

Universiteit Leiden

Master Computer Science

Evaluating the Transferability of EchoCLIP to LUMC Echocardiography Data

Name:	Max Jacobs
Student ID:	s4225414
Date:	27/07/2026
Specialisation:	Artificial Intelligence
1st supervisor:	Joost Batenburg
2nd supervisor:	RJ van der Geest
3rd supervisor:	Li-Hsin Chen

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Abstract

Echocardiography produces large amounts of ultrasound video and accompanying clinical report text, but much of this data is difficult to use for supervised machine learning because structured labels are often unavailable. Vision-language foundation models such as EchoCLIP offer a way to learn reusable representations from routinely collected scan-report pairs. However, it remains unclear how well such models transfer across institutions, where imaging protocols, patient populations, and reporting practices may differ.

This thesis evaluates the transferability of EchoCLIP to echocardiography data from Leiden University Medical Center (LUMC). We apply the pretrained EchoCLIP model directly to LUMC data, finetune the image and text encoders separately, and test whether integrating multiple echocardiographic views improves performance beyond the original A4C-centered setup. Model performance is evaluated using image and text embedding-space visualization, zero-shot classification, and cross-modal retrieval.

The results show that EchoCLIP transfers meaningful visual structure to LUMC data, as seen in the diagnostic-code organization of the image embedding space and in zero-shot device-detection performance. However, this transfer is incomplete: cross-modal retrieval remains weak, indicating limited alignment between individual LUMC scans and reports. Finetuning improves performance mainly on the text side, with text-encoder finetuning improving retrieval while broader image-encoder finetuning reduces zero-shot LVEF performance and weakens clinically meaningful embedding structure. The multi-view experiments are inconclusive, but show that simply adding A2C and PLAX views through late fusion is not enough to overcome the underlying scan-report alignment problem.

Overall, the findings indicate that EchoCLIP transfers meaningful visual representations to LUMC data, but that cross-institutional transfer is limited by differences in report structure and image-text alignment. Successful medical vision-language transfer depends not only on model architecture and image data, but also on the specificity and structure of the clinical reports used as supervision.

Contents

1 Introduction **1**

2 Related Work **3**

2.1 Deep Learning in Echocardiography: Task-Specific Approaches 3

2.2 Foundation Models and Representation Learning in Cardiac Imaging 3

2.3 Contrastive Vision-Language Learning 4

2.4 Vision-Language Models for Echocardiography 5

3 Methods **6**

3.1 Echocardiography Images and Views 6

3.2 Clinical Reports and Dataset Differences 7

3.3 Dataset and Label Construction 9

3.4 EchoCLIP Model Architecture 11

3.5 Evaluation Methods 13

3.5.1 Embedding Space Visualization 13

3.5.2 Zero-shot Classification 15

3.5.3 Cross-modal Retrieval 15

3.6 Multi-view Experimental Setup 16

4 Baseline Model Transfer Evaluation **18**

4.1 Embedding Space Analysis 18

4.1.1 Image Embedding Space 18

4.1.2 Text Embedding Space 20

4.2 Zero-shot Testing 21

4.2.1 LVEF Prediction 21

4.2.2 Device Classification 23

4.3 Retrieval Evaluation 24

5 Finetuning **25**

5.1 Text-encoder Finetuning 26

5.2 Image-encoder Finetuning 29

6 Multi-view Integration **32**

6.1 Multi-view Results 32

7 Discussion **33**

7.1 Summary of Research Questions 33

7.2 Why Scan-report Alignment is Limited 35

7.3 Reporting Practices 36

7.4 Future Work 37

8 Conclusion **38**

References **42**

1 Introduction

Echocardiography laboratories generate vast amounts of clinical data each year, including high-dimensional ultrasound video and accompanying diagnostic reports. Despite this abundance of data, relatively little of it exists in a structured, machine-readable format. Most contemporary deep learning models for echocardiography rely on explicitly labeled datasets and are trained for narrowly defined tasks. For example, a model may be trained to predict left ventricular ejection fraction (LVEF) using scans paired with manually curated EF values. While such task-specific models can achieve strong performance, their development depends on structured annotations that are expensive, time-consuming, and require expert cardiologists to produce. As a result, large portions of routinely acquired clinical data remain underutilized.

Many recent advances in machine learning have been driven by the emergence of foundation models, which are large-scale models trained on broad and diverse data using general learning objectives that enable transfer across tasks rather than being optimized for a single objective. In computer vision and natural language processing, models such as CLIP have demonstrated that aligning images with natural language descriptions through contrastive learning enables zero-shot generalization and flexible usage across tasks. This framework enables models to use larger amounts of routinely collected data and to be reusable across multiple tasks. Because deep learning performance improves with the amount of data and size of the model, expanding the ability to leverage larger amounts of clinical data is critical for improving model capability. Applying CLIP to echocardiography enables supervision at scale by utilizing existing scan-report pairs without requiring additional manual annotation.

Building on this, EchoCLIP [3] applies contrastive vision-language learning to echocardiography. The model is trained on paired ultrasound videos and clinical reports, learning to embed scans and their corresponding text into a shared representation space. EchoCLIP associates visual cardiac features with the semantic content of diagnostic reports, enabling zero-shot classification and cross-modal retrieval without task-specific structured labels. Christensen et al. introduced EchoCLIP and demonstrated promising zero-shot classification and retrieval performance across multiple tasks, suggesting that clinically meaningful representations can emerge from large-scale weak supervision. However, the extent to which such representations generalize across institutions

remains unclear. Furthermore, the original model was trained exclusively on the apical four-chamber view, potentially limiting the breadth of visual information available to the learned representation.

In this work, we investigate the transferability of EchoCLIP to echocardiography data from a new institutional setting. We use Leiden University Medical Center (LUMC) echocardiography scans and their associated clinical reports as the target dataset, and apply the pretrained EchoCLIP model to these scan-report pairs. The goal is not to train a new foundation model from scratch, but to evaluate how well an existing pretrained model can be reused, adapted, and extended on local clinical data. The thesis focuses on three main questions. First, how well does the pretrained EchoCLIP model generalize to the LUMC dataset? Second, what degree and type of finetuning is required to make EchoCLIP perform better on LUMC scan-report pairs? Third, does incorporating multiple echocardiographic views provide useful information beyond the original A4C-centered setup? These questions are addressed sequentially: the baseline evaluation establishes the extent of transfer, the finetuning experiments test how the model can be adapted, and the multi-view experiments examine whether additional visual information improves the representation. Together, these experiments evaluate both the promise and the limitations of EchoCLIP as a transferable foundation model for echocardiography.

2 Related Work

2.1 Deep Learning in Echocardiography: Task-Specific Approaches

A major focus of deep learning in echocardiography has been automated assessment of cardiac function. Left ventricular ejection fraction (LVEF) is especially important because it summarizes systolic function and is widely used in diagnosis, prognosis, and treatment decisions. Ouyang et al. introduced video-based deep learning methods capable of beat-to-beat assessment of cardiac function [8]. Zhang et al. demonstrated fully automated echocardiogram interpretation in clinical practice, showing that deep learning systems could estimate multiple clinically relevant measurements from routine echocardiographic data [10]. Subsequent multicohort work further validated automated assessment of systolic and diastolic function using deep learning [9].

Deep learning models have also been developed for disease detection and structural abnormality identification. Ghorbani et al. demonstrated deep learning-based interpretation of echocardiograms across multiple diagnostic categories [5]. Holste et al. reported automated detection of severe aortic stenosis from single-view echocardiography [6]. Together, these studies show that echocardiographic images and videos contain enough visual information for neural networks to learn clinically useful features related to ventricular function, valve disease, chamber size, and other pathological findings.

The limitation is not that task-specific models fail, but that they scale poorly across clinical questions. These systems are usually trained for defined objectives using structured labels. Each new task therefore requires a suitable labeled dataset and often a separate training procedure. This increases both annotation requirements and computational cost, especially in clinical settings where many potentially useful tasks exist but only some are available in a structured form. This motivates representation-learning approaches that can make broader use of routinely collected clinical data and produce more broadly applicable models.

2.2 Foundation Models and Representation Learning in Cardiac Imaging

Foundation models address the limitations of task-specific learning by shifting the goal from training one model for one clinical objective to learning representations that can be reused across multiple

downstream tasks. Rather than being optimized only for a predefined label set, these models are pretrained using broad learning objectives and are then adapted for specific applications [1]. This reduces the need to train a new model from scratch for every task and makes the pretrained representation itself the central object of interest.

Large-scale representation learning is also motivated by scaling behavior in modern machine learning, where performance often improves with increases in data, model size, and compute [2]. In medical imaging, this has encouraged the development of modality-specific foundation models that learn general-purpose visual representations from large imaging datasets. For example, Jacob et al. pretrained a vision transformer using self-supervised learning on 36 million cardiovascular magnetic resonance (CMR) images and demonstrated improved performance across segmentation, classification, and landmark localization tasks [7]. These results support the idea that large-scale pretraining can improve robustness and data efficiency in cardiac image analysis.

For echocardiography, this foundation-model framing could support several downstream tasks, such as functional assessment, disease classification, retrieval, and similarity analysis. However, the practical usefulness of such a model depends on whether the learned representation remains meaningful outside the original training setting and whether it can be adapted efficiently to local data. This makes transferability and finetuning central questions when applying foundation models to clinical imaging datasets.

2.3 Contrastive Vision-Language Learning

One way to build a transferable representation model is to use paired images and text during pretraining. This is especially relevant in clinical imaging, where images are routinely accompanied by written reports. Contrastive vision-language learning provides a framework for using this text as supervision by aligning visual inputs with their corresponding language descriptions.

A CLIP-style architecture contains two main components: an image encoder and a text encoder. The image encoder maps an image, frame, or video-derived input into a fixed-dimensional embedding. The text encoder maps a caption, report, or prompt into an embedding of the same dimension. These embeddings are usually normalized and compared using cosine similarity. During contrastive training, matching image-text pairs are treated as positives, while other image-text combinations in the batch are treated as negatives. The training objective encourages the similarity of the true

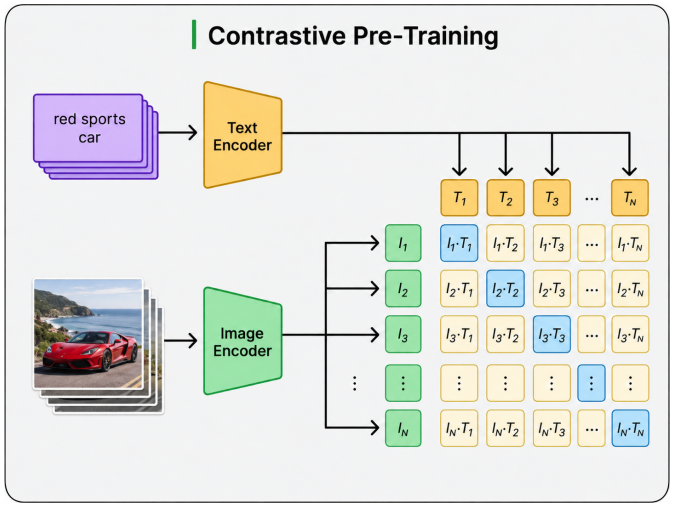


Figure 1: Schematic overview of CLIP-style contrastive vision-language learning. An image encoder and text encoder map paired inputs into a shared embedding space, where cosine similarity is used to align matching image-text pairs.

pair to be higher than the similarity of non-matching pairs.

The shared embedding space can then be reused in several ways. In zero-shot classification, each class is represented by a text prompt, and an image is assigned to the class whose prompt embedding is most similar to the image embedding. In cross-modal retrieval, an image can be used to retrieve the most similar text, or a text query can be used to retrieve the most similar image. These capabilities are central to the appeal of vision-language foundation models because they allow the same pretrained representation to be adapted to a wide variety of downstream tasks.

2.4 Vision-Language Models for Echocardiography

EchoCLIP extends contrastive vision-language learning to echocardiography [3]. The model was trained on paired echocardiogram videos and clinical reports collected from routine care. Following the CLIP-style framework, it uses an image encoder to represent echocardiographic visual input and a text encoder to represent clinical report text. Both modalities are mapped into a shared embedding space, where scans and reports from the same study are encouraged to have similar representations.

Christensen et al. showed that EchoCLIP can perform several downstream tasks without task-specific supervised training. These included left ventricular ejection fraction estimation, detection of intracardiac devices, assessment of chamber dilation, and identification of conditions such as

pericardial effusion and hypertrophy [3]. The reported LVEF performance included a mean absolute error of 7.1% on an external validation dataset [3]. In addition to classification and regression tasks, the model enabled cross-modal retrieval and study-level similarity analysis.

These results demonstrate that clinically meaningful representations can be learned from scan-report pairs rather than from manually curated labels for every separate task. A model that can use this supervision effectively may reduce the need to construct a separate labeled dataset for every downstream objective.

However, the extent to which EchoCLIP transfers beyond its original training setting remains uncertain. A model trained in one institution may not behave the same way when applied to another hospital. This motivates the first research question of this thesis: how well the pretrained EchoCLIP model generalizes to LUMC data without local training. If transfer is incomplete, it is also necessary to determine what kind of adaptation is most useful, motivating the second research question on model finetuning. Finally, because EchoCLIP was developed around apical four-chamber inputs while routine echocardiography studies contain multiple complementary views, the third research question asks whether additional views such as A2C and PLAX provide useful information beyond the original A4C-centered setup. The common thread connecting these questions is how a foundation model developed for echocardiography transfers to a new institutional setting, what limits that transfer, and whether targeted adaptation or additional views can improve its usefulness.

3 Methods

3.1 Echocardiography Images and Views

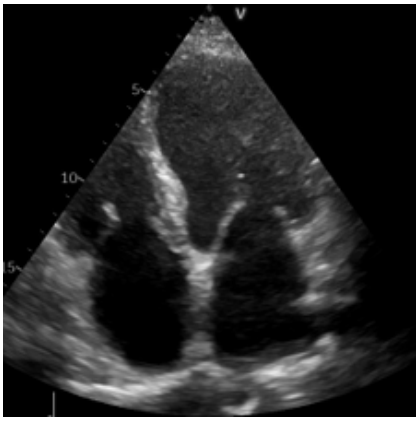
In echocardiography, each examination consists of multiple ultrasound video clips. A full study contains several standard views of the heart, each showing different anatomical structures and functional information. In this thesis, the primary focus is the apical four-chamber view (A4C), which is also the main view used in the original EchoCLIP setup. Additional views, including the apical two-chamber view (A2C) and parasternal long-axis view (PLAX), are considered later as an extension to test whether multi-view information provides useful signal beyond the A4C baseline.

Figure 2 shows representative examples of these views. A4C shows both ventricles and atria from the apex and is commonly used to assess global ventricular function. A2C is also an apical

view, but shows the left atrium and left ventricle from a different plane. PLAX shows the heart from beside the sternum along its long axis, providing a view of the left ventricle, aortic valve, mitral valve, and outflow tract. These views are visually and anatomically distinct, which is why the question of whether EchoCLIP can benefit from additional views is treated separately in the multi-view experiments.

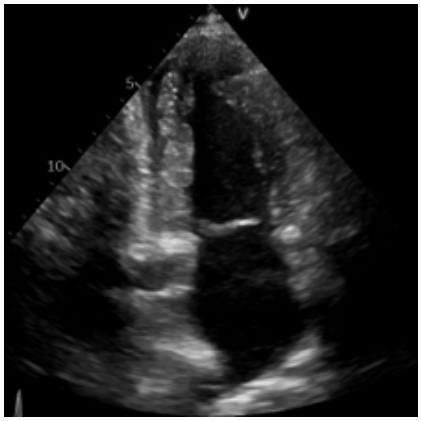
Each echocardiography examination is accompanied by a clinical report written by the interpreting cardiologist. The report summarizes the findings of the study as a whole, drawing on multiple ultrasound views, measurements, and clinical observations. In an EchoCLIP-style vision-language model, these reports provide the textual supervision.

(a) Apical four-chamber view



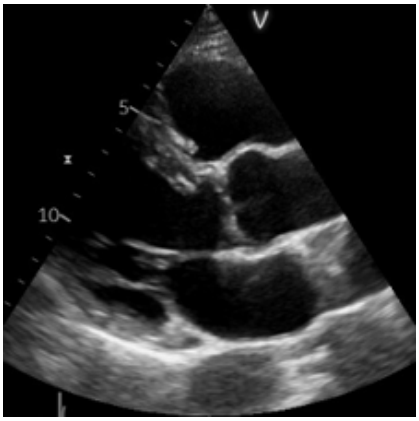
Corresponding report: Non-dilatated LV with good systolic function. Borderline-dilatated RV with good function. High normal dimension aortic root and ascendants.

(b) Apical two-chamber view



Corresponding report: Non-dilated LV with reasonably good function Dilated LA Low mitral valve insufficiency

(c) Parasternal long-axis view



Corresponding report: 1. non-dilated non-hypertrophic LV with overall good function and normal diastolic function for age. 2. biatrial dilation, floppy intra-atrial septum without signs of PFO. Chiari network in the right atrium. 3. no clot vitae of importance. no pulmonary pressure increased.

Figure 2: Examples of echocardiography views used in this thesis. Each image shows a representative frame from a video clip, while the text below gives the corresponding report excerpt. The reports are study-level, meaning they describe all scans for that study.

3.2 Clinical Reports and Dataset Differences

The experiments in this thesis depend on the relationship between two data settings: the data used to train the original EchoCLIP model and the LUMC echocardiography data used for transfer evaluation.

The original EchoCLIP reports appear to be particularly well suited for contrastive image-text learning. The authors describe the report text as highly structured and repetitive, with recurring phrases, severity terms, and numerical values. This structure allowed them to develop a custom report tokenizer for EchoCLIP-R, an extension of EchoCLIP, compressing common echocardiography concepts into a shorter token sequence [3]. For example, report phrases containing findings such as ventricular hypertrophy, ejection fraction, or severity descriptors could be represented using standardized tokens. This matters because contrastive learning benefits from text that is both clinically informative and consistent enough to align with the visual input. This tokenizer extension is beyond the scope of this project, we will be testing and adapting only the base EchoCLIP model, but it indicates how highly structured and consistent their reports are.

The LUMC dataset differs from this pretraining setting in several important ways. Like the EchoCLIP data, it consists of routinely collected echocardiography examinations with associated clinical reports. However, the reporting style is different. The LUMC reports are relatively concise free-text summaries written by the interpreting cardiologist, whereas the reports used in the original EchoCLIP study were structured reports produced using Syngo Dynamics. Syngo Dynamics is a reading and reporting platform in which the cardiologist reviews the study, records measurements and findings in structured fields or templates, and then generates a final report from those entered values and selected findings.

This difference is quantifiable through report length. Using BPE-style tokenization, the 325,681 LUMC reports contain a mean of 50.5 ± 20.6 tokens. In comparison, Christensen et al. report that the original EchoCLIP reports contained 530.3 ± 154.7 BPE tokens per report [3]. The LUMC reports are therefore much shorter by direct token-count comparison, which suggests that the textual supervision available for contrastive learning is more concise.

In the EchoCLIP paper, zero-shot LVEF estimation can be formulated using explicit numerical prompts such as "LVEF is 73%". In the LUMC data, LVEF information is more often available through broader diagnostic categories or short descriptive statements, such as normal function, mild dysfunction, or severe dysfunction. The information is therefore still clinically meaningful, but is less standardized and less directly numerical than in the original EchoCLIP setting.

These dataset differences are central for interpreting the results. If EchoCLIP performs worse on LUMC data than in the original paper, this does not necessarily mean that the pretrained visual

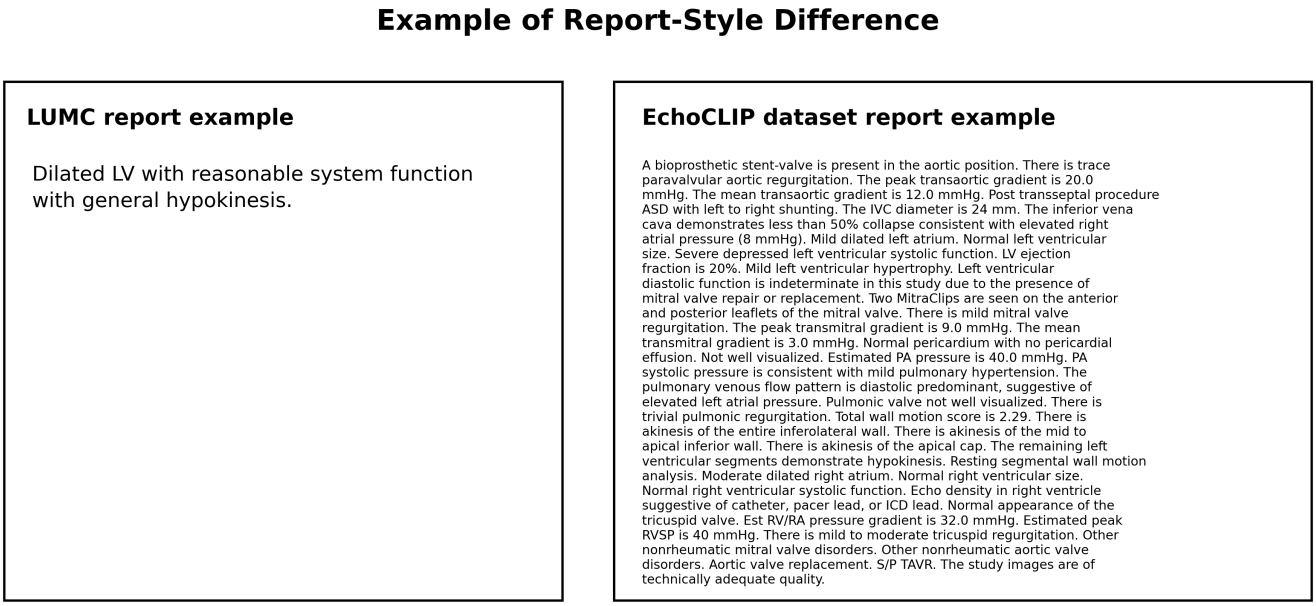


Figure 3: Side-by-side comparison of example report styles from the LUMC dataset and the original EchoCLIP training setting. The LUMC example is shorter and more concise, while the EchoCLIP-style example contains more structured and detailed report text.

representation has failed. It may reflect differences in reporting style, language, label availability, and the specificity of the textual supervision.

3.3 Dataset and Label Construction

The final paired LUMC dataset contains 325,681 scan-report pairs from 84,541 echocardiography studies and 40,749 unique patients. Each scan is linked to the corresponding study-level clinical report. This means that multiple scans from the same examination can share the same report, since the report summarizes the findings of the full echocardiography study rather than describing each individual scan separately.

For comparison, the original EchoCLIP model was pretrained on a larger dataset of paired echocardiogram videos and clinical report text. Christensen et al. report 1,032,975 video-text pairs derived from 224,685 echocardiography studies across 99,870 patients [3].

In addition to reports, the LUMC dataset contains diagnostic-code metadata. These codes provide structured information about clinical findings, procedures, and diagnostic categories associated with each study. They are especially important in this project because the LUMC reports often do not contain precise numerical measurements, such as exact LVEF values, in a

consistent form. The diagnostic codes therefore provide an alternative set of structured evaluation targets, allowing model performance to be tested against labels even when the corresponding information is described qualitatively in the report text. A single report can contain multiple diagnostic codes. Therefore, diagnostic-code groups are not always mutually exclusive, and some overlap between categories is expected.

Diagnostic codes are used in three main ways in this thesis. First, they are used for embedding space analysis. After image or text embeddings are generated, embeddings are grouped by diagnostic code to test whether clinically related codes occupy related regions of the pretrained image embedding space. The codes are not provided as inputs to the model; they are used only after embedding generation to interpret the structure of the visual representation.

Second, diagnostic codes are used to construct the zero-shot LVEF classification task. Because precise numerical LVEF values are not consistently available in the LUMC reports, LVEF evaluation is formulated as a three-class task using diagnostic-code categories. Code 000 corresponds to normal LVEF, code 400 corresponds to mild dysfunction, and code 401 corresponds to severe dysfunction. These categories provide a clinically meaningful but coarse evaluation of whether the pretrained image representation contains information about left ventricular systolic function.

Third, diagnostic codes are used to construct device-detection tasks. These tasks evaluate whether EchoCLIP can identify visual evidence of implanted or interventional cardiac devices in a zero-shot setting. Separate binary labels are constructed for pacemaker, Impella, TAVR, and MitraClip findings, as well as an aggregate any-device label. The resulting evaluation sets are highly imbalanced: pacemaker labels contain 3,353 positive and 321,484 negative examples, Impella labels contain 548 positive and 324,289 negative examples, TAVR labels contain 400 positive and 324,437 negative examples, MitraClip labels contain 299 positive and 324,538 negative examples, and the any-device task contains 4,735 positive and 320,102 negative examples.

These diagnostic-code labels are used only for evaluation and analysis. The baseline zero-shot experiments do not train a supervised classifier on LUMC labels. Instead, the model produces image and text embeddings using the pretrained EchoCLIP encoders, and diagnostic codes are used afterward to evaluate whether the resulting predictions or embedding structures correspond to clinically meaningful categories.

3.4 EchoCLIP Model Architecture

The experiments in this thesis use the pretrained EchoCLIP model as the baseline representation model. EchoCLIP follows a two-tower architecture with separate image and text encoders. The image tower is based on a ConvNeXt-Base encoder, which maps an echocardiographic image frame to a fixed-dimensional visual embedding. Although echocardiography data are acquired as videos, the original EchoCLIP training procedure samples individual frames from these videos during training.

The text tower is a decoder-only Transformer based on the original CLIP text encoder architecture, with a context length of 77 tokens. Report text and class prompts are tokenized and passed through the text encoder to produce embeddings in the same dimensional space as the image embeddings. Both image and text embeddings are L2-normalized before cosine similarity is computed.

During contrastive training, the similarity between every image embedding and every text embedding in a batch is computed, producing an image-text similarity matrix. The diagonal entries correspond to matching scan-report pairs, while the off-diagonal entries correspond to non-matching pairs. The contrastive objective encourages matching pairs to have higher similarity than non-matching pairs. In this thesis, the pretrained EchoCLIP checkpoint is used without local training for the baseline transfer experiments and as the starting point for the finetuning experiments. Figure 4 visualizes this architecture and training objective.

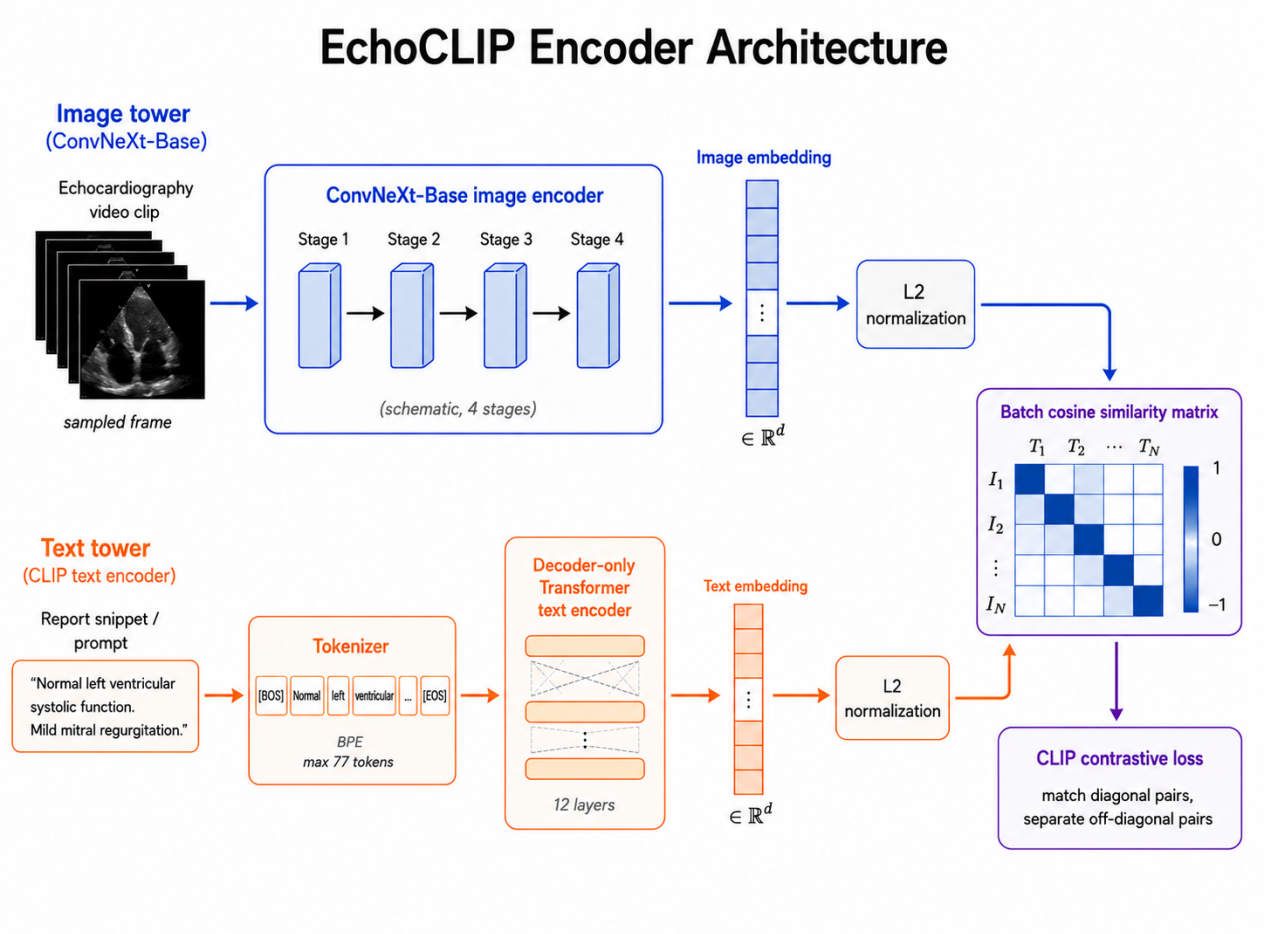


Figure 4: Schematic overview of the EchoCLIP encoder architecture and contrastive training objective. A frame is encoded by a ConvNeXt-Base image tower with four stages, while a report or prompt is tokenized and encoded by a 12-layer CLIP-style decoder-only Transformer text tower. The resulting image and text embeddings are projected into the same d-dimensional space, L2-normalized, and compared using cosine similarity. During contrastive training, similarity is computed over a batch so that matching image-text pairs are encouraged to align, while non-matching pairs are separated.

3.5 Evaluation Methods

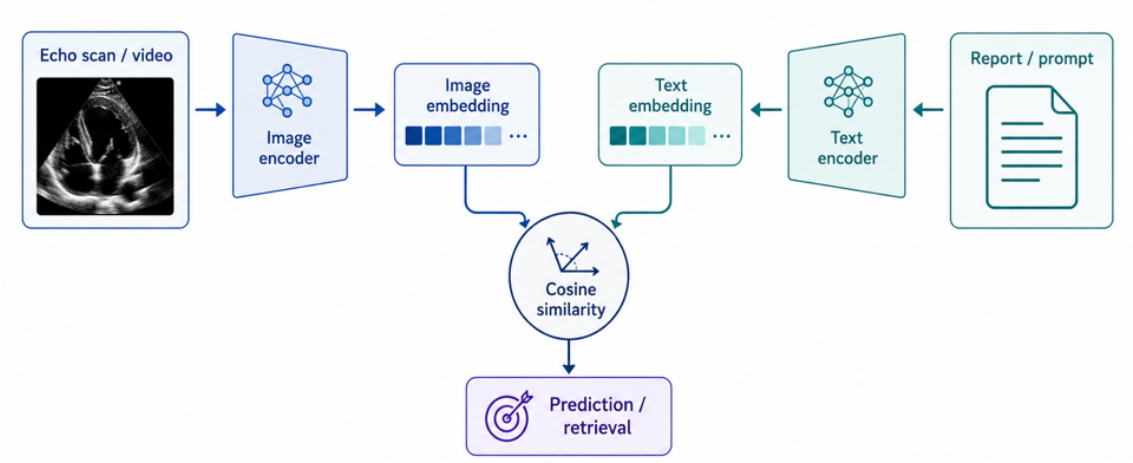
EchoCLIP is evaluated using three complementary methods: embedding space visualization, zero-shot classification, and cross-modal retrieval. These methods provide the main performance measures used across the experimental stages of this thesis. Embedding space visualization evaluates whether the pretrained image and text encoders preserve clinically meaningful structure on LUMC scans and reports, respectively. Zero-shot classification evaluates whether this structure can be accessed through text prompts for clinically relevant tasks such as LVEF estimation and device detection. Cross-modal retrieval evaluates whether LUMC scans and reports are aligned in the shared image-text embedding space.

These evaluations are used to answer the research questions at each stage of the experimental pipeline. Baseline transfer is assessed by applying embedding analysis, zero-shot classification, and retrieval to the pretrained EchoCLIP model without local training. Finetuning is assessed by testing whether model adaptation improves these metrics while preserving clinically meaningful structure. Multi-view integration is assessed by testing whether additional A2C and PLAX views improve zero-shot LVEF classification beyond A4C alone. The evaluation metrics therefore define how model performance is measured for each research question.

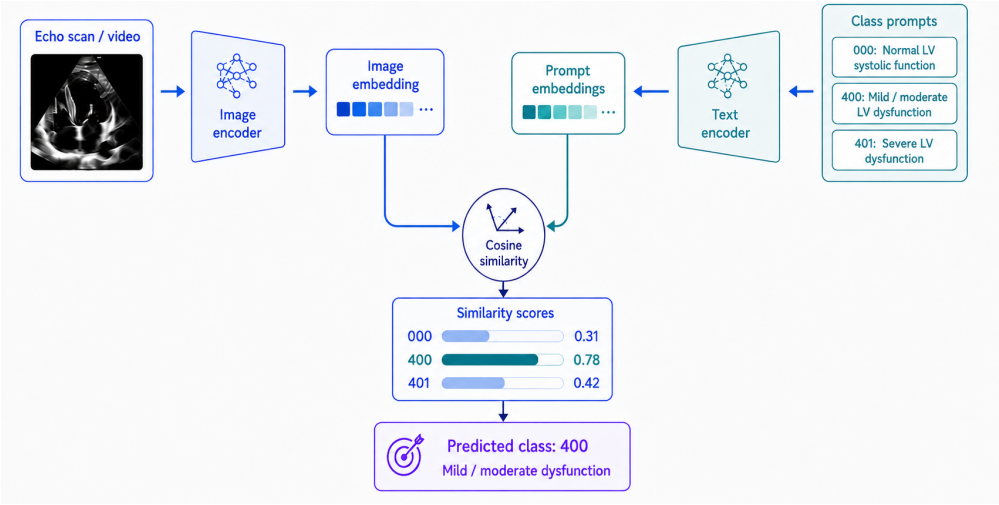
3.5.1 Embedding Space Visualization

Embedding space visualization is used to evaluate the pretrained image and text encoders independently from each other. For the image encoder, frame-level image embeddings are first computed for each scan and then mean-pooled to obtain a single scan-level representation. Diagnostic codes are used after embedding generation to group and interpret the resulting image embeddings, but they are not provided as inputs to the model. For the text encoder, because there is only one report per study, the reports are simply embedded and then grouped by their corresponding diagnostic codes.

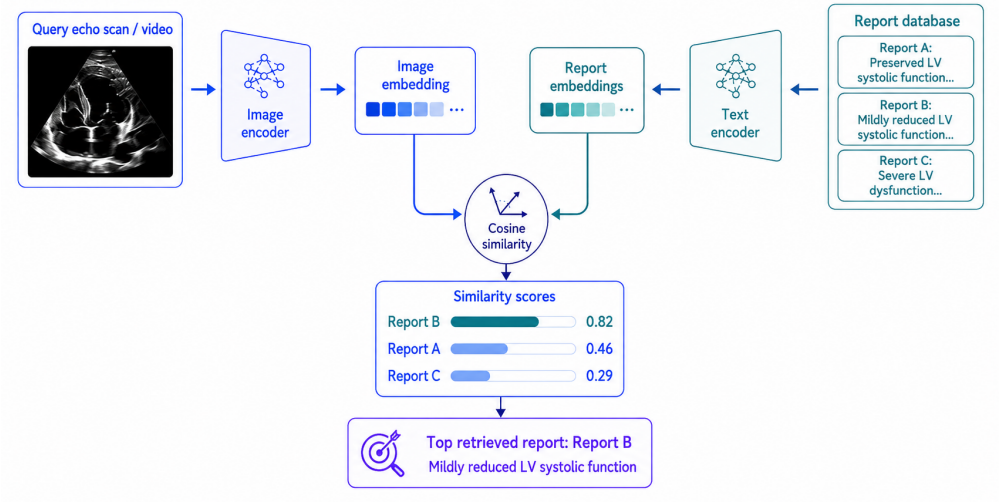
Two dimensionality-reduction methods are used: principal component analysis (PCA) and uniform manifold approximation and projection (UMAP). PCA provides a linear projection of the embedding space and is useful as a conservative view of global structure. UMAP is nonlinear and is used to visualize local neighborhood structure, but its clusters should be interpreted cautiously because it can exaggerate separation when high-dimensional points are projected into



(a) General EchoCLIP image-text embedding setup.



(b) Zero-shot classification using class prompts.



(c) Cross-modal retrieval using report embeddings.

Figure 5: Overview of the main EchoCLIP evaluation tasks used in this thesis. Echocardiographic scans and reports are encoded into a shared image-text embedding space. This shared representation can be evaluated through zero-shot classification by comparing image embeddings with class-prompt embeddings, or through cross-modal retrieval by ranking report embeddings according to their similarity to a query image, and vice versa.

two dimensions. This analysis addresses the first research question at the level of representation transfer: if EchoCLIP transfers to LUMC data, clinically related diagnostic-code groups should show meaningful organization in the embedding spaces.

3.5.2 Zero-shot Classification

Zero-shot classification is performed by comparing image embeddings with text embeddings generated from class prompts. For each task, a set of prompts is created to describe the possible clinical classes. The text encoder maps these prompts into the shared embedding space, and the predicted class is selected based on the highest cosine similarity between the image embedding and the class-prompt embeddings (Figure 5b).

For LVEF classification, the task is formulated as a three-class problem using diagnostic-code labels for normal LVEF, mild dysfunction, and severe dysfunction. Performance is evaluated using accuracy and Macro-F1. Accuracy measures the overall fraction of correct predictions, while Macro-F1 computes the F1-score separately for each class and averages them equally. Macro-F1 is emphasized because the LVEF labels are imbalanced, so accuracy alone can be misleading.

For device detection, each device category is evaluated as a binary zero-shot task. Each device is represented using text prompts that describe the presence or absence of that device. For example, pacemaker detection can be evaluated by comparing an image embedding with prompts such as "an echocardiogram with a pacemaker" and "an echocardiogram without a pacemaker." Performance is measured using ROC AUC and average precision (AP). ROC AUC measures whether positive examples tend to receive higher scores than negative examples across possible thresholds. AP summarizes precision-recall performance and is especially important for rare findings, because it is sensitive to false positives among the highest-scoring predictions.

3.5.3 Cross-modal Retrieval

Cross-modal retrieval evaluates whether image and report embeddings are aligned at the level of individual scan-report pairs. After all images and reports are encoded, cosine similarity is computed between a query embedding and a set of candidate embeddings from the other modality. In image-to-report retrieval, an image embedding is used as the query, and all candidate report embeddings are ranked from most to least similar. The correct report is the report associated with

that image's study. In report-to-image retrieval, the direction is reversed: a report embedding is used as the query, and candidate image embeddings are ranked by similarity to that report. A retrieval result is successful when the true paired item appears near the top of the ranked list (Figure 5c).

This task is stricter than zero-shot classification because it requires fine-grained alignment between individual scans and their corresponding reports. In zero-shot classification, the model only needs to assign an image to a broad clinical category, such as normal LVEF or severe dysfunction. In retrieval, the model must distinguish the correct report from many other clinically similar reports. This is especially challenging in echocardiography because many reports describe common findings using similar language.

Retrieval performance is evaluated using mean reciprocal rank (MRR), median rank, and recall at K. MRR measures how highly the correct match is ranked on average, giving larger values when the correct match appears near the top of the ranked list. Median rank reports the median position of the correct match across all queries. Recall@K measures the fraction of queries for which the correct match appears within the top K retrieved candidates. In this thesis, retrieval is especially important for evaluating whether the pretrained image-text alignment transfers to LUMC reports, and for testing whether text-encoder finetuning improves that alignment.

3.6 Multi-view Experimental Setup

The multi-view experiments use A4C, A2C, and PLAX views from the LUMC dataset. Each view is first evaluated separately using the same zero-shot LVEF classification setup described in Section 3.5.2. This makes it possible to test whether the pretrained EchoCLIP image encoder represents each view in a clinically useful way. Because the original EchoCLIP setup is centered on A4C, separate view-level evaluation is important: a view may contain useful clinical information in principle, but still be less effective if the pretrained model does not represent that view as well.

Multi-view integration is then performed using late fusion. Each available view is encoded separately by the image encoder, producing view-specific image embeddings and class-prompt similarity scores. These scores are then combined across views, and the final prediction is made from the combined class scores. This approach tests whether A2C and PLAX provide complementary signal when added to A4C, without changing the underlying image encoder or training a new

study-level architecture.

Late fusion is used as a conservative extension of the baseline model. More complex approaches, such as early fusion, multiple-instance learning, or attention-based pooling, could combine information from multiple clips or views more flexibly. However, these methods would require additional model design and training, making it harder to separate the effect of additional visual information from the effect of architectural changes. Since the image-encoder finetuning experiments already test whether modifying the visual representation is reliable, the multi-view experiment is designed as a simpler suitability test: it asks whether the existing EchoCLIP representation can benefit from additional LUMC views when those views are combined at the score level.

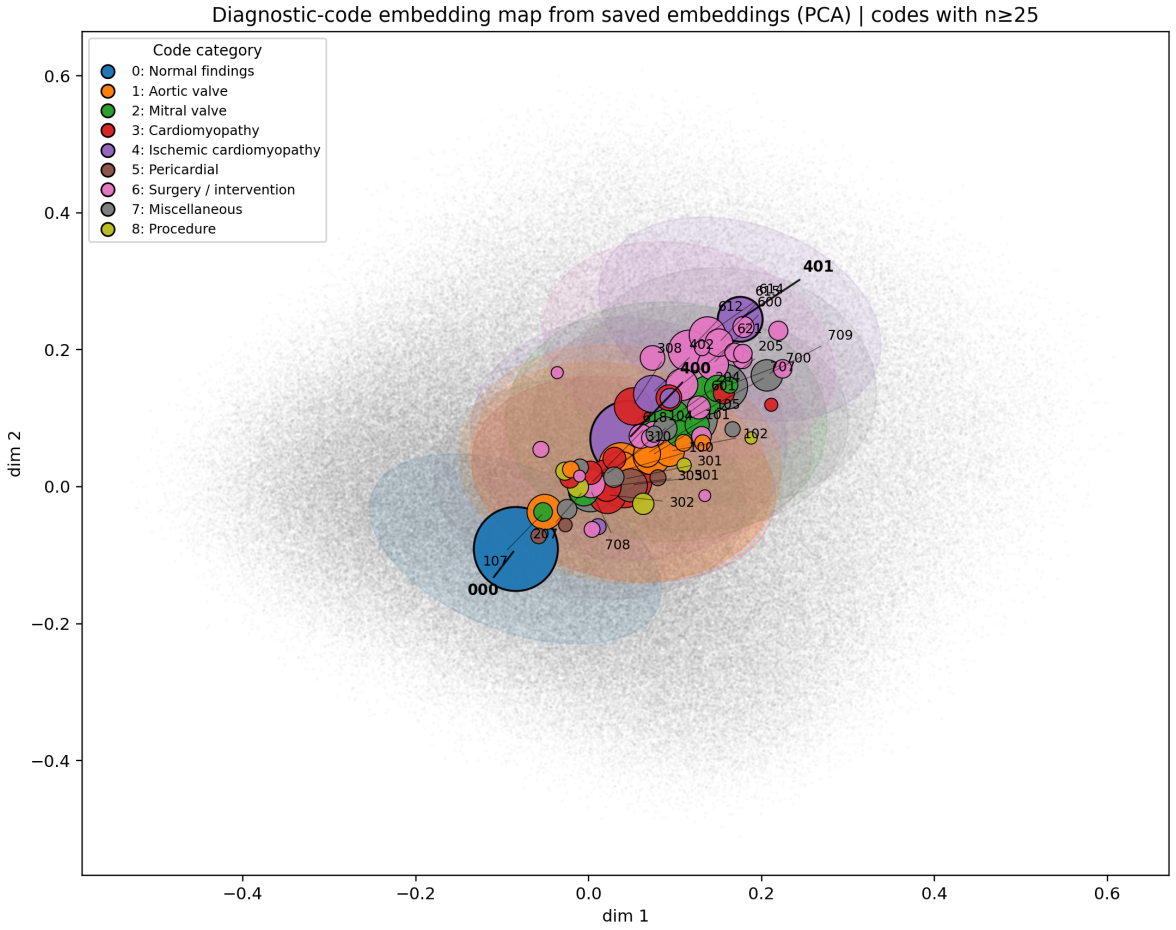
4 Baseline Model Transfer Evaluation

4.1 Embedding Space Analysis

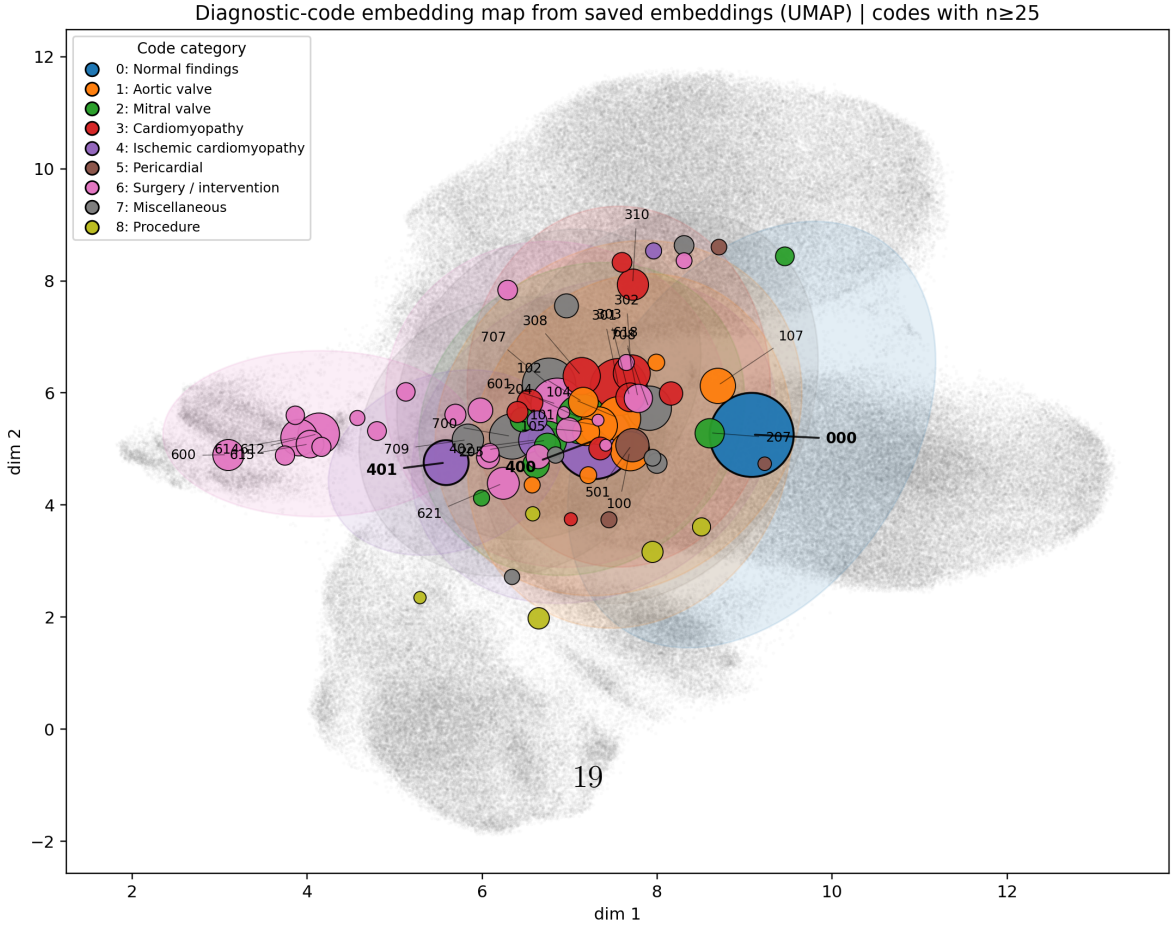
We begin the baseline evaluation with embedding space analysis to examine whether EchoCLIP produces clinically meaningful representations on LUMC data before evaluating explicit prediction or retrieval tasks. As described in Section 3.5.1, this analysis is performed separately for image and text encoders using diagnostic codes as an external reference to interpret embedding spaces. Comparing these two embedding spaces helps to distinguish whether the transfer is stronger on the visual side or on the textual side.

4.1.1 Image Embedding Space

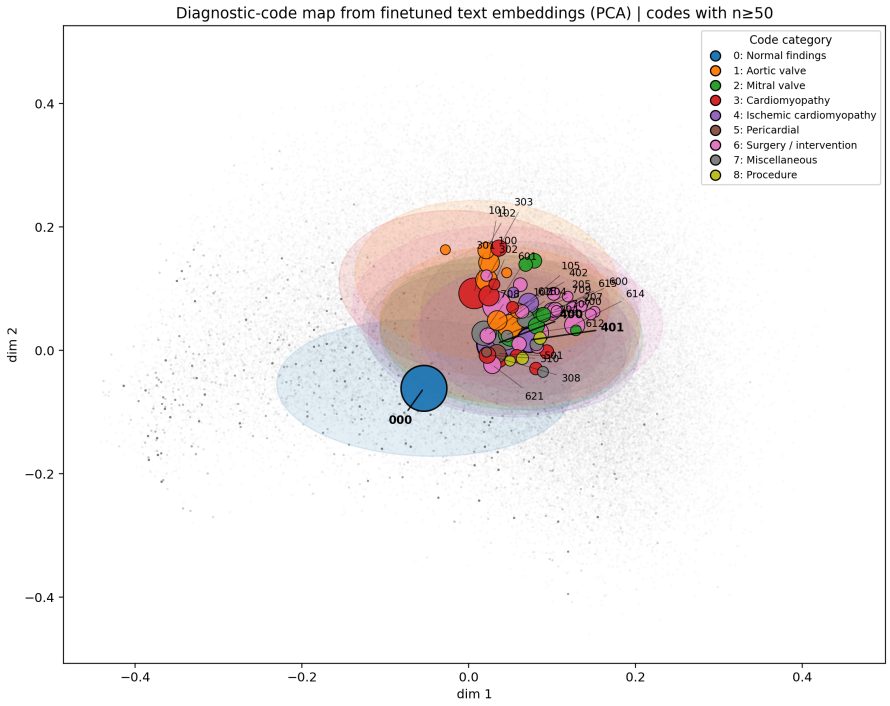
Both projections suggest that the image encoder captures some clinically meaningful structure. A simple example is given by codes 401, 400, and 000, which correspond to severe LVEF dysfunction, mild LVEF dysfunction, and normal findings, respectively. In Figures 6a and 6b, code 000 is relatively well separated from the other diagnostic codes, code 400 lies in a more intermediate region, and code 401 is shifted furthest away. The separation is not perfect, but it is in the direction that we would expect if the embedding contains information related to cardiac function.



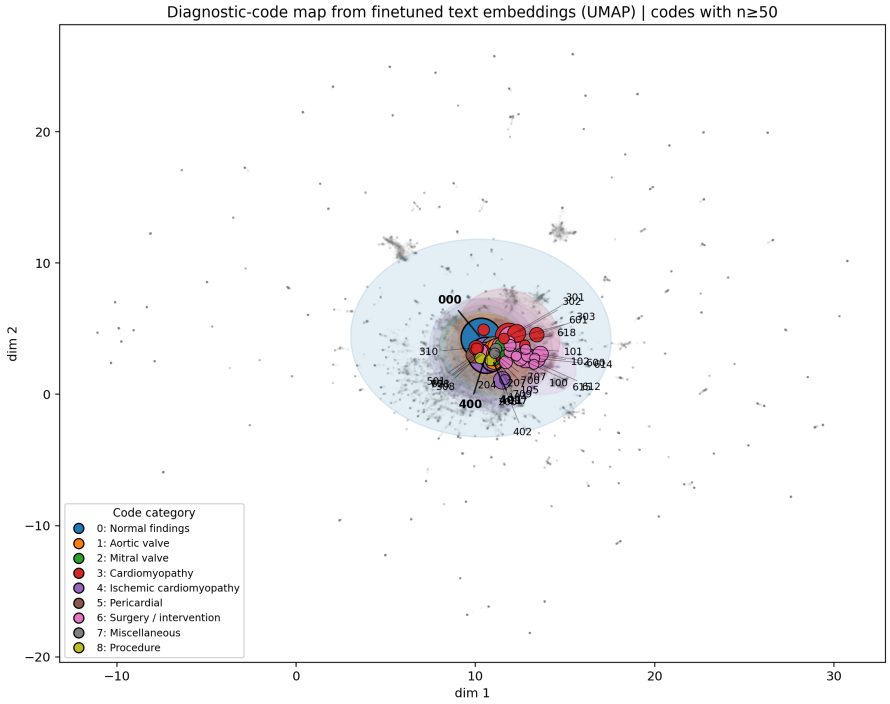
(a) PCA projection



4.1.2 Text Embedding Space



(a) PCA projection



(b) UMAP projection

Figure 7: Comparison of PCA and UMAP projections of the diagnostic-code embedding space.

The text embedding space shows much less diagnostic-code structure than the image embedding space. In Figures 7a and 7b, the centroids are clustered close together and show substantial overlap. Unlike the image embedding space, where some clinically meaningful separation is visible between normal findings and LVEF dysfunction codes, the report embeddings do not form clearly separated regions corresponding to different clinical categories.

This lack of separation is especially visible in the UMAP projection. Most report embeddings are concentrated in a dense central region, with diagnostic-code centroids lying very close to one another. Although there are some scattered outlying report embeddings, these do not correspond to a clearly organized structure. The PCA projection shows a similar pattern: the main diagnostic-code groups overlap strongly, and the broad ellipses around the categories occupy largely the same region of the space.

These results suggest that the pretrained text encoder does not organize LUMC reports into a clinically structured embedding space. This is concerning, but there will be more concrete numerical tests in Section 4.3 to verify whether the text encoder struggles to distinguish between LUMC reports.

4.2 Zero-shot Testing

The second baseline experiment evaluates whether the pretrained embedding space can be used for zero-shot clinical classification on LUMC data. Instead of training a classifier on LUMC labels, each class is represented by a text prompt and the predicted class is selected by comparing image embeddings with prompt embeddings (Figure 5b). Zero-shot performance therefore tests whether information in the image representation can be accessed through language.

4.2.1 LVEF Prediction

A key limitation of the LUMC dataset is the absence of consistent, precise numerical LVEF values in the reports. As mentioned earlier, the EchoCLIP paper often uses explicit percentage estimates (e.g., "LVEF is 55%") while the LUMC reports typically contain qualitative descriptions (e.g., "mildly reduced function") or broad ranges. As a result, we reformulated the task as a 3-way classification problem using diagnostic codes:

- **000**: Normal LVEF
- **400**: Mild dysfunction
- **401**: Severe dysfunction

Table 1: Zero-shot LVEF 3-way classification results

Metric	**Value**
Number of samples (N_{valid})	199,776
Accuracy	0.5046
Macro F1-score	0.4146

Confusion Matrix (rows = true, cols = predicted)

	000	400	401
000	65820	29142	18572
400	13609	30895	32960
401	40	4646	4092

Overall, the model achieves an accuracy of 0.50 and a macro F1-score of 0.41. The dataset is highly imbalanced. The majority class (000) accounts for approximately 57% of samples, meaning a naive classifier that always predicts this class would achieve higher accuracy (0.57). This highlights that accuracy alone is not an appropriate metric in this setting, and motivates the use of macro F1-score, which more fairly reflects performance across all classes.

The confusion matrix reveals several important patterns. First, the model performs best on the normal (000) class, with a relatively large number of correct predictions. However, there is substantial confusion between normal and mild dysfunction (400), suggesting that the model struggles to distinguish subtle reductions in cardiac function. This aligns with clinical reality, where mild dysfunction can be difficult to identify even for human observers.

Second, there is strong confusion between mild (400) and severe (401) dysfunction, with a large number of mild cases being predicted as severe and vice versa. This indicates that the embedding space does not cleanly separate different levels of dysfunction severity. Instead, it appears that the model captures a more general notion of "abnormal vs normal" rather than fine-grained gradations.

Taken together, these results suggest that the pretrained model retains some clinically meaningful signal in a zero-shot setting, particularly in distinguishing normal from abnormal function. However, it struggles with finer distinctions in severity. These findings are consistent with observations in

the image embedding space, where a continuous gradient from severe to normal LVEF function is present, but with substantial overlap between the normal to mild and mild to severe classes.

4.2.2 Device Classification

Device classification provides a complementary zero-shot evaluation. While LVEF classification tests whether the model captures functional information, device detection tests whether the pretrained image representation contains more localized visual signals associated with implanted intracardiac devices.

Table 2: Zero-shot device detection performance. The number of positive and negative samples reflects the ground truth labels derived from diagnostic codes. (Pos: Positive, Neg: Negative, Prev: Prevalence, AP: Average Precision)

Task	**Pos (GT)**	**Neg (GT)**	**Prev.**	**ROC AUC**	**AP**
Pacemaker	3,353	321,484	0.0103	0.851	0.055
Impella	548	324,289	0.0017	0.905	0.026
TAVR	400	324,437	0.0012	0.875	0.021
MitraClip	299	324,538	0.0009	0.927	0.104
Any device	4,735	320,102	0.0146	0.731	0.039

The results in Table 2 show that the pretrained model captures a meaningful signal in a zero-shot setting. Across the individual device tasks, ROC AUC values are comparable to the results in the original paper, ranging from 0.85 to 0.93. This indicates that scans containing a device generally receive higher similarity scores with the prompt than scans without that device. In other words, the pretrained image encoder appears to contain visual features that are useful for identifying device-related findings, even though no task-specific classifier was trained on the LUMC data.

The *any-device* task was constructed as an aggregate zero-shot concept. Instead of using a single prompt, the text embedding was formed by combining the prompt embeddings for the individual device categories. Each device prompt was encoded separately, the resulting text embeddings were averaged, and the combined vector was then used as the text representation for the broader concept of "any device." This demonstrates one of the advantages of a vision-language model: new clinical concepts can be defined by composing text prompts, without training a new classifier from scratch.

Performance on the aggregated *any-device* task is lower than on most individual device tasks. This is likely because the task combines device types with very different visual appearances.

Pacemaker or ICD leads, TAVR valves, Impella devices, and MitraClips do not correspond to a single consistent visual pattern. Averaging their prompt embeddings creates a broad device concept, but it also makes the class less visually coherent.

In contrast to the relatively high ROC AUC values, average precision (AP) remains low for all tasks. This is an important clinical limitation. The low AP is partly explained by the extreme class imbalance: device-positive cases are rare, with prevalence below 1.5% for all device categories and below 0.2% for several individual devices. In such settings, even a model that ranks positives above negatives reasonably well can still produce many false positives among its highest-scoring predictions. Therefore, a high ROC AUC does not imply that the model would be reliable as a standalone clinical detection tool.

Overall, the device classification results provide evidence that the pretrained EchoCLIP image representation transfers to the LUMC dataset. At the same time, the low AP values show that strong ranking performance is not sufficient for clinical deployment, especially when there are significant class imbalances.

4.3 Retrieval Evaluation

Retrieval is evaluated on a test set of 21,484 paired scan-report examples. It is evaluated in both directions. In image-to-report retrieval, each image embedding is used as a query and all candidate report embeddings in the test set are ranked by cosine similarity. The correct item is the report associated with that image's study. In report-to-image retrieval, the direction is reversed: each report embedding is used as a query and all candidate image embeddings are ranked by similarity. A successful retrieval result places the true paired item near the top of the ranked list.

Table 3: Baseline cross-modal retrieval performance of the pretrained EchoCLIP model on the LUMC retrieval test set of 21,484 paired scan-report examples. Lower MCMRR indicates better retrieval, while higher Recall@10 and normalized MCMRR indicate better retrieval.

Model	MCMRR	I→R R@10	R→I R@10	Norm. MCMRR
Baseline	8849.4	0.0001	0.0002	0.4119

Overall, retrieval performance is limited (Table 3). The baseline model obtains an MCMRR of 8849.4 on a retrieval set with 21,484 candidates, with Recall@10 values of 0.0001 for image-to-report

retrieval and 0.0002 for report-to-image retrieval. This means that the correct paired report or image is almost never retrieved among the top 10 candidates. Although the normalized MCMRR value of 0.4119 indicates performance above a random baseline, the very low Recall@10 values and high MCMRR show that the pretrained model does not provide reliable scan-report matching on the LUMC data.

This contrasts with the zero-shot classification results, where meaningful clinical signal is present. These results, alongside the embedding space analysis, suggest that the pretrained image encoder transfers better than the text encoder to our dataset.

In relation to the first research question, the baseline experiments therefore show partial transfer. The pretrained model retains clinically meaningful visual structure, as shown by the image embedding space and the device-detection results, but the alignment between LUMC scans and reports is weak under retrieval evaluation and the text embedding space shows little separation. This distinction motivates the second research question: whether targeted finetuning can improve alignment with the local data while preserving the useful structure already present in the pretrained representation.

5 Finetuning

EchoCLIP has separate image and text encoders, which makes it possible to adapt the two modalities independently. This is important because the domain shift between the original EchoCLIP data and the LUMC data may affect the two towers differently. Separate finetuning experiments therefore allow us to test whether the main mismatch lies in the text representation, the image representation, or the alignment between the two.

The finetuning experiments are designed conservatively. Rather than retraining the full model, we perform an ablation study where progressively larger parts of each encoder are unfrozen while the remaining parameters are kept fixed. This allows the effect of increasing trainable capacity to be evaluated. This is especially important in the LUMC setting because the reports are brief, which may make the scan-report supervision noisier than in the original EchoCLIP pretraining data.

For the ablation studies, each finetuning direction is selected using the metric most directly tied to the intended effect of that finetuning. Text-encoder finetuning is primarily evaluated using

retrieval metrics, because its purpose is to improve the correspondence between LUMC report embeddings and the images. Image-encoder finetuning is primarily evaluated using zero-shot LVEF classification because it directly changes the visual representation that this task depends upon. Both will also be evaluated through embedding space analysis. These metrics are used for the ablation analysis, but the final selected models are evaluated across the full set of tasks in Table 4.

5.1 Text-encoder Finetuning

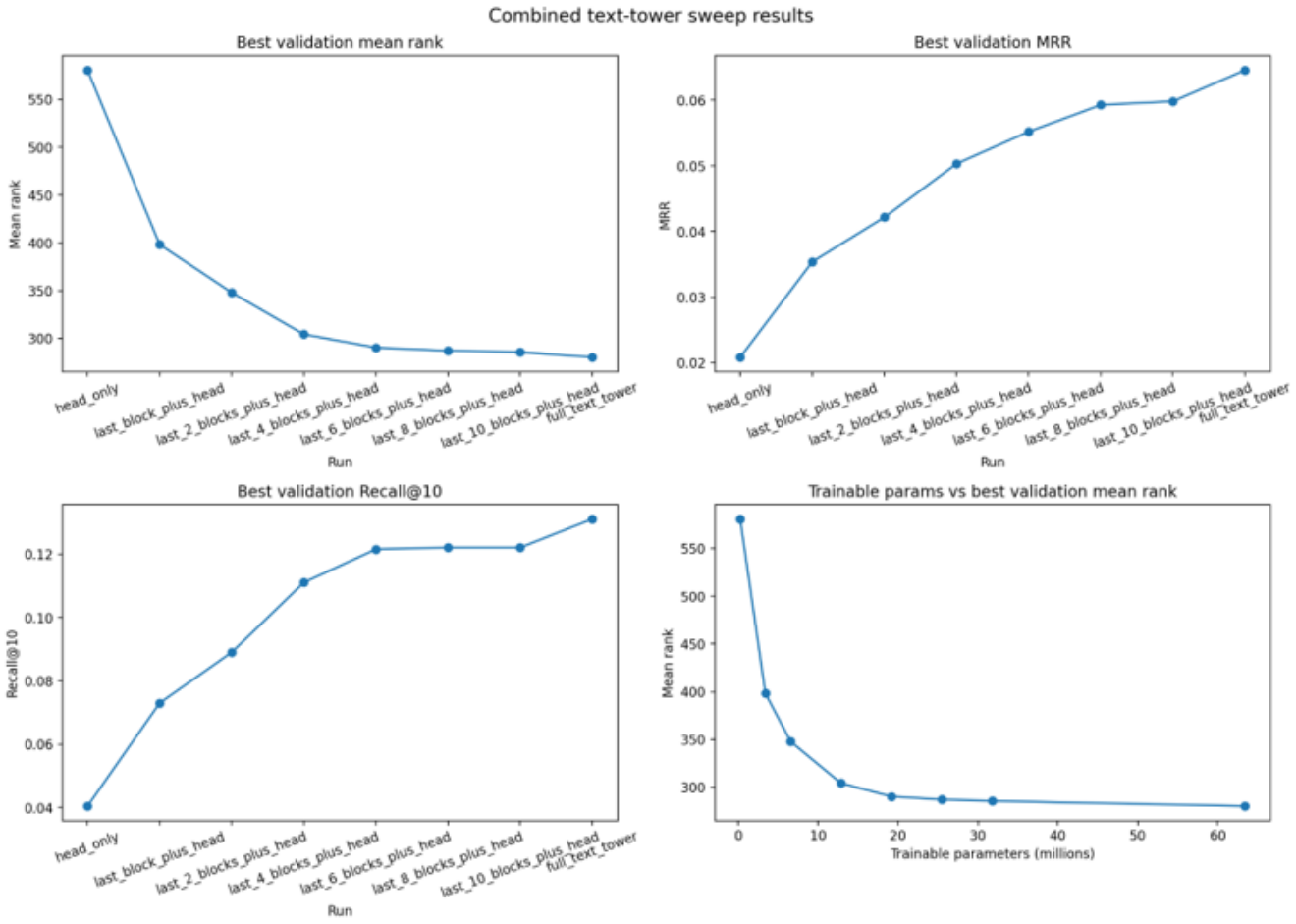


Figure 8: Text encoder finetuning ablation results. Each subplot reports a validation retrieval metric for a different unfreezing configuration of the text encoder: mean rank, MRR, Recall@10, and validation mean rank as a function of the number of trainable parameters. The X axis of each plot is in order of least to most unfrozen layers/parameters.

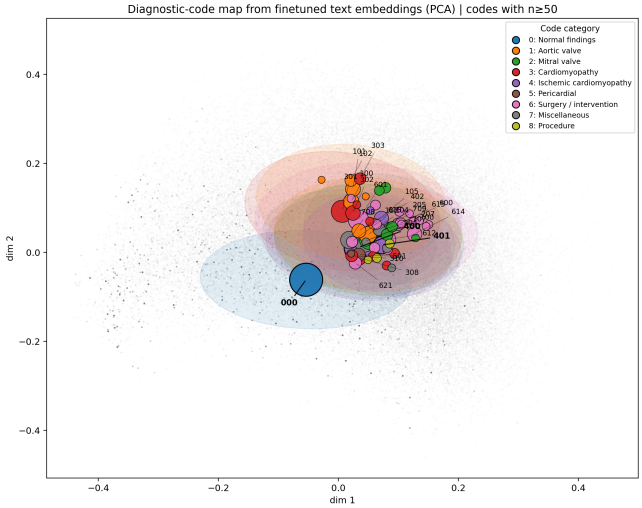
The text encoder ablation shows a clear relationship between the amount of adaptation and validation retrieval performance. When only the projection head is trained, retrieval performance

improves only modestly. As progressively more Transformer blocks are unfrozen, validation mean rank decreases, while MRR and Recall@10 increase. This is exactly the pattern we would hope to see from controlled finetuning: giving the text encoder more trainable capacity allows it to better adapt to the LUMC report distribution, while still preserving the general structure learned during pretraining.

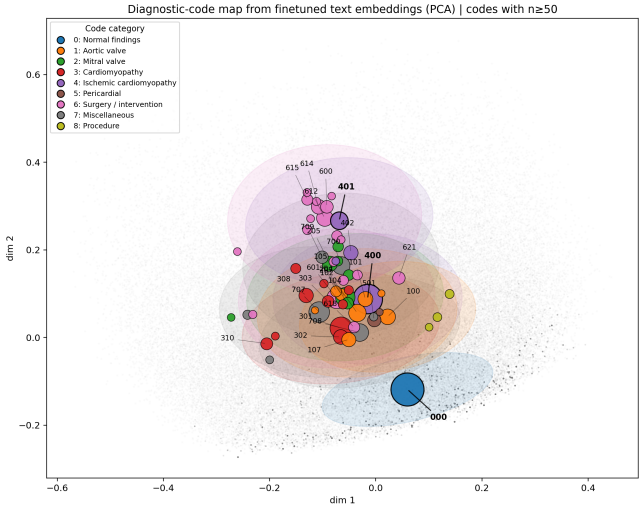
The improvement is also gradual across the sweep rather than erratic. This is important because it suggests that the gains are not simply caused by overfitting to a small subset of training pairs. If the extra trainable parameters were mainly causing memorization, we would expect the training loss to improve without a corresponding improvement in validation retrieval, or for validation performance to become unstable as model capacity increases. Instead, all three validation metrics move in the expected direction as more of the text tower is unfrozen.

The largest improvements occur when moving beyond the projection head and final one or two blocks, suggesting that the mismatch between EchoCLIP and the LUMC dataset is not limited to the final embedding projection. More of the text encoder needs to adapt in order to represent LUMC report phrasing effectively. At the same time, the gains begin to flatten as the number of unfrozen layers increases, which suggests diminishing returns from unfreezing the full text tower.

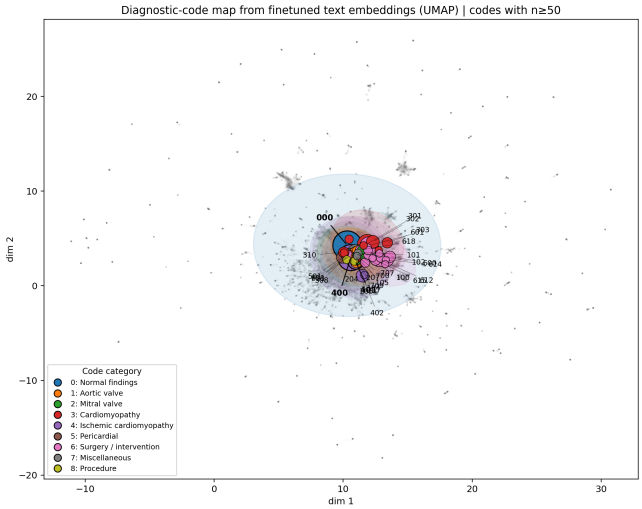
Figure 9 demonstrates that this improvement is also visible in the text embedding space. In both projections, the baseline text embeddings are concentrated in a dense central region, with diagnostic-code centroids lying close together and showing substantial overlap. After finetuning, the embedding space becomes more spread out, and some diagnostic-code groups show clearer relative separation. The separation is still limited, and the diagnostic-code groups remain strongly overlapping. However, the change in structure is consistent with the retrieval results: finetuning appears to make the LUMC report embeddings more informative and better adapted to the local report distribution, even though the resulting text space remains noisy and only partially organized by diagnostic category.



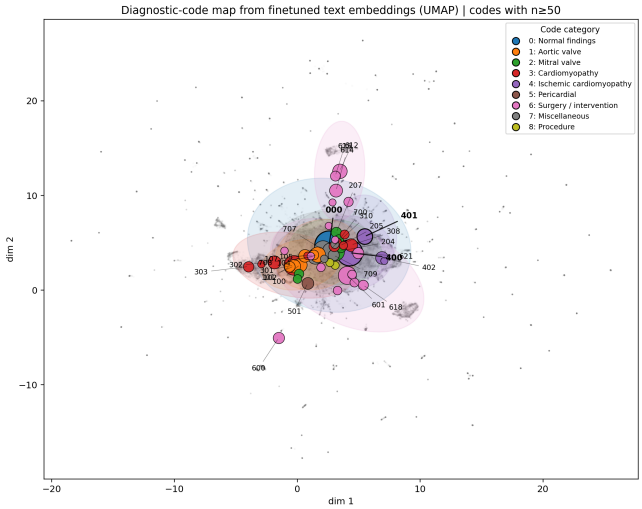
(a) Baseline PCA



(b) Post-finetuning PCA



(c) Baseline UMAP



(d) Post-finetuning UMAP

Figure 9: Comparison of diagnostic-code structure in the text embedding space before and after text-encoder finetuning. The top row shows PCA projections and the bottom row shows UMAP projections. Background points show individual report embeddings, while labeled markers indicate diagnostic-code centroids.

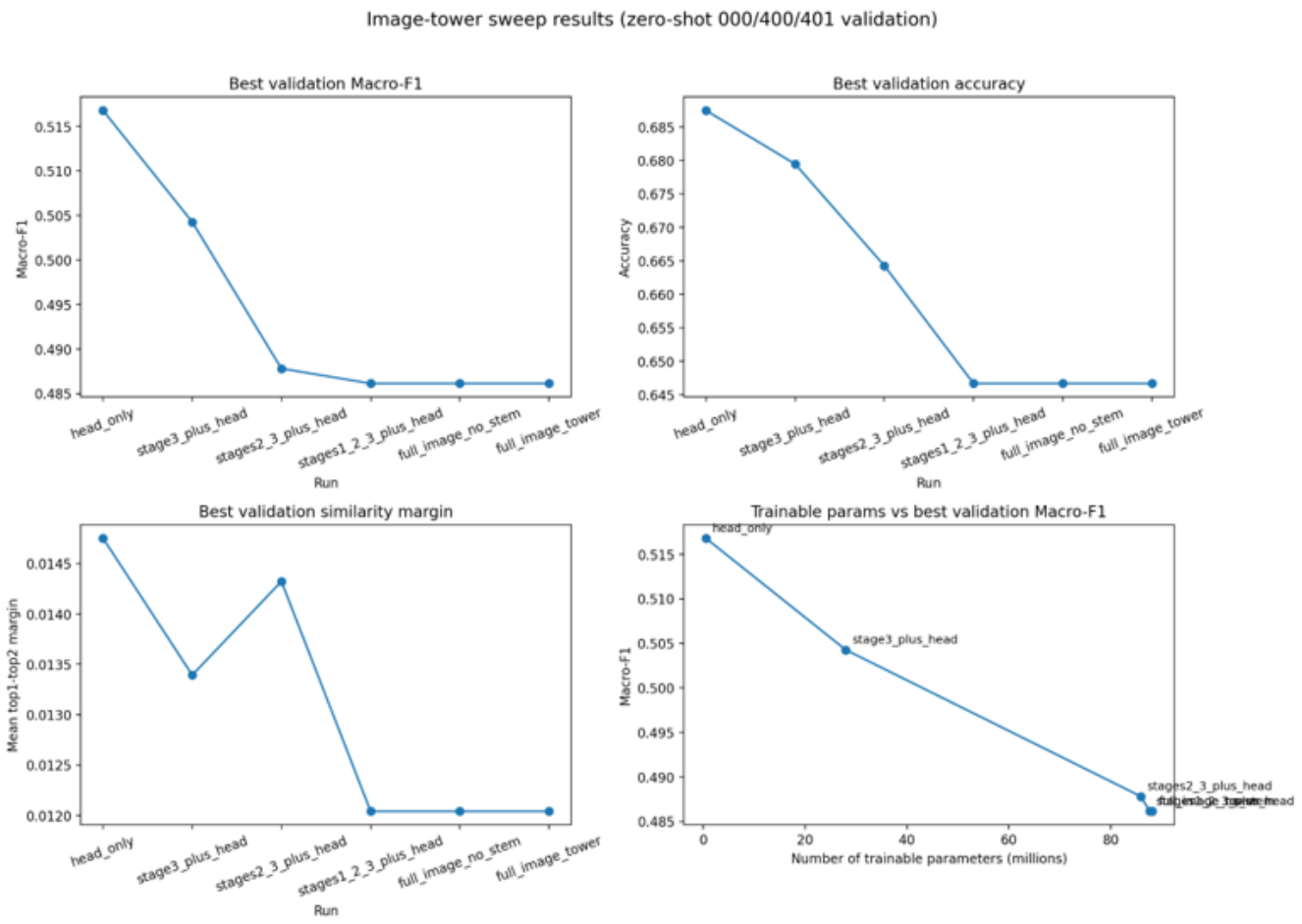


Figure 10: Image encoder finetuning ablation results on the zero-shot LVEF classification task. Each subplot reports a validation metric for a different image encoder unfreezing configuration: Macro-F1, accuracy, similarity margin, and validation Macro-F1 as a function of the number of trainable parameters.

5.2 Image-encoder Finetuning

As demonstrated by Figure 10, the image encoder ablation shows the opposite pattern from the text encoder sweep. The best zero-shot LVEF performance is obtained when only the image head is unfrozen, while progressively unfreezing larger parts of the image tower reduces both Macro-F1 and accuracy. This suggests that increasing the number of trainable image parameters does not improve the clinical usefulness of the image embeddings for this task.

The similarity margin shows a similar overall trend. Although there is some variation between intermediate configurations, the margin is highest when only the image head is updated and lowest once most or all of the image tower is unfrozen. Since the margin measures the separation between

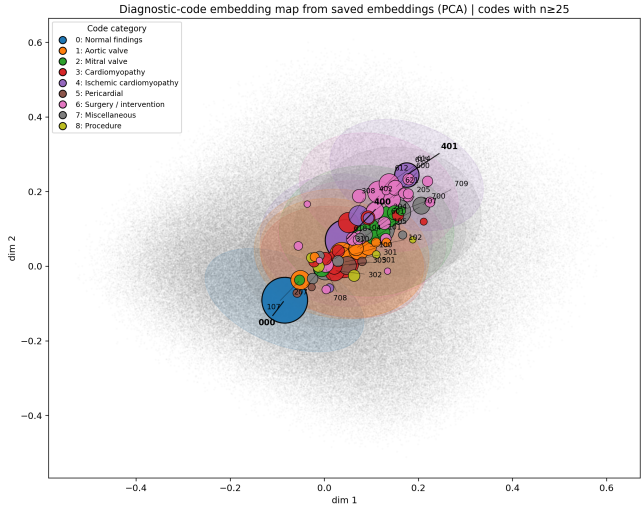
the top two class-prompt similarities, this suggests that larger image-side updates do not make the model more confident or better separated in the zero-shot LVEF setting.

An analysis of the embedding space after image-encoder finetuning provides additional, but qualitative, support for this interpretation. As shown in Figure 11d, the post-finetuning UMAP projection appears more compressed than the baseline projection. Diagnostic-code centroids are drawn closer together, and the grey background points, which represent individual scan embeddings, form a denser structure. This suggests that local separation between diagnostic-code groups is reduced after finetuning.

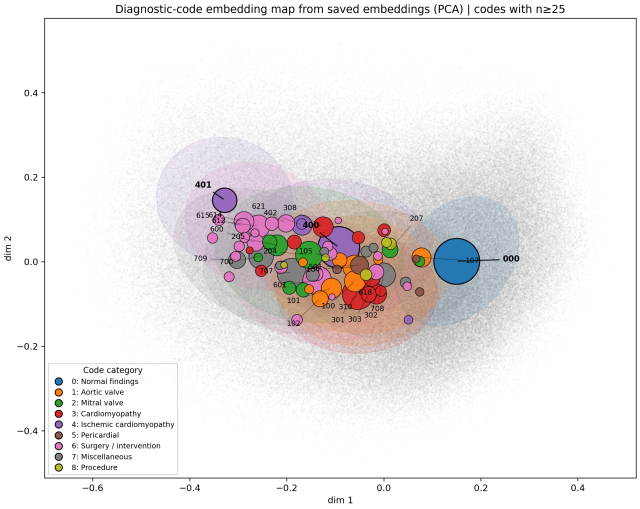
The PCA projection in Figure 11b shows this effect less clearly, since the broad arrangement of diagnostic-code centroids is still partly preserved. However, the post-finetuning PCA projection occupies a smaller coordinate range, suggesting reduced spread along the main directions of variation. Because PCA captures global linear structure while UMAP emphasizes local neighborhood relationships, the two projections together suggest that image-side finetuning does not completely remove the coarse organization of the embedding space, but may reduce the finer separation between clinically meaningful groups.

This visual pattern is consistent with the quantitative results from the image-encoder ablation. As more of the image tower is unfrozen, the similarity margin decreases, indicating that the model becomes less able to clearly separate the most likely LVEF class from the next-best alternative. At the same time, zero-shot LVEF Macro-F1 and accuracy decline, showing that this reduced separation corresponds to poorer distinction between normal, mild, and severe dysfunction. Together, these results suggest that image-side finetuning does not adapt the visual representation to LUMC data, but may instead compress or blur the structure of the embedding space. One likely explanation is that the contrastive scan-report supervision is too noisy to provide a reliable training signal for changing visual features. This is discussed further in Section 7.2.

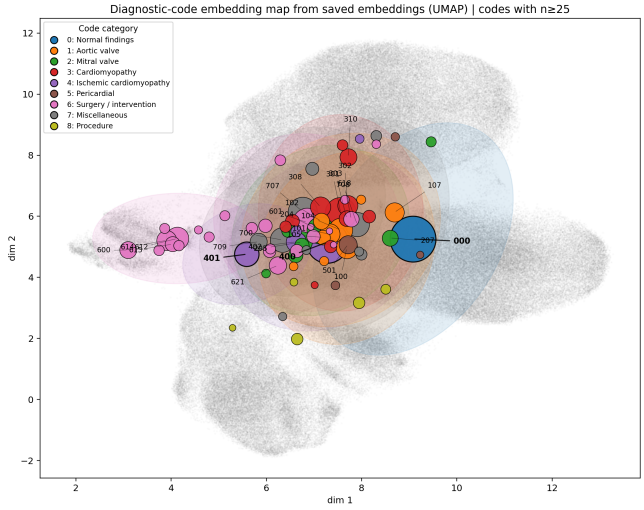
In relation to the second research question, the finetuning experiments suggest that adaptation is more effective on the text side than on the image side for our dataset. Text-encoder finetuning improves retrieval alignment and creates more separation and diagnostic code structure in the embedding space. Conversely, going beyond very limited image-encoder finetuning reduces zero-shot LVEF performance and compresses the embedding space.



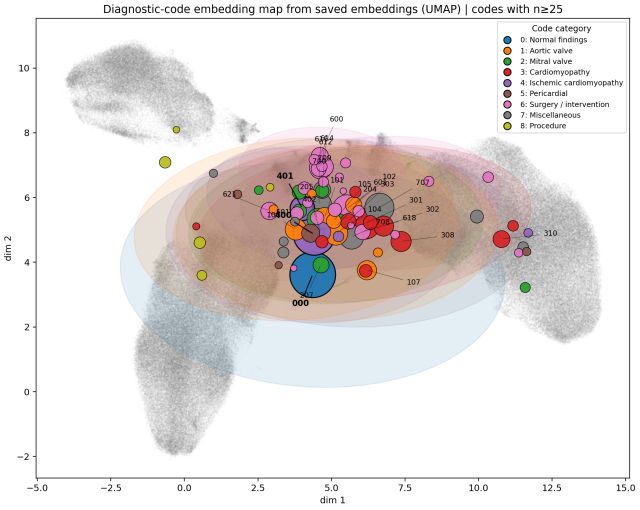
(a) Baseline PCA



(b) Post-finetuning PCA



(c) Baseline UMAP



(d) Post-finetuning UMAP

Figure 11: Comparison of diagnostic-code structure in the image embedding space before and after image-encoder finetuning. The top row shows PCA projections and the bottom row shows UMAP projections. Background points show individual scan embeddings, while labeled markers indicate diagnostic-code centroids for codes with at least 25 samples.

6 Multi-view Integration

The final experiment addresses the third research question: whether incorporating additional echocardiographic views provides useful information beyond the original A4C-centered EchoCLIP setup. As described in Section 3.6, A4C, A2C, and PLAX views are evaluated individually and in late-fusion combinations using zero-shot LVEF classification.

6.1 Multi-view Results

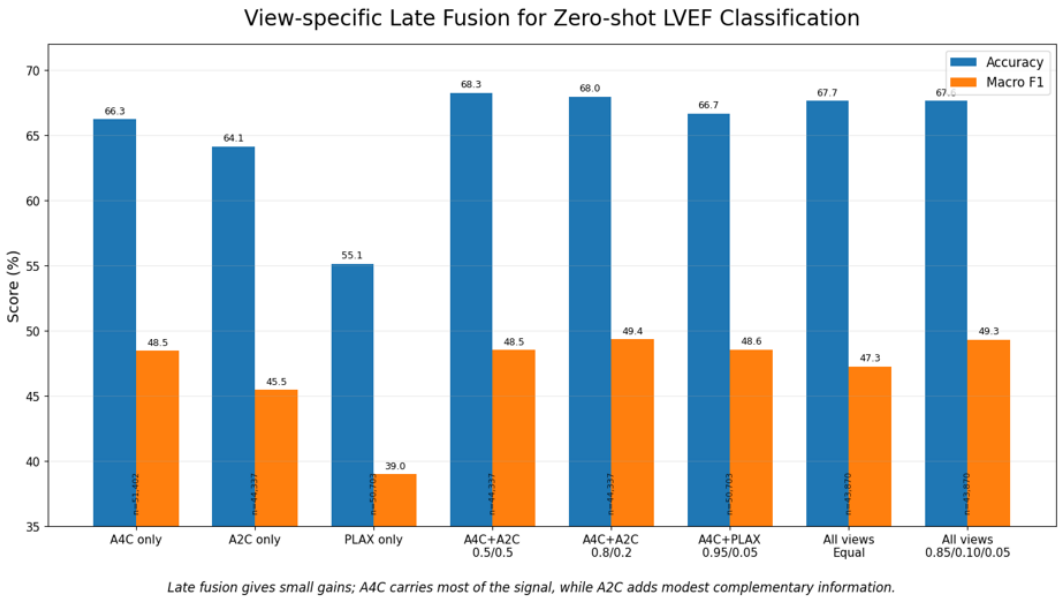


Figure 12: View-specific late fusion results for zero-shot LVEF classification using A4C, A2C, and PLAX views. Bars show accuracy and Macro-F1. Sample sizes differ between configurations because each setting requires the relevant views to be available: A4C only (n = 51,402), A2C only and A4C+A2C (n = 44,337), PLAX only and A4C+PLAX (n = 50,703), and all-view fusion (n = 43,870). Sample sizes refer to the number of valid labeled examples (400, 401, 000) not the total number of scan-report pairs in the dataset.

The multi-view results are shown in Figure 12. The main comparison is between A4C-only performance, single-view performance for A2C and PLAX, and late-fusion combinations of multiple views. Sample sizes differ between configurations because each setting requires the relevant views to be available for the same labeled examples.

Across the fusion configurations, performance remains broadly similar to the A4C-only baseline. The best multi-view settings produce only small changes in Macro-F1, indicating that simple

late fusion does not substantially change the model's ability to distinguish LVEF function. The single-view results follow the expected pattern. A4C achieves the strongest single-view performance. A2C performs slightly worse, but remains close to A4C, which is consistent with both views being acquired from an apical perspective and containing information about left ventricular function. PLAX performs substantially worse, likely because it differs more strongly from A4C in both appearance and anatomical content.

Overall, the multi-view experiment provides an inconclusive answer to the third research question. The results should not be interpreted as strong evidence that additional views are unhelpful. Rather, they show that the specific multi-view strategy used here does not provide a reliable improvement under the conditions tested. The experiment depends on image-side adaptation, which was already shown to be unstable in the finetuning experiments, making it difficult to determine whether the multi-view results reflect the value of A2C and PLAX or the limitations of the adapted image representation.

7 Discussion

7.1 Summary of Research Questions

The central theme of this thesis is how well a foundation model trained in one institutional setting transfers to another, and what can be learned from this case about the broader problem of cross-institutional transfer in medical vision-language models. This is addressed through three research questions: whether the pretrained EchoCLIP model transfers to LUMC data, what type of finetuning is needed to adapt it, and whether additional echocardiographic views provide useful information beyond the original A4C-centered setup. The experiments provide a mixed but coherent answer. EchoCLIP does transfer meaningful clinical structure to the LUMC dataset, but this transfer is incomplete and depends strongly on the type of task being evaluated.

For the first research question, the baseline experiments show that the pretrained model retains clinically useful visual information on LUMC data. This is most visible in the image embedding space analysis and in the zero-shot device-detection results, where the model achieves relatively strong ROC AUC scores comparable to the EchoCLIP paper despite not being trained on LUMC images. The zero-shot LVEF task also shows some clinically meaningful signal, especially in

separating normal from abnormal function, but the model struggles with finer severity distinctions. In contrast, cross-modal retrieval performance is very weak, indicating that the pretrained image and text embeddings are not well aligned for individual LUMC scan-report pairs, and the text embedding space is highly compressed and shows little meaningful structure. The answer to the first research question is therefore that EchoCLIP generalizes partially: the image representation transfers better than the text representation when applied to our dataset.

For the second research question, the finetuning experiments suggest that, for our dataset, adaptation is most useful on the text side. Text-encoder finetuning improves retrieval performance in a stable and gradual way as more of the text tower is unfrozen. This indicates that the LUMC report embeddings can be adapted toward the pretrained image space. Image-encoder finetuning shows the opposite pattern. Updating larger parts of the image tower reduces zero-shot LVEF performance and compresses the diagnostic-code structure in the embedding space. This suggests that the pretrained visual representation is useful, but that changing it using the available scan-report supervision can damage embedding space structure. The answer to the second research question is therefore that local adaptation is needed, but should be targeted primarily at the text encoder rather than the image encoder when there is poor image-text alignment.

For the third research question, the multi-view experiments show that adding A2C and PLAX views does not substantially improve performance when using late fusion. A4C remains the strongest single view, A2C performs relatively close to A4C, and PLAX performs worse, consistent with its greater visual and anatomical difference from the A4C view used in the original EchoCLIP setup. The late-fusion combinations produce only limited changes in Macro-F1, suggesting that additional views are not automatically beneficial unless the model can represent and combine them more effectively. The answer to the third research question is therefore that multi-view information may be useful in principle, but because we cannot effectively finetune the image encoder it is difficult to make any conclusive statements about this approach.

Taken together, these findings suggest that the central issue in this project is the alignment between LUMC scans and LUMC reports. The image encoder appears to retain transferable clinical structure, while the text supervision available in the LUMC reports is shorter, less standardized, and less directly aligned with individual scans. This interpretation connects the three sets of experiments: baseline transfer is strongest for visual tasks, text-side finetuning improves alignment,

text-reliant image-side finetuning is unstable, and multi-view integration provides limited benefit when the underlying image-text supervision remains weak.

7.2 Why Scan-report Alignment is Limited

The central issue throughout this project is that the LUMC reports are much less descriptive than the reports used in the original EchoCLIP study (Figure 3). This creates a problem for contrastive learning, because the method depends on a meaningful correspondence between each image and its paired text. In a CLIP-style objective, the matching scan-report pair is pulled closer together, while all other reports in the batch are treated as negatives and pushed farther away. This works best when the text descriptions contain enough detail and variety to distinguish one report from another. If many reports are short, generic, or clinically similar, then a non-matching report may still be a reasonable description of the scan. In that case, the model treats a semantically similar report as a negative, even though it should remain close in the embedding space.

This creates two related problems. First, the positive signal is weak: the paired report may not contain enough scan-specific information to tell the image encoder which visual features should be emphasized. This is reinforced by the fact that the reports describe the result of the full echocardiography study, while each training example contains only one scan or view. Second, the negative signal is noisy: many other reports may describe similar findings, such as reduced LV function or general hypokinesis, but are still treated as negatives by the contrastive loss. The model is therefore asked to pull the image toward a brief and incomplete description, while pushing it away from other reports that may be clinically similar.

This helps explain why image-side finetuning does not improve the visual representation. The contrastive objective may still separate individual scan-report pairs, but the resulting gradients are less likely to correspond to meaningful clinical structure. Instead of sharpening distinctions, image finetuning may blur or distort the structure already present in the pretrained image space. This is consistent with the reduced similarity margin, poorer zero-shot LVEF separation, and denser embedding space observed after image encoder adaptation. The model receives pressure to move embeddings, but the report text does not provide enough information to guide those movements in a useful direction.

This also explains why text encoder finetuning was more successful than image encoder finetuning.

When the image encoder is kept fixed, the model can adapt the report embeddings to a stable visual representation that already contains useful clinical structure. This directly addresses the weak retrieval performance of the baseline model, which depends on the alignment between image embeddings and report embeddings. In contrast, when the image encoder is updated, the visual space is changed using inadequately descriptive report supervision. As a result, image-side finetuning risks introducing noisy signals and disrupting the useful pretrained structure instead of improving it.

The results therefore suggest that the main limitation lies in the quality and specificity of the textual reports. When the supervision signal is weak or underspecified, increasing model complexity is unlikely to address the underlying problem. For example, adding multi-view features may provide more visual information, but it does not resolve the core issue that the model is still being trained against insufficiently descriptive reports. This is reflected in the rather lackluster multi-view results, where additional views did not notably improve performance. Meaningful improvement may therefore require improving the report supervision or reformulating the learning objective, rather than simply adding architectural complexity.

7.3 Reporting Practices

A broader takeaway from these results is that clinical reporting practices should be considered part of the data pipeline for medical vision-language models. In a contrastive vision-language model, the report is not merely metadata attached to the scan; it is the supervision signal that defines what visual information the model is encouraged to preserve. Differences in report structure, length, terminology, and clinical emphasis can therefore become a major source of domain shift.

This issue is natural because reports are written for clinical communication, not for machine learning. A concise report may be sufficient for a cardiologist who has access to the full study, patient history, measurements, and local clinical context. However, the same report may provide weak supervision for a model if important findings are abbreviated or omitted because they are obvious to the clinician. This creates a mismatch between what is clinically useful for human reporting and what is informative for representation learning.

This has several implications for future medical vision-language research. Institutions aiming to train or adapt such models should preserve as much report information as possible, including

structured measurements, anatomical descriptions, diagnostic codes, device information, and metadata linking reports to studies, views, and acquisition times. Explicit fields such as LVEF and device presence provide much stronger supervision than broad descriptive statements alone, while also making it easier to define evaluation labels and test model performance. However, this should not be interpreted as a call to increase the reporting burden on clinicians. The European Society of Radiology (ESR) notes that structured reporting is difficult to adopt when it disrupts the normal reporting workflow. If radiologists have to stop interpreting images to manually fill in rigid templates, structured reporting becomes less practical in routine clinical use [4].

For this reason, the recommendation is not simply that clinicians should write more structured reports, but that institutions should invest in reporting infrastructure that captures structured information with minimal disruption to clinical workflow. The ESR paper highlights several developments that support this direction, including interoperable data standards such as FHIR, integration of AI results into reporting systems, and the use of natural-language processing or large language models to extract structured fields from dictated or free-text reports [4].

As medical AI advances, hospitals will increasingly need to make use of models and methods developed outside their own institution. This requires clinical information to be stored in a form that supports reuse, adaptation, and validation. Standardized and structured reporting systems therefore play an important role in improving the transferability and practical adoption of medical AI. They make routine clinical data more informative for model development while also reducing the need to rebuild or extensively adapt models for each hospital separately. Structured reporting will not remove all sources of domain shift, since things like imaging protocols and patient populations will still differ between institutions. However, it can make the text supervision more explicit, reusable, and comparable across settings. For this reason, institutions interested in benefiting from medical machine learning should treat structured reporting infrastructure as a core part of their long-term AI strategy, not merely as a documentation tool.

7.4 Future Work

The most immediate direction for future work is to repeat the experiments using newer and more informative LUMC reports. LUMC has recently implemented a more structured reporting system, in which reports are generated from standardized clinical information and measurements associated

with the patient record. These reports were not available for the present project because the data is recent and access to sensitive clinical information requires additional governance. However, evaluating EchoCLIP on these newer reports would be highly informative. If retrieval and finetuning performance improve substantially, this would support the central interpretation of this thesis.

Once the report supervision has been improved, more advanced modeling strategies become more meaningful to explore. One important direction is study-level modeling. In the current setup, individual scans or views are paired with reports that describe the full echocardiography examination. A more appropriate approach would be to represent each study as a collection of scans or videos and learn a single study-level embedding. Multiple instance learning or attention-based pooling could be used to combine information across views while allowing the model to learn which scans are most relevant for the report.

Another direction is to incorporate temporal information more directly. EchoCLIP uses frame-based image embeddings, but echocardiography is inherently dynamic. Important clinical information, especially related to ventricular function, valve motion, and wall motion abnormalities, depends on movement across the cardiac cycle. Future work could therefore evaluate video encoders or temporal aggregation methods that represent full clips rather than individual frames or mean-pooled frame embeddings.

Finally, future researchers should consider alternative contrastive learning objectives that are better suited to clinical data. Standard CLIP-style training treats all non-matching reports in a batch as negatives, even when many may be clinically similar to the true report. This is especially problematic when reports are brief or repetitive. Multi-positive contrastive learning, soft labels based on diagnostic-code overlap, or objectives that group clinically similar reports could reduce the false-negative problem.

8 Conclusion

This thesis investigated the transferability of EchoCLIP to echocardiography data from LUMC. The results show that the pretrained model transfers meaningful visual structure to this new institutional setting, as seen in the diagnostic-code organization of the image embedding space and in zero-shot device-detection performance. However, this transfer is incomplete: cross-modal

retrieval remains weak, indicating that individual LUMC scans and reports are not well aligned in the shared embedding space. Finetuning results suggest that adaptation is more effective on the text side than on the image side, with text-encoder finetuning improving retrieval while broader image-encoder finetuning reduces clinically meaningful structure. The multi-view experiments are inconclusive, but show that simply adding A2C and PLAX views through late fusion is not enough to overcome the underlying alignment problem. Overall, the findings suggest that successful transfer of medical vision-language models depends not only on model architecture and image data, but also on the structure and specificity of the clinical reports used as supervision. Reporting infrastructure should therefore be considered part of the foundation-model pipeline: structured, measurement-rich reports may make routine clinical datasets more useful for contrastive learning and cross-institutional model transfer.

Table 4: Comparison of baseline and finetuned EchoCLIP models across retrieval, zero-shot, and patient-similarity metrics. Green indicates the best value for a metric, and yellow indicates the second-best value. For MCMRR, normalized MCMRR, and different-patient similarity, lower is better; for all other metrics, higher is better. The EchoCLIP paper row is included only as an external reference (mainly for device AUC scores) and is not directly comparable, since it was evaluated on the original EchoCLIP dataset rather than the LUMC dataset.

	Retrieval				LVEF zero-shot		Device zero-shot AUC					Patient similarity		
Model	$MCMRR_{eq}$	I→R R@10	R→I R@10	Norm. MCMRR	Acc.	Macro-F1	Pacemaker	Impella	TAVR	MitraClip	Any-device	AUC	Same sim.	Diff sim.
Baseline	8849.4	0.0001	0.0002	0.4119	0.6444	0.4808	0.8509	0.9053	**0.8753**	0.9273	0.7317	**0.7844**	0.8178	**0.7337**
Text-tuned	**1887.6**	**0.0105**	**0.0078**	**0.0878**	**0.7094**	**0.5607**	**0.8770**	**0.9309**	0.8682	**0.9360**	0.7481	**0.7844**	0.8178	**0.7337**
Image-tuned	8444.0	0.0002	0.0002	0.3930	0.6880	0.4934	0.8695	0.9055	0.8720	0.9307	**0.7544**	0.7823	**0.8346**	0.7588
EchoCLIP paper	*571.3*	N/A	N/A	N/A	N/A	N/A	*0.84*	N/A	*0.92*	*0.97*	N/A	N/A	N/A	N/A

References

[1] Rishi Bommasani et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[2] Mehdi Cherti et al. Reproducible scaling laws for contrastive language-image learning. *arXiv preprint arXiv:2212.07143*, 2022.

[3] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30:1481–1488, 2024.

[4] European Society of Radiology (ESR). Esr paper on structured reporting in radiology—update 2023. *Insights into Imaging*, 14:199, 2023.

[5] Amirata Ghorbani et al. Deep learning interpretation of echocardiograms. *NPJ Digital Medicine*, 3:10, 2020.

[6] Gregory Holste et al. Severe aortic stenosis detection by deep learning applied to echocardiography. *European Heart Journal*, 44:4592–4604, 2023.

[7] Athira J. Jacob, Indraneel Borgohain, Teodora Chitiboi, Puneet Sharma, Dorin Comaniciu, and Daniel Rueckert. Towards a cardiovascular magnetic resonance foundation model for multi-task cardiac image analysis. *Journal of Cardiovascular Magnetic Resonance*, 27:101967, 2025.

[8] David Ouyang, Bing He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P Langlotz, Paul A Heidenreich, Robert A Harrington, David H Liang, and Euan A Ashley. Video-based ai for beat-to-beat assessment of cardiac function. *Nature*, 580:252–256, 2020.

[9] Jasper Tromp, Paul J Seekings, Chih-Cheng Hung, et al. Automated interpretation of systolic and diastolic function on the echocardiogram: a multicohort study. *Lancet Digital Health*, 4(1):e46–e54, 2022.

[10] Jiancheng Zhang, Sravya Gajjala, Pranav Agrawal, Geoffrey H Tison, Lee A Hallock, Lauren Beussink-Nelson, Michelle H Lassen, Eric Fan, Mustafa A Aras, Charles Jordan, et al. Fully

automated echocardiogram interpretation in clinical practice. *Circulation*, 138(16):1623–1635, 2018.