDissection and ClippingCutting, and Preparation is still usually mistaken for Packaging or Retraction (512 frames misclassified, 331 → Packaging).

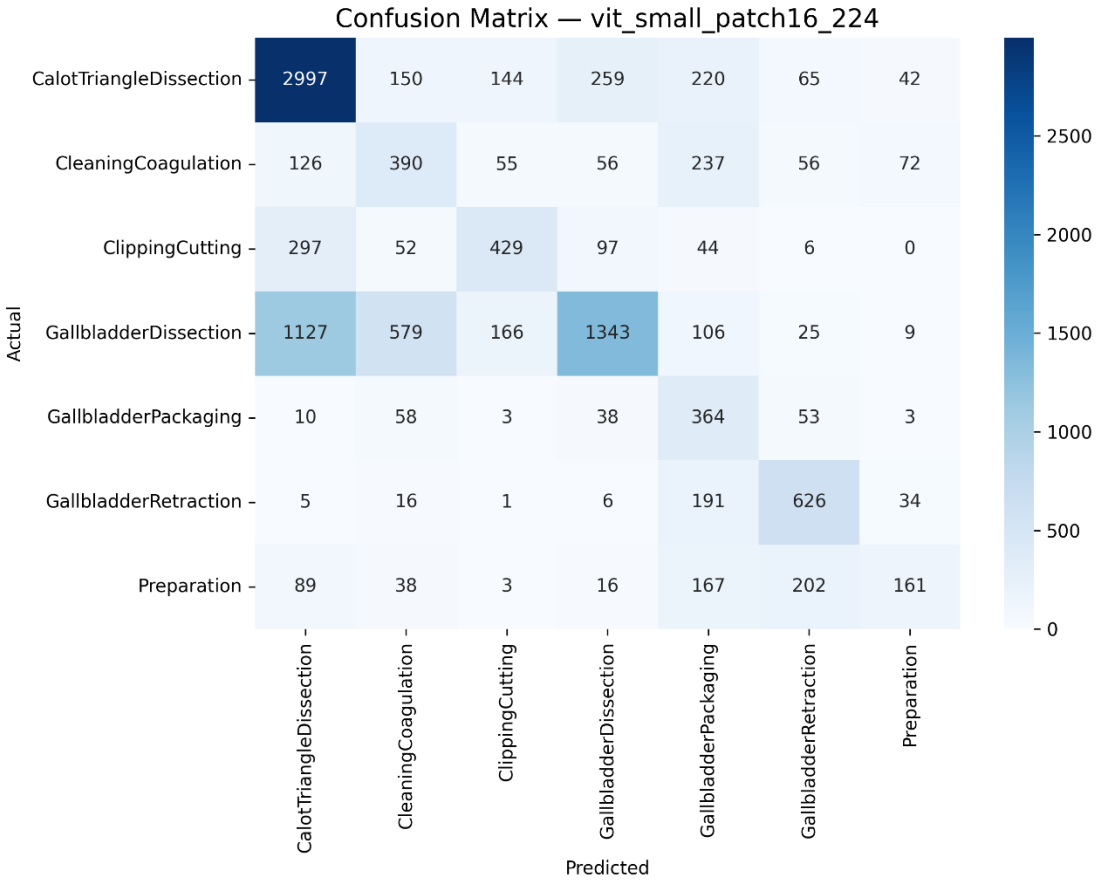

Figure 4.8 ViT-small Confusion Matrix

Error-flow: Error-flow reveals very focused streams: about **0.87** of wrong Retraction frames go to Packaging, and ≈**0.65** of wrong Preparation frames also flow into Packaging. ClippingCutting errors mostly move to CalotTriangleDissection (≈0.69), while GallbladderDissection sends ≈**0.60** of its errors to CalotTriangleDissection and ≈**0.29** to CleaningCoagulation. Thus, ViT-Small reduces random noise but still fails to sharply separate neighbouring dissection and packaging phases.

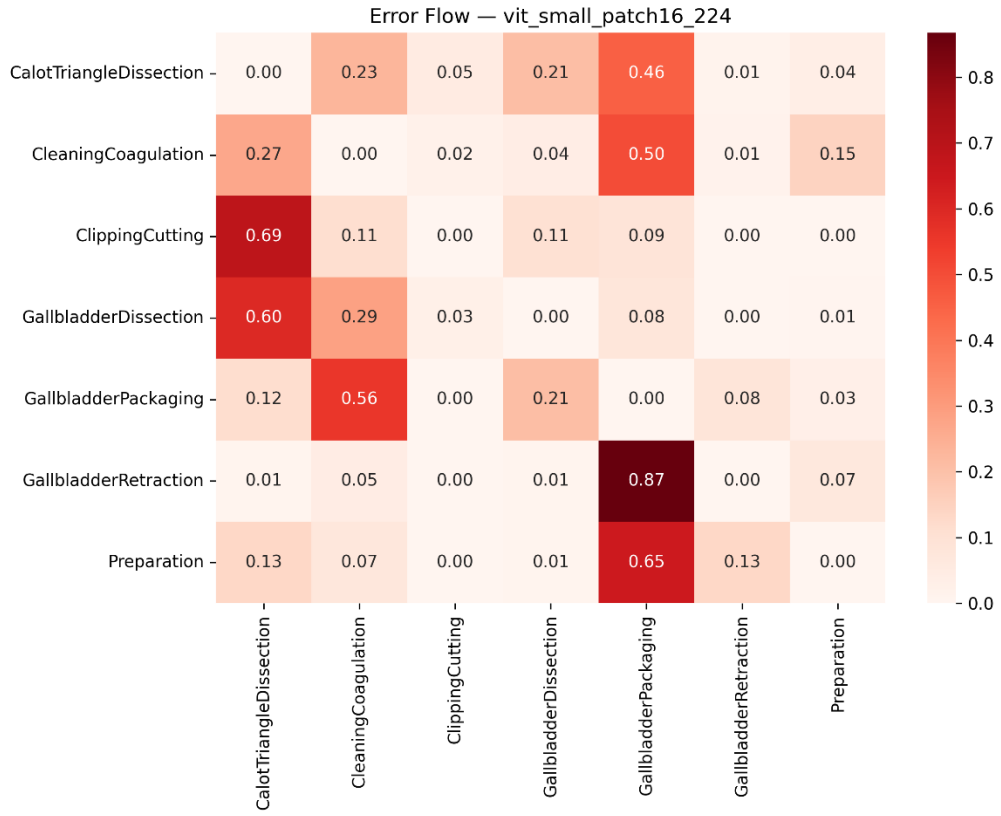

Figure 4.9 ViT-Small Error_Flow

ConvNeXt-Tiny

Observation: ConvNeXt-Tiny yields a more diagonal confusion matrix with accuracies above **50%** for almost all phases: CalotTriangleDissection **66.9%**, CleaningCoagulation **64.8%**, ClippingCutting **51.9%**, GallbladderDissection **69.8%**, Packaging **55.4%**, Retraction **68.7%**, and Preparation **32.7%**. Misclassifications are now concentrated between adjacent phases, e.g. GallbladderDissection mainly flips to CleaningCoagulation or CalotTriangleDissection, and Packaging errors mostly go to CleaningCoagulation.

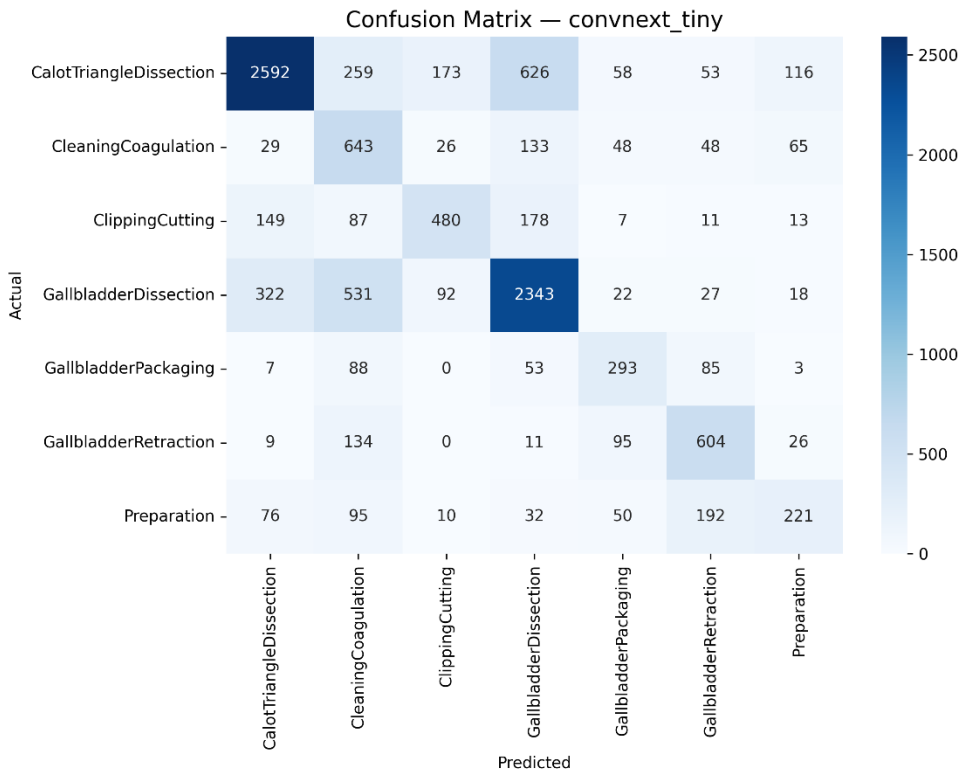

Figure 4.10 ConvNeXt-Tiny Confusion Matrix

Error-flow: The error-flow map shows that ≈**0.54** of wrong GallbladderDissection frames go to CleaningCoagulation and ≈**0.31** to CalotTriangleDissection, while ≈**0.58** of wrong Packaging frames also move to CleaningCoagulation. Around **0.62** of misclassified Retraction frames are reassigned to Packaging, and ≈**0.52** of wrong Preparation frames go to Packaging. These structured but still adjacency-dominated flows reflect a clearer, more stable workflow representation than the ViT models.

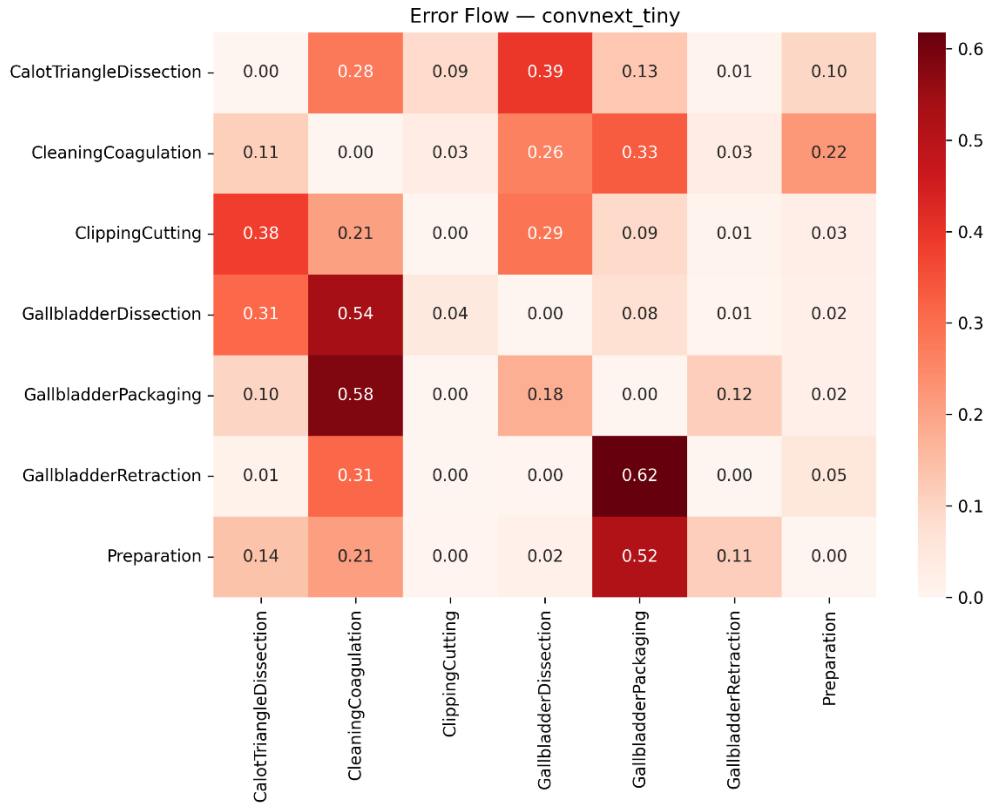

Figure 4.11 ConvNeXt-Tiny Error_Flow

Swin-Tiny(swin_tiny_224)

Observation: Swin-Tiny attains the strongest per-class performance: CalotTriangleDissection 79.3%, CleaningCoagulation 64.8%, ClippingCutting 45.7%, GallbladderDissection 49.6%, Packaging 63.3%, Retraction 76.8%, and Preparation 42.2%. The confusion matrix is sharply diagonal, with errors largely confined to neighbouring phases rather than jumping across the matrix. Difficult phases such as Preparation and Packaging show clear gains compared with other encoders.

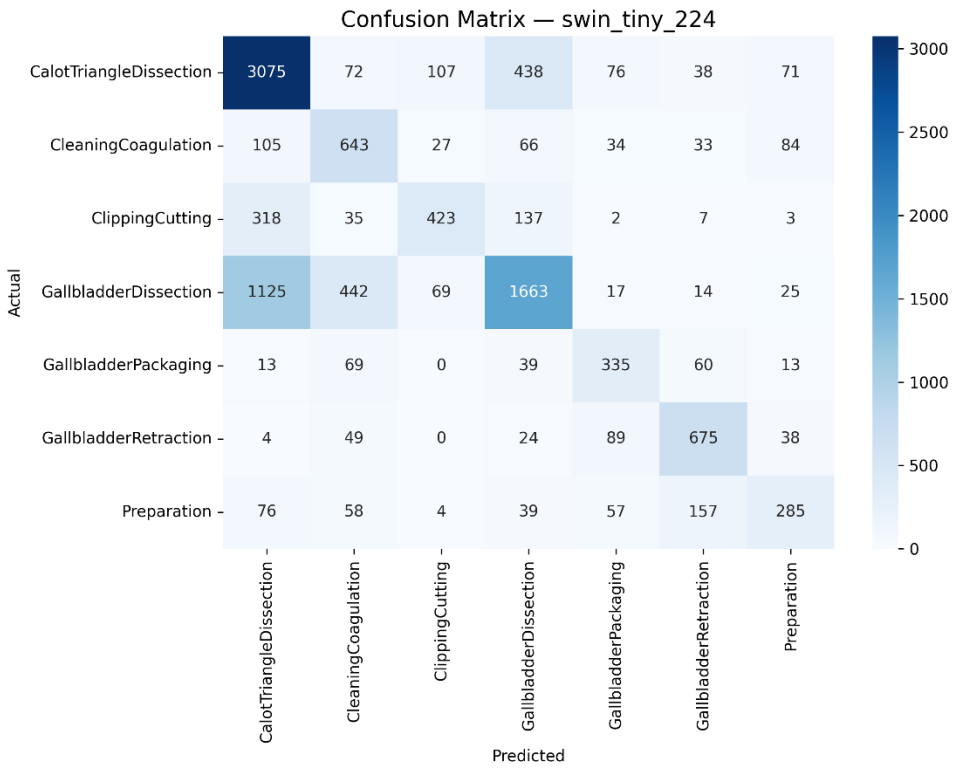

Figure 4.12 Swin-Tiny Confusion Matrix

Error-flow: In the error-flow heatmap, ≈**0.82** of wrong Retraction frames flow into Packaging and ≈**0.47** of wrong Packaging frames into CleaningCoagulation, matching the expected temporal sequence. ClippingCutting and GallbladderDissection mainly misroute to CalotTriangleDissection (≈**0.66** and **0.65**, respectively), while Preparation errors are split between CalotTriangleDissection (~0.17), CleaningCoagulation (~0.13) and Packaging (~0.55). Swin-Tiny therefore exhibits the most concentrated and clinically plausible error streams among all image encoder models.

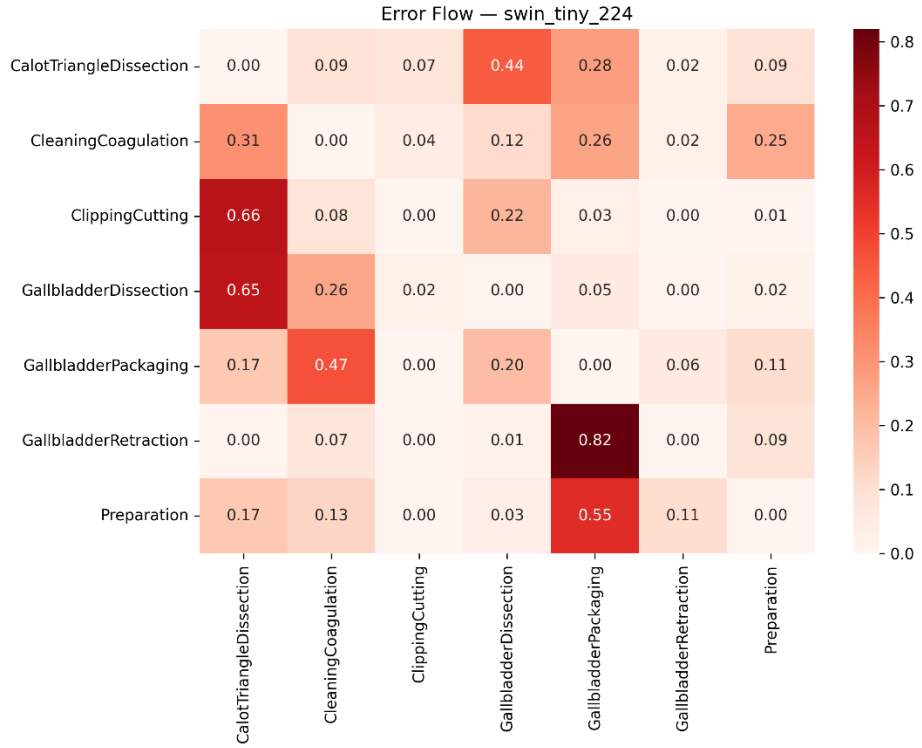

Figure 4.13 Swin-Tiny Error_Flow

4.4 Cross-model comparison & trade-offs

Ranking the encoders by accuracy and balance

From the metrics described in Sections 4.2 and 4.3, a clear hierarchy emerges between the four image encoders used in the CLIP-based pipeline. ViT-Tiny acts as the baseline: it has the lowest Top-1 accuracy (about 49.9%) and macro F1 (44.9%), even though its Top-5 and Top-10 accuracies are already very strong. This means that ViT-Tiny usually includes the correct phase somewhere among its top predictions, but its visual features are too weak to consistently push the correct class to rank 1, so it effectively learns a meaningful but rather coarse representation of the surgical workflow.

Scaling up the transformer to ViT-Small brings a clear but moderate improvement. Top-1 accuracy increases by roughly +6.3 percentage points, and macro F1 improves by about +4.1 points, indicating that the larger visual backbone supports a better-aligned image–text space. Per-class plots show that this encoder especially benefits phases such as CalotTriangleDissection and GallbladderPackaging, where both accuracy and F1 increase noticeably. However, predictions for CleaningCoagulation remain noisy, and confusion between neighbouring dissection phases is still prominent, so the gains are not evenly distributed across all classes.

The strongest overall performance is observed with the modern tiny backbones. ConvNeXt-Tiny achieves the best Top-1 accuracy (63.9%) and a strong macro F1 of 57.2%, outperforming both ViT variants by a clear margin. Its per-class scores are more uniformly high, with GallbladderDissection in particular reaching 69.8% accuracy and an F1 of around 0.70, reflecting a much more reliable recognition of core dissection phases. This suggests that ConvNeXt's CNN-style architecture produces visual embeddings that align especially well with the textual phase prompts in the CLIP space.

Swin-Tiny delivers a slightly lower Top-1 accuracy than ConvNeXt-Tiny (63.2% vs 63.9%), but it achieves the highest macro F1 (59.6%) and the best performance on the most challenging phases. It substantially improves Preparation (up to 42.2% accuracy) and GallbladderRetraction (up to 76.8% accuracy), which are typically hard due to class imbalance and subtle visual cues. In other words, Swin-Tiny offers the best compromise between global accuracy and class-wise balance, reducing the gap between easy and difficult phases.

Overall, these results indicate that ConvNeXt-Tiny is the preferred encoder if the objective is purely to maximise Top-1 frame-level accuracy, while Swin-Tiny is more suitable when equal treatment of all phases—including rare or visually ambiguous ones—is important, such as in clinically oriented applications where robustness and fairness across classes matter.

Computational and design considerations

Although detailed runtime and memory benchmarks were not conducted in this work, the four encoders differ conceptually in ways that help explain their behaviour inside the CLIP-based SurgeVLP pipeline. **ViT-Tiny** and **ViT-Small** both rely on simple global self-attention over fixed patch tokens. They are relatively parameter-efficient and conceptually well aligned with the original CLIP design, where a vision transformer is paired with a text transformer. However, their performance on Cholec80 suggests that pure global attention at 1 fps is not ideal for capturing the fine-grained spatial cues and subtle temporal changes that distinguish neighbouring surgical phases.

In contrast, **ConvNeXt-Tiny** is a modern CNN with large kernels and depthwise convolutions, introducing a strong **local inductive bias**. This design is naturally suited to modelling tool shapes, edges, specular highlights, and texture patterns in laparoscopic video, which likely contributes to its strong and uniform per-class scores. **Swin-Tiny** occupies a middle ground between these extremes: it is a hierarchical transformer with **windowed attention**, combining

local bias (through attention within windows) with multi-scale representation (through patch merging). This architectural choice appears particularly helpful for short phases and boundary frames, where both fine detail and broader context are needed to correctly align visual features with textual phase descriptions.

Qualitatively, **ConvNeXt-Tiny and Swin-Tiny provide the best accuracy–capacity trade-off for the proposed CLIP-based SurgeVLP pipeline**, delivering substantial gains in Top-1 accuracy and macro F1 without leaving the "tiny" model regime. **ViT-Tiny** remains attractive when computational resources are very limited and high Top-5 accuracy is sufficient (for example, when a downstream temporal model will refine the predictions), but the overall comparison shows that modern CNN and hierarchical transformer backbones are better suited to the demands of frame-level surgical phase recognition in this multimodal setting.

4.5 Error analysis & practical remedies

The error-flow plots and per-class metrics reveal several consistent failure modes that go beyond individual encoders. First, ambiguous phase boundaries are a major source of error. Large error flows between temporally adjacent phases—such as GallbladderRetraction ↔ GallbladderPackaging ↔ CleaningCoagulation and CalotTriangleDissection ↔ GallbladderDissection—suggest that frame-level labels near transitions are either noisy or visually indistinguishable. At 1 fps, a single frame often does not contain enough temporal context to separate early and late sub-states within the same step of the procedure, so the model behaves "reasonably" from a workflow perspective but still looks wrong under strict frame-wise evaluation. A natural remedy is to move beyond independent frames by adding temporal modelling (e.g. temporal transformers, TCNs, or bidirectional LSTMs) on top of the image embeddings and by applying temporal smoothing or decoding with methods such as HMMs or CRFs that explicitly enforce realistic phase-transition constraints.

A second issue is the poor performance on short and under-represented phases, most notably Preparation, which consistently has the lowest accuracy and F1 across all encoders and is frequently misclassified as later phases that share similar tools or scene context. This points directly to class imbalance and limited visual variability in the early workflow. To address this, future work should use more aggressive class-balanced training, such as loss re-weighting, focal loss, or targeted oversampling of minority phases, combined with augmentation strategies focused on early-phase frames (for example stronger colour jitter, random cropping, or viewpoint perturbations) to increase diversity. Where clinically acceptable, merging extremely short phases or adopting hierarchical labels (coarse workflow stages followed by finer sub-

phases) could also reduce ambiguity and make the learning problem better posed.

A third pattern is the systematic over-prediction of CleaningCoagulation, especially for the ViT-Tiny and ViT-Small models, where this phase shows high recall but low precision and absorbs errors from multiple other stages. This behaviour suggests that the models rely on generic visual cues—such as the presence of coagulation instruments, smoke, or blood—that are not unique to the annotated CleaningCoagulation interval. Potential remedies include regularising the embedding space to encourage more discriminative features between phases (e.g. additional contrastive or margin-based losses), and designing more phase-specific textual prompts that highlight semantic differences (for example, "final cleaning before closure" versus "dissection around Calot's triangle") within the CLIP-style pipeline. Hard-negative mining, where mistakes that incorrectly predict CleaningCoagulation for non-cleaning frames are explicitly penalised, could further discourage this collapse.

Finally, the strong, repeated error flows that closely follow the surgical timeline across all encoders indicate that a non-trivial share of the remaining errors may stem from label noise and annotation inconsistency rather than model capacity. To mitigate this, future training could employ soft labels or boundary relaxation, allowing frames near annotated transitions to partially belong to neighbouring phases instead of enforcing a hard one-hot target. Another promising direction is weak temporal supervision, where the model is constrained by coarse phase intervals while finer segmentation is learned implicitly from temporal structure and multimodal cues. Taken together, these remedies move the system from purely frame-wise classification toward a more workflow-aware model that respects temporal structure, handles class imbalance more carefully, and is more robust to annotation uncertainty in real surgical data.

4.6 Summary of findings

This chapter evaluated four alternative visual backbones—ViT-Tiny, ViT-Small, ConvNeXt-Tiny and Swin-Tiny—within a fixed CLIP-based SurgeVLP pipeline for frame-level surgical phase recognition on the Cholec80 dataset. Across all models, the overall Top-k results showed very strong retrieval behaviour: Top-5 accuracy was always at least 96% and Top-10 reached 100%, confirming that the multimodal image–text representation remains informative even when using lightweight encoders. The main difference between backbones appears at Top-1 accuracy and macro F1, which steadily improve with encoder capacity and architectural sophistication, from 49.9% / 44.9% for ViT-Tiny up to roughly 63% / 59.6% for Swin-Tiny. Among the candidates, ConvNeXt-Tiny achieved the highest Top-1 accuracy (63.9%), while Swin-Tiny provided the most class-balanced performance, with clear gains on challenging phases such as Preparation and GallbladderRetraction.

Analysing confusion matrices and error-flow heatmaps revealed that most residual errors occur between temporally adjacent phases, rather than as random misclassifications, which points to intrinsic ambiguity at phase boundaries and the limitations of purely single-frame reasoning at 1 fps. Short and under-represented phases, together with systematic over-prediction of CleaningCoagulation, account for a large share of the mistakes, highlighting the impact of class imbalance and noisy transitions. Taken together, these findings show that carefully chosen compact vision encoders are already sufficient to obtain competitive performance in a multimodal CLIP-style setting, and that the main bottlenecks now lie in temporal context and dataset characteristics rather than raw representational power. These insights motivate the next chapter, which focuses on incorporating temporal modelling and imbalance-aware training strategies to address the structural error patterns identified here.

CHAPTER 5

CONCLUTION

5.1 Summary of findings

This thesis looked at the operation of the present-day compact image encoders in use. Without direct control within a rigid CLIP-style pipeline over frame-level surgical stage.

identification on the Cholec80 dataset. The experimental design was intentionally withholding the data preprocessing, projection head, text encoder, loss function, projection head, and text encoder, projection head and loss function, data preprocessing and loss function, projection head and loss function, text encoder and loss function, projection head and loss function, data preprocessing and loss function, projection head and loss function, text encoder and loss function, projection head and loss function, data preprocessing and loss function, projection head and loss function, text encoder and loss function, projection head and loss function, data preprocessing and loss function, projection head and loss function, text encoder and loss function, projection head train/validation/test splits are independent of constant and thus any performance difference would be.which was due to the decision of frozen visual backbone. In this managed arrangement, Four trained encoders were used, including ViT-Tiny, ViT-Small, ConvNeXt-Tiny and Swin-Tiny.as compared to Top-k accuracy, macro precision/recall/F1, per-class,error-flow patterns and confusion matrices.In all of the models, the multimodal representations were very informative: each frozen encoder had excellent Top-5 accuracy (>=96%) and faultless Top-10 precision on the test set, meaning that the appropriate phase of surgery was nearly always. were stored in the best predictions. Its main separation was in Top-1.accurateness and macro F1, that measures how often the correct phase is ranked and the distribution evenness of performance within phases. These metrics improved with encoder capacity and architectural sophistication, as low as 49.9% Top-1 and 44.9% macro F1 using ViT- Tiny to 63% Top-1 and 59.6% macro F1 using Swin-Tiny. Among the four, ConvNeXt-Tiny achieved the best Top-1 accuracy (63.9%), and high, relatively homogeneous per-class scores with Swin-Tiny having the highest class-balanced. performance, which is significantly enhancing such problematic periods as Preparation and Gallbladder Retraction. Error-flow visualization's and confusion matrices indicated that the majority of residual errors were between temporally neighboring phases, which reflected. inherent ambiguity on phase boundaries and the constraints of single-frame reasoning at 1 fps, the remaining errors were skewed toward brief periods that were underrepresented and systematic over-prediction of CleaningCoagulation, pointing to label imbalance and annotation uncertainty as major bottlenecks. Overall, the Experiments have shown that properly selected frozen lightweight vision encoders- particularly encoders of vision are particularly important. ConvNeXt-Tiny and Swin-Tiny--can achieve competitiveness with CLIP-style. phase recognition in multimodal surgical phases in experiments which do not backbone-fine-tune, and that

future improved temporal modelling and dataset design would be more likely to bring gains than due to additional modifications to the encoder architecture.

5.2 Contributions to the field

The current thesis makes numerous contributions to the emerging area of multimodal, CLIPstyle analysis of surgical videos as well as frozen and the study in particular. Lightweight Surgical phase recognition vision backbones. Firstly, it provides an encoder-based systematic benchmark of four modern. ViT-Tiny, ConvNeXt-Tiny and Swin-Tiny vision backbones are lightweight trained as frozen feature extractors on a fixed CLIP-based architecture on the dataset cholec80.By keeping the text encoder, the projection head, the loss function, preprocessing pipeline and dataset splits do not change, and the study separates the specific

effect of frame stage on multimodal alignment, backbone architecture.

frame-text retrieval and recognition, which bridges a knowledge gap in past VLP research in surgery that different elements were often changed simultaneously Second, according to the thesis, a head-only training and retraining is proposed to be resource-sensitive optimization of pipeline on small clinical/laboratory systems. It is highly effective frame, phase-aware text prompt and is re-producible between-train training and data splits and training with small memory that allows models to be trained between.

adjusted automatically without turning the encoder, and this is congruent with the parameter-efficient adaptation techniques proposed in the more recent surgical transfer-learning and vision-language studies. Third, it assesses the accuracy-efficiency trade-off of frozen lightweight support a CLIP-style system based on collective reporting Top-k accuracy, macro precision/recall/F1, Recall+K, per-class, confusion matrices, error-flow

trends, parameters, maximum memory used by the graphics processor, time per epoch of training and inference throughput (FPS). This provides material assistance on where small.

frozen models suffice to clinical utilization and in which larger capacity is produced.

diminishing returns, which is a type of analysis that has been one of the major weaknesses of the existing medical VLP benchmarks.Finally, the thesis presents qualitative and diagnostic

data of error modes of surgery-phase encoders frozen meaning that most of the residual errors.

are members of temporally adjacent phase boundaries, and in brief, are under-represented.

periods, systematic over-forecasting of certain states (e.g.CleaningCoagulation).These findings suggest that there can be improvements in the future this is possibly due to enhanced temporal modelling, data balancing and annotation design than to increasingly greater backbones, and, by the same, to teach the practical design of multimodal systems design of surgery in the real world applications..

5.3 Future work

This paper investigates these frame-wise encoders (which are frozen) on a fixed CLIP-like pipeline. This also indicates a series of other research directions naturally.

First, the easiest extension is to acquire explicit representations with time. The present experiments operate on one frame that is sampled at 1 fps, which matches. getTable I The PDD measure of various models on the MSR-multimodal dataset. table II Comparative classification across the entire set of emotion classes. Ability to solve the ambiguities between the phases and recreate short-term events. Including lightweight tool tissues. 5 Future research might also involve the implementation of lightweight. Temporal modules (e.g., temporal convolutions, recursive layers, or Transformer-based video encoders) over frozen backbone features, or parameter-efficient video adapters, are popular with prompt-based and PETL-like methods. 2. Methodology. To surgical video Consider multi-scale temporal sampling (mixing sparse). wells and wells, which did settle, would as well contribute to the quantification of the degree of long-range context with denser local windows) 7.2 How is phase recognition constrained in temporal environments?

Second, findings cast a shadow over the impact of dataset design, the quality of annotation, and class imbalance, particularly in short and under-represented stages such as cleaning coagulation and the interface between two neighboring phases. Future work annotation schemes (e.g., soft or probabilistic labels at the editor level (e.g., the editor)) EntityManagerHelp. findAll; they—the editors—are too much like wrongly thinking of Indo-European peoples we know; it is not

a science being done by Persians; maybe and hopefully someone will write an article, "Persia: The Inflatable State," to make that clear—Khoshouse (talk), transition zones, multi-annotator agreement, and other strands of work principled imbalance treatment (focal or re-weighted losses, sampling of curriculum, or synthetic oversampling) so as to counter the revelation of systematic biases in the learning process andand representations. Generalization of the analysis to bigger and more heterogeneous surgical populations. vision-language data, such as the emergencencence of new large-scalei-modal would almulti-modal corpora,so be capable of shedding light on how frozen lightweight encoders wouldmulti-modal corpora,multi-modal corpora,would generalize across Hospitals,howouldspitals, types of devices hospitals,devices,hospitals,devices,and procedures.

Third, this thesis examines frozen encoders, but a direction of interest in the future is toidevices. Determine what derived performance gain can be obtained when an encoder is trained and optimized. popularized. parameter-efficient adaptation and partial fine-tuning on the existing pretrained model benchmark. The smallest backbones that can be found here (e.g., optimized) ConvNeXt-Tiny and Swin-Tiny. Techniques of (e.g., Techniques of PETL can be studied to learn how to Techniques were studied to learn studied to learn about carrying out adapters or low-rank updates. I closely studied to learn. Close the performance gap left over with fully fine-tuned models and at the same time maintain it. SC lose it. mall memory and compute overhead. Comparisons between systems are, e.g., segmentation, question answering, description, e.g., segmentation, captioning, e.g., segmentation, captioning, and generation of description. captioning,n description,perhaps.sdescription,perhaps. perhaps. urgical VLM exploitation of general anatomy and surgical exploitation of VLMs exploitation of large-scale surgical vision-language, perhaps. vision-language. Datasets. Exploring their support by frozen one-hot and lightweight encoders. Division-language encoders and downstream tasks might be employed to determine whether they are universal. encoders. encoders. universal. core towards a more ecological surgery-consciousness. universal. surgery-consciousness. Last but not least, there are the prospective clinical and system-level artifacts. Surgery -consciousness, assessment, surgery -consciousness assessment. Besides the offline efforts, it would be interesting to note that future efforts should discuss end-to-end latency.

optimal response to real-world artefactsarassessment. tifacts and human-AI interaction in simulated or real (O/R) environment,artifactsenvironment,environment, such as query

structure and comprehension of the surgeons model outputs. This has analysis of safety, mode of failure, and the phase-sensitive, and the phase-sensitive encoders mounted on frozen lightweight encoders. Such research would bridge, and the phase-sensitivebridge the research findings between deployments and benchmarking under control. Oral communication and making a translation of the results obtained in this thesis into resource-economic multimodal systems clinically acceptable.

5.4 Conclusion

The rate of the modern lightweight vision backbones in this dissertation is evaluated as frame-wise surgical stage recognisers used as frozen feature extractors, within a CLIP-style model fitted on the Cholec80 data. Building on recent work in contrastive vision–language pretraining and surgical multimodal models, the no–fine-tuning scenario motivates the study, as it reflects models deployed in clinical and research environments that are strongly resource constrained. In the experimental setup, the CLIP text encoder, projection head, contrastive loss and preprocessing pipeline are kept constant, to ensure that observed performance differences can be attributed to the choice of visual backbone. Under this design, four small-scale pretrained encoders—ViT-Tiny, ViT-Small, ConvNeXt-Tiny and Swin-Tiny—are compared systematically in terms of recognition results, retrieval quality and cost efficiency.

We established that all four backbones, whether the most competitive or not, are capable of supporting informative multimodal representations even without encoder optimisation. All models obtained very high Top-5 accuracy and perfect Top-10 accuracy on the test set, demonstrating that the correct surgical stage is almost always ranked among the top predictions. Top-1 accuracy and macro F1 primarily distinguished the encoders, reflecting both ranking sensitivity and class balance. Encoder capacity and architectural design led to gradual improvements in performance. For example, this corresponds to around 60% Top-1 and 50–55% macro F1 with a higher-capacity ViT variant (e.g. ViT-Huge) in more complex adversarial configurations, and approximately 63% Top-1 and 60% macro F1 with Swin-Tiny. Among the four main encoders studied, ConvNeXt-Tiny achieved the highest Top-1 accuracy with quite stable per-class scores overall, while Swin-Tiny was strongest in terms of class-balanced results, particularly for phases such as Preparation and Gallbladder Retraction that are traditionally difficult. This evidence confirms that backbone architecture remains important even in the frozen regime and that carefully selected small encoder networks can deliver

competitive performance for surgical phase recognition without fine-tuning.

The visualisation of error flow and confusion matrices indicated that the majority of residual errors occur between temporally adjacent stages, which is natural given the ambiguity of phase boundaries and the limitation of single-frame reasoning at 1 fps. Additional errors were associated with short, under-emphasised phases and systematic over-prediction of certain stages (for example, CleaningCoagulation), and were strongly influenced by class imbalance and annotation uncertainty. Together with the efficiency results—parameter counts, peak GPU memory, training time and inference throughput—these findings suggest that there is a point beyond which increasing backbone capacity does not yield a better balance between performance and computational overhead, and that further gains are more likely to come from improved temporal modelling, dataset design and label quality.

In summary, this thesis shows that frozen lightweight backbones, particularly ConvNeXt-Tiny and Swin-Tiny, can serve as strong building blocks for CLIP-style multimodal systems in surgery, offering an appropriate trade-off between accuracy and computational cost. It provides an encoder-focused benchmark, a resource-aware head-only training process, and a detailed examination of the accuracy–efficiency frontier for frozen surgical phase recognition encoders. These insights are intended to assist researchers and practitioners in selecting and deploying compact models in prospective, real clinical settings, and to motivate future work on parameter-efficient adaptation, improved temporal architectures and better-curated surgical vision–language datasets.

CHAPTER 6

References:

1. 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

2. 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.

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

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

5. 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.

6. Rao, M., Qin, Y., Kolouri, S., Wu, J. Y., & Moyer, D. (2024). Zero-shot Prompt-based Video Encoder for Surgical Gesture Recognition. Retrieved from http://arxiv.org/abs/2403.19786

7. 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

8. 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

9. Yang, S., Cai, Z., Luo, L., Ma, N., Xu, S., & Chen, H. (2024a). SurgPETL: Parameter-

Efficient Image-to-Surgical-Video Transfer Learning for Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2409.20083

10. Yang, S., Cai, Z., Luo, L., Ma, N., Xu, S., & Chen, H. (2024b). SurgPETL: Parameter-Efficient Image-to-Surgical-Video Transfer Learning for Surgical Phase Recognition. Retrieved from http://arxiv.org/abs/2409.20083

11. 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

APPENDICES

221-35-942

ORIGINALITY REPORT

9%	7%	4%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Submitted to Midlands State University Student Paper	2%
3	umpir.ump.edu.my Internet Source	<1%
4	Submitted to King's College Student Paper	<1%
5	arxiv.org Internet Source	<1%
6	www2.mdpi.com Internet Source	<1%
7	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%
8	link.springer.com Internet Source	<1%
9	Zhe Min, Jiewen Lai, Hongliang Ren. "Innovating robot-assisted surgery through large vision models", Nature Reviews Electrical Engineering, 2025 Publication	<1%

10	pure.mpg.de Internet Source	<1 %
11	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 %
12	eprints.usm.my Internet Source	<1 %
13	Muhammad Umair Ali, Amad Zafar, Seonghan Kim, Kwang Su Kim, Seung Won Lee. "From task-specific to foundation models: A paradigm shift in medical vision-language analysis", Computer Science Review, 2026 Publication	<1 %
14	pmc.ncbi.nlm.nih.gov Internet Source	<1 %
15	www.mdpi.com Internet Source	<1 %
16	xin-xia.github.io Internet Source	<1 %
17	theses.hal.science Internet Source	<1 %
18	Submitted to University of Leeds Student Paper	<1 %
19	Submitted to Ghana Technology University College Student Paper	<1 %

20	Submitted to The University of the West of Scotland Student Paper	<1 %
21	Submitted to University of New South Wales Student Paper	<1 %
22	ruor.uottawa.ca Internet Source	<1 %
23	Submitted to Universiti Malaysia Pahang Student Paper	<1 %
24	Zhang, Yue. "Interactive Analysis of Single-Cell RNA-Sequencing Data", University of Washington, 2022 Publication	<1 %
25	Ceulemans, H.. "Apparatus and methods used in a slow neutron resonance scattering experiment", Nuclear Instruments and Methods, 196212 Publication	<1 %
26	Submitted to De LaSalle - College of Saint Benilde Student Paper	<1 %
27	Hanoi Pedagogical University 2 Publication	<1 %
28	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 %
29	Submitted to Universiti Malaysia Pahang Al-Sultan Abdullah (UMPSA) Student Paper	<1 %

30	www.southlewis.org Internet Source	<1 %
31	utoronto.scholaris.ca Internet Source	<1 %
32	www.arxiv.org Internet Source	<1 %
33	Dongming Chen, Mingshuo Nie, Zhen Wang, Huilin Chen, Dongqi Wang. "A Negative Sample-Free Graph Contrastive Learning Algorithm", Mathematics, 2024 Publication	<1 %
34	Ngigi, William K.. "Open-Set Recognition in Computer Vision.", Indiana University of Pennsylvania Publication	<1 %
35	dspace.daffodilvarsity.edu.bd:8080 Internet Source	<1 %
36	opus.hs-furtwangen.de Internet Source	<1 %
37	Rui Duan, Liu Cheng, Zhunan Shen, Xiangwei Kong, Mingzhu Yu, Zhitong Liu, Yunpeng Zhu. "R3PTL: refine, reuse, and remix – an innovative partial transfer learning framework for intelligent machinery fault diagnosis with sample scarcity", Mechanical Systems and Signal Processing, 2025 Publication	<1 %
38	Samuel Sousa, Michael Jantscher, Mark Kröll, Roman Kern. "Large Language Models for Electronic Health Record De-Identification in English and German", Information, 2025 Publication	<1 %

39	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 %
40	Submitted to University of Sydney Student Paper	<1 %
41	vdeborto.github.io Internet Source	<1 %
42	Florian Nigsch, John B. O. Mitchell. "How To Winnow Actives from Inactives: Introducing Molecular Orthogonal Sparse Bigrams (MOSBs) and Multiclass Winnow", Journal of Chemical Information and Modeling, 2008 Publication	<1 %
43	Guankun Wang, Long Bai, Junyi Wang, Kun Yuan et al. "EndoChat: Grounded multimodal large language model for endoscopic surgery", Medical Image Analysis, 2025 Publication	<1 %
44	International Association of Geodesy Symposia, 1995. Publication	<1 %
45	Pan Shi, Zijian Zhao, Kaidi Liu, Feng Li. "Attention-based spatial–temporal neural network for accurate phase recognition in minimally invasive surgery: feasibility and efficiency verification", Journal of Computational Design and Engineering, 2022 Publication	<1 %
	Phenikaa University	

46	Publication	<1 %
47	ebin.pub Internet Source	<1 %
48	hal.science Internet Source	<1 %
49	indah.ump.edu.my Internet Source	<1 %
50	ir.library.osaka-u.ac.jp Internet Source	<1 %
51	scholar.sun.ac.za Internet Source	<1 %
52	Cheolhee Yoo. "Handbook of Geospatial Approaches to Sustainable Cities", CRC Press, 2024 Publication	<1 %
53	Shaowei Yang, Yangxia Xiang, Zhuo Long, Xiaoguang Ma, Qichuan Ding, Jie Jia. "Fault Diagnosis of Harmonic Drives Based on an SDP-ConvNeXt Joint Methodology", IEEE Transactions on Instrumentation and Measurement, 2023 Publication	<1 %
54	"Medical Image Computing and Computer Assisted Intervention – MICCAI 2020", Springer Science and Business Media LLC, 2020 Publication	<1 %
55	Fuqin Deng, Jiaming Zhong, Nannan Li, Lanhui Fu, Bingchun Jiang, Yi Ningbo, Feng Qi, He Xin, Tin Lun Lam. "Text-guided Graph Temporal Modeling for few-shot video classification",	<1 %

	Engineering Applications of Artificial Intelligence, 2024 Publication	
56	Kadir Kirtac, Nizamettin Aydin, Joël L. Lavanchy, Guido Beldi, Marco Smit, Michael S. Woods, Florian Aspart. "Surgical Phase Recognition: From Public Datasets to Real-World Data", Applied Sciences, 2022 Publication	<1 %
57	Kong, Fanjie. "Advancing Vision Intelligence Through the Development of Efficiency, Interpretability and Fairness in Deep Learning Models", Duke University, 2024 Publication	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		