Fake It Till You Make It: Using Synthetic Data and Domain Knowledge for Improved Text-Based Learning for LGE Detection

Athira J Jacob [1,2], Puneet Sharma [1], Daniel Rueckert [2,3]

[1]Digital Technology and Innovation, Siemens Healthineers, Princeton, NJ, USA
[2] AI in Healthcare and Medicine, Klinikum rechts der Isar, Technical University of Munich, Germany
[3] Department of Computing, Imperial College London, UK

Abstract

Detection of hyperenhancement from cardiac LGE MRI images is a complex task requiring significant clinical expertise. Although deep learning-based models have shown promising results for the task, they require large amounts of data with fine-grained annotations. Clinical reports generated for cardiac MR studies contain rich, clinically relevant information, including the location, extent and etiology of any scars present. Although recently developed CLIP-based training enables pretraining models with image-text pairs, it requires large amounts of data and further finetuning strategies on downstream tasks. In this study, we use various strategies rooted in domain knowledge to train a model for LGE detection solely using text from clinical reports, on a relatively small clinical cohort of 965 patients. We improve performance through the use of synthetic data augmentation, by systematically creating scar images and associated text. In addition, we standardize the orientation of the images in an anatomy-informed way to enable better alignment of spatial and text features. We also use a captioning loss to enable fine-grained supervision and explore the effect of pretraining of the vision encoder on performance. Finally, ablation studies are carried out to elucidate the contributions of each design component to the overall performance of the model.

Introduction

Late Gadolinium Enhancement (LGE) imaging—also referred to as Delayed Enhancement Imaging—plays a critical role in assessing myocardial viability. By highlighting affected areas with increased contrast uptake, it allows non-invasive detection and assessment of myocardial infarction, as well as ischemic and non-ischemic cardiomyopathies and other cardiac pathologies (Jenista et al. 2023). However, detecting areas of hyperenhancement from LGE images (henceforth referred to as LGE or scar detection) is a challenging task. Enhancement can be subtle and is often influenced by spatial and temporal variations across different scanners, sequences, and study protocols. Moreover, image noise, artifacts, and varying LGE patterns add to the difficulty. These complexities make developing robust, generalizable automated solutions for LGE detection particularly challenging. Although deep learning (DL) approaches have achieved impressive performance in a range of medical imaging applications, they are heavily dependent on access to large, good quality, annotated datasets for increased accuracy and generalization. The annotation of LGE data, which requires precise delineation of both myocardial and enhancement regions, is especially complex due to its variability and requires substantial clinical expertise. These demands pose a limitation on large-scale DL training efforts in this area.

Meanwhile, clinical reports present a source of valuable information about LGE assessment. The reports are created from the cardiac magnetic resonance (CMR) study and contain clinician-written notes about the LGE, including location, extent, etiology, and other impressions. While these could be manually converted to binary labels to train a DL model, that is impractical for large amounts of data. Moreover, translating some of the information into discrete labels could potentially be a non-intuitive task. In such a situation, training directly with text offers an attractive alternative.

Recently, Contrastive Language Image Pre-training (CLIP) (Radford et al. 2021) has achieved notable success in incorporating text supervision into vision models, for a wide range of downstream tasks in natural image processing, such as classification (Zhou et al. 2022b), object detection(Lin and Gong 2023), and segmentation (Luo et al. 2023). CLIP's approach aligns image and corresponding textual description embeddings within a shared latent space, facilitating a unified representation. This training paradigm has also been applied within medical imaging (Zhao et al. 2023). However, CLIP models often require large datasets—often millions of image-text pairs—to learn robust associations. Collecting such data, especially in specialized fields like medical imaging, is challenging due to limited availability and high annotation costs. These data demands can limit accessibility. In addition, despite the impressive zero-shot or few-shot performance, fine-tuning with task-specific data is often required to reach state-of-the-art (SoTA) performance. Many studies explore adapting CLIP models to downstream tasks through prompt tuning and linear probing. However, these require further training stages for the model, and is dependent on the size of the fine-tuning dataset.

In this study, we train a model for LGE detection on a clinical cohort of 965 patients. We use text supervision from their corresponding clinical reports, without any further fine-tuning with discrete labels. We use domain knowledge to systematically augment the limited dataset with synthetic scar images and text. Moreover, we normalize the orientation of the images in an anatomy-informed way to facilitate correspondence between image and text features. Additionally, we add captioning loss to provide more granular supervision from the text. Furthermore, the model's encoder is pre-trained on a related task involving LGE images. We obtain an average balanced accuracy of 0.83 on the test set. Though the model is supervised with patient-level descriptions, it can produce slice-level predictions due to the training strategy, which aids in interpretability. We also conduct ablation studies to assess the impact of each design choice and explore the effect of alternative encoder initialization strategies on model performance.

Related Work

Vision-Language Training in the medical domain. Many studies have explored building specialized models in the medical domain (Zhang et al. 2022; Wang et al. 2022; Zhang et al. 2024). BiomedCLIP (Zhang et al. 2023b), a vision-language model (VLM) was trained on 15M image-text pairs from the biomedical domain. Adapting these models to downstream tasks has been explored in various ways: a) End to end fine-tuning (Zhang et al. 2022; Ikezogwo et al. 2024), b) Prompt tuning (Zhou et al. 2022c,a), c) Linear probing (Radford et al. 2021; Zhou et al. 2022c; Shakeri et al. 2024). These methods still require additional training stages on a labeled "support" dataset for the downstream task, and can underperform in very low data cases, especially with complex tasks. Recently, multi-modal language models (MLLMs) have emerged as generalist models for various tasks (Moor et al. 2023; Singhal et al. 2023; Li et al. 2024). These are generative models capable of generating and processing long-form text. For instance, Med-Flamingo (Moor et al. 2023) is an open-source, multimodal few-shot learner adapted to the medical domain. These models demonstrate impressive performance across a wide variety of tasks, such as visual question answering, classification, report generation, etc., on a wide range of modalities.

Synthetic Data Generation. Augmentation with synthetic data has the potential to mitigate the issue of limited domain data in medical imaging. However, generating high-quality synthetic medical images remains challenging. Clip-Medfake (Chen et al. 2024) leverages Stable Diffusion (Rombach et al. 2022) to generate synthetic images, which are then used to pretrain a CLIP model before fine-tuning on actual medical data. Latte-CLIP (Cao et al. 2024) uses MLLMs to generate descriptive text for domain-specific images, which are then used for CLIP model fine-tuning. Data synthesis in these methods is done by pre-trained models or models trained from a small portion of the original training data. They lack fine grained control over the synthesis process. CtrlSynth (Cao, Najibi, and Mehta 2024) utilizes pre-trained models to identify concepts within an image, systematically alter these attributes, and generate synthetic text with LLMs. This synthetic text is then converted into images through Stable Diffusion, creating a comprehensive synthetic dataset that supports model training on controlled variations.

LGE Detection. There is no consensus regarding the optimal method for LGE analyses. Commonly used manual and semi-automatic methods clinically include manual planimetry, the Full Width Half Maximum (FWHM) approach (Amado et al. 2004; Hsu et al. 2006), and n-std from remote myocardium. Comparative studies examining these methods (Flett et al. 2011; Heiberg et al. 2022) reveal significant variability in quantification results, highlighting issues with both reliability and reproducibility.

Recently, there have been many DL-based studies focused on automated scar detection from cardiac LGE images (Zhang 2021; Kim, Chung, and Choe 2024; Girum et al. 2021; Yang and Wang 2021). For instance, Kim et al. (Kim, Chung, and Choe 2024) leveraged segmental information to train models that identify the presence or absence of LGE by transforming the images into polar coordinates based on the left ventricular (LV) center and the right ventricular (RV) insertion points. Additionally, several studies have explored infarct segmentation on the publicly available EMIDEC dataset (Lalande et al. 2020) containing 150 patients (100 for training), achieving classification accuracies as high as 0.92 (Lalande et al. 2022) and a Dice coefficient of up to 0.71 for infarct segmentation (Zhang 2021). These models are trained on dense pixel-wise annotations of myocardial and scar tissue provided by clinical experts, thereby enhancing their ability to accurately segment scar regions.

Methodology

Preliminaries: CLIP-based training CLIP training framework consists of a vision encoder and a text encoder to extract features from pairs of images and text, respectively. Each encoder is followed by a set of projection layers to project the features into a common embedding space. More specifically, the input image x_{img} is encoded by the encoder E_{img} into feature vector $f_{img} \in \mathbb{R}^n$. A projection module P_{img} maps the features into the embedding $v \in \mathbb{R}^p$.

$$v = P_{img}(E_{img}(x_{img}))$$

Similarly, the text encoder E_{txt} encodes the input text into feature vector $f_{txt} \in \mathbb{R}^m$. A projection module P_{txt} maps the features into the embedding $t \in \mathbb{R}^p$.

$$t = P_{txt}(E_{txt}(x_{txt}))$$

The similarity score is calculated using dot product as

$$s = v_n \cdot t_n$$

where v_n, t_n represent L2 normalized vectors. Then, cross-entropy loss (CE) is used to maximize the similarity score within the same pair, and minimize the same across pairs (Radford et al. 2021).

Preliminaries: Related tasks on LGE images. Prior to this study, we trained DL networks to segment the myocardium and detect anterior and posterior RV Insertion Points on LGE MRI Images. Ground truth (GT) annotations for these tasks are relatively easy to create. The myocardium

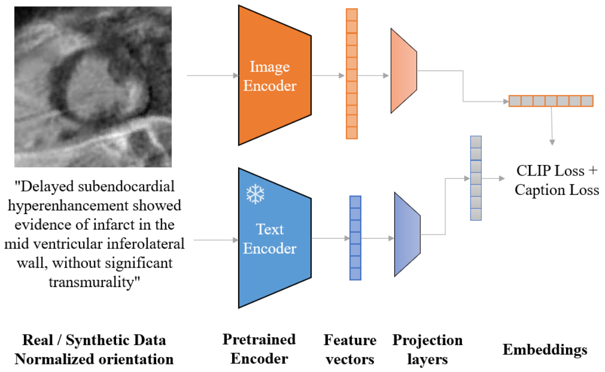
Image Encoder
"Delayed subendocardial hyperenhancement showed evidence of infarct in the mid ventricular inferolateral wall, without significant transmurality"
Text Encoder
CLIP Loss + Caption Loss
Real / Synthetic Data Normalized orientation
Pretrained Encoder
Feature vectors
Projection layers
Embeddings

Figure 1: Overview of the proposed model

segmentation network consisted of a UNet architecture, with DenseNet121 (Huang et al. 2017) encoder. It was trained on 7401 manually annotated images from 786 patients, from 3 centers. The model achieved an average Dice score of 0.88 on the test set of 131 patients (1197 images) from the 3 centers. The landmark detection model consisted of a UNet architecture, with ResNet18 (He et al. 2016) encoder. It was trained on 3478 manually annotated images from 377 patients, from a single center to predict heatmaps centered around the landmark points. It achieved an average error 3.1 and 3.3 mm for the anterior and inferior RV Insertion Points, respectively.

Proposed Method

We train a model to detect myocardial hyperenhancement from LGE images using relevant text from the clinical reports (Figure 1). Due to the limited dataset size and the long-tailed distribution of LGE etiologies, we use domain knowledge to systematically augment the training data with synthetic image-text pairs. During training, the image-text pairs are aligned using global CLIP loss and a local caption loss. The image encoder is initialized with the weights from the myocardial segmentation network described in the previous section. Each of these parts is explained in detail in the following sections. During inference, we query the model using the following text: `there is hyperenhancement in the myocardium` and `there is no hyperenhancement in the myocardium`, denoting the positive and negative LGE classes respectively.

Synthetic Data Generation

We add synthetic scars to real LGE images and create associated text descriptions (Figure 2a). This allows us to augment the limited data, and systematically cover a wide variety of scar distributions. The scar is applied only to images with no prior LGE (as determined from the clinical report). A Controller module randomly chooses from a set of scar parameters. Then, a synthetic scar is added to the image, and text is created with these parameters. This is done at every training iteration, with a probability of λ.

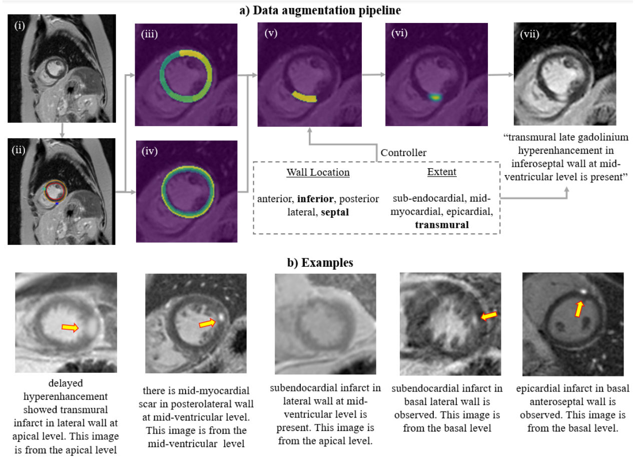
a) Data augmentation pipeline
(i)
(ii)
(iii)
(iv)
(v)
(vi)
(vii)
Controller
Wall Location
anterior, inferior, posterior lateral, septal
Extent
sub-endocardial, mid-myocardial, epicardial, transmural
"transmural late gadolinium hyperenhancement in inferoseptal wall at mid-ventricular level is present"
b) Examples
delayed hyperenhancement showed transmural infarct in lateral wall at apical level. This image is from the apical level
there is mid-myocardial scar in posterolateral wall at mid-ventricular level. This image is from the mid-ventricular level
subendocardial infarct in lateral wall at mid-ventricular level is present. This image is from the apical level.
subendocardial infarct in basal lateral wall is observed. This image is from the basal level
epicardial infarct in basal anteroseptal wall is observed. This image is from the basal level.

Figure 2: Synthetic Data Generation: a) Generation pipeline, b) Examples of synthetically generated images and corresponding captions. Text may refer to a region other than the one of the paired image (image 3).

Controller. Wall location is randomly chosen as one of [”anterior”, ”inferior”, ”posterior”], or [”lateral”, ”septal”], or a combination of the words chosen from the two sets (Eg: ”inferoseptal”). Scar extent is randomly chosen from [”sub-endocardial”, ”mid-myocardial”, ”epicardial”, ”transmural”]. ”Transmural” signifies the scar spanning over more than 50% of the myocardial thickness.

Image Generation. For every negative LGE image (Figure 2a.i) at a given slice location, synthetic scars are added as follows:

1. Myocardial mask and anterior RVIP are determined using the previously trained DL networks (Figure 2a.ii).
2. Anterior RVIP is used to divide the myocardium into AHA segments: four if apical layer, or six segments if basal or mid (Figure 2a.iii).
3. The myocardial mask is divided equally into 3 concentric sections to represent endocardial, mid-myocardial and epicardial layers (Figure 2a.iv).
4. Wall location (chosen by the controller), along with slice location, is translated into AHA segments through hard-coded values. For eg, inferoseptal on basal level denotes segment 3.
5. A scar candidate region mask is created using an intersection between the identified AHA segments (Step 4) and chosen scar extent (by the controller). This is the ”allowed” region for scar creation (Figure 2a.v).
6. Synthetic scar is created in the candidate region mask as a randomly placed, oriented and sized ellipse (Figure 2a.vi). A random pixel is chosen from the candidate region as the center of the synthetic scar. The radii of the ellipse are chosen randomly between set minimum and maximum values. The minimum and maximum values are determined as a fraction of the myocardial thickness at that point, depending on whether the scar is chosen to be transmural or within specific myocardial layers. The

created ellipse is smoothed with a Gaussian filter for a more natural and continuous appearance.

$$r_{min} = max(0.01, th * \rho_{min})$$
$$r_{max} = th * \rho_{max}$$
$$\mathbf{r} = rand(r_{min}, r_{max}, 2)$$
$$\alpha = rand(0, \pi)$$
$$\sigma = rand(0, 1) * s_1 + s_2$$

where, th represents myocardial thickness at that point, $0 < \rho_{min}, \rho_{max} <= 1$ are hyperparameters representing ratios relative to myocardial thickness, $\mathbf{r} \in \mathbb{R}^2$ are the radii of the major and minor axes of the ellipse, α represents the orientation of the major axis of the ellipse relative to the positive x-axis, and σ represents the standard deviation of the Gaussian kernel used for smoothing. The created scar M is min-max normalized to the range of $[0, 1]$.

7. The scar image M is then blended with the image I as,

$$I_{synth} = I * (1 - M) + \gamma * max(I) * M$$
$$\gamma = rand(b_1, b_2)$$

where γ controls the brightness of the scar, and is randomly chosen between preset minimum and maximum values b_1, b_2 (Figure 2a.vii).

Text Generation. Given the chosen scar parameters (slice location, wall location, wall extent) from the controller module, the associated text description is synthesized using preset templates such as the following:
"there is <extent> delayed enhancement in <location> wall. This image is from <slice location> level."
or variations of this, where the words within <> are replaced with chosen scar parameters. The slice location of the input image is appended to stay consistent with real clinical text and contextualize the spatial location of the slice for the model (further explained in the Section "Implementation Details: Text Encoder"). To add further variation to the text, the words "delayed enhancement", "delayed hyperenhancement", "late enhancement", "scar", "infarct" are used interchangeably.

Examples of images with synthetically generated scar and corresponding text are shown in Figure 2b.

Normalization of the LV orientation

Clinical descriptions of LGE use words that describe location relative to the orientation of the LV, defined by the RV insertion point. The orientation of the LV can vary across patients, and even across images within the same patient. While this could potentially be learnt implicitly from a large cohort of image-caption pairs, this could prove to be a difficult challenge in a dataset of limited size. To help the model better associate position descriptors with image features, we standardize the orientation of the LV using the anterior and inferior RVIPs (Figure 3). The RVIPs for each image are obtained from the landmark detection model described previously. Using the two insertion points, each image is rotated to position the line connecting them along the vertical axis of the image, with anterior RVIP being on the top.

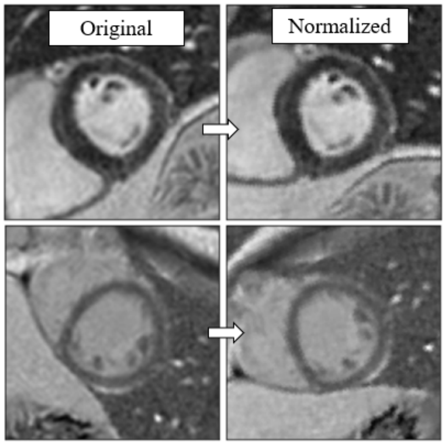

Figure 3: Anatomically informed normalization of LV orientation

Caption loss

CLIP loss aligns image and text embeddings on a global level. However, LGE descriptions from clinical reports are information-dense, with multiple words providing critical information about the location, extent and etiology of the scar. To encourage granular supervision on the level of the individual text tokens, we use a captioning loss similar to contrastive captioner models (Yu et al. 2022). A multi-modal decoder is applied to the text tokens, consisting of layers of multi-headed, self-attention layers, followed by cross-attention layers attending to features from the vision encoder. The final layer is the classification layer that predicts the distribution of the next token over the supported vocabulary set.

Task specific encoder

We initialize the vision encoder with the weights of LGE myocardium segmentation model trained as described previously. Segmenting LGE myocardium is a closely related task to LGE detection, and is hypothesized to aid convergence. We later study the effect of this design choice by conducting ablation studies with various other image encoders.

Experiments

Data

The data consists of 965 patients with cardiac MRI studies and clinical reports from a single center, of which 404 patients have reported LGE. The scans were performed on 1.5 T magnets (MAGNETOM Avanto, Siemens Healthcare, Erlangen, Germany) using a T1-weighted, phase sensitive inversion-recovery (PSIR), gradient-echo sequence, and were acquired 10 min after injection of a gadolinium-based contrast agent. The acquisition parameters are as follows, TR/TE: 2.4 /1.1 ms; flip angle: 50, slice thickness: 8 mm; in plane resolution between 1.5 × 1.5 6 mm^2 and 2.6 × 2.6 mm^2. The patients were divided in train, validation and testing splits of 772, 91 and 102 patients, respectively, while maintaining class distribution (LGE presence/absence).

Implementation Details

Vision Encoder. The vision encoder consists of a Densenet121 (Huang et al. 2017) encoder with UNet decoder, with 5 downsampling layers. A MaxPool layer is added after the last layer of the DenseNet encoder to get a feature vector size of n = 1024. The projection module consists of two Linear layers, separated by GelU non-linearity(Hendrycks and Gimpel 2016) and followed by Dropout and LayerNorm (Ba 2016).

For every patient, we select the segmented PSIR DICOM series (Muehlberg et al. 2018) using information in their DICOM tags, as this sequence theoretically has the higher spatial resolution required for LGE detection. In this dataset, these typically consist of 3 slices, covering apical, mid and basal regions of the heart. Each image is preprocessed according to the following steps: a) resizing to $1mm \times 1mm$ resolution, b) cropping to 112×112 dimension, centered around the LV. LV mask is obtained from the LGE segmentation network described previously, d) Upsampling $2\times$ to 224×224 e) capping intensities at the 98 percentile and f) normalizing to the range of [0,1].

Text Encoder. We use the publicly available BiomedBERT (Zhang et al. 2023a). Feature vector has size $m = 768$. The text encoder is held frozen for all experiments in this study. The projection module has the same architecture as described in the previous section and is optimized end-to-end during training.

For each patient, text relevant to LGE imaging is extracted from the respective clinical report, from both the ”Findings” and the ”Impressions” sections, using a simple keyword search. The extracted text is split into individual sentences (henceforth also referred to as captions), from which one is sampled randomly for every iteration of training. However, this approach introduces an issue: the clinical text annotations are provided at the patient level, whereas the corresponding images represent specific heart sub-regions. To address this, we implement a straightforward solution by appending the phrase ”`This image is from <slice location> level`” to each input text, where <> is replaced with basal, mid or apical, as per the image location. This is done consistently during both training and inference to help the model contextualize the image within the anatomical structure. While we limit our method here to three pre-selected slices for simplicity, this framework is adaptable to any number of images by adjusting the corresponding section tag.

Scar augmentation parameters. In all experiments, we use $\lambda = 0.7$, $[\rho_{min}, \rho_{max}] = [0.1, 0.4], [0.3, 0.6], [0.7, 0.1]$ for single-layer, two-layer, and transmural extents respectively, $s_1 = s_2 = 2$, $b_1 = 0.8, b_2 = 1$. These were selected empirically to ensure image fidelity, anatomical relevance, and diversity across generated outputs.

Baselines and Metrics

We compare the proposed method against the following:

a) BiomedCLIP (Zhang et al. 2023b): We test the model on the same test set, using the same query text: `there is hyperenhancement in the myocardium` and `there is no hyperenhancement in the myocardium`.

Table 1: Balanced accuracies, throughput and memory consumption of the compared methods (n = 102 patients)

Method	FPS ↑	Mem (GB) ↓	Accuracy
BiomedCLIP (Zhang et al. 2023b)	74.4	0.76	0.58
MedFlamingo (Moor et al. 2023)	1.4	31.09	0.50
Image-only Classifier	84.5	0.04	0.77
Proposed method	53.0	0.88	**0.83**

b) MedFlamingo (Moor et al. 2023): The inference query is constructed as a prompt for few-shot, visual question-answering using examples: `<image 1> Does this cardiac LGE MR image show hyperenhancement in the LV myocardium? Answer: No. <image 2> Does this cardiac LGE MR image show hyperenhancement in the LV myocardium? Answer: Yes. <image 3> Does this cardiac LGE MR image show hyperenhancement in the LV myocardium? Answer: ,` where `image 1,image 2` are example images from the training set, without and with LGE respectively, and `image 3` is the query image.

c) Image-only classifier: A model is constructed to be identical to the CLIP based model, but without the text part, i.e the image encoder and a projection module followed by a classification head. The image encoder is initialized with the same pretrained weights. The classification head consists of two Linear layers separated by a GeLU non-linearity and a Dropout layer. The baseline model is trained with the GT binary labels extracted from the clinical reports, using binary cross entropy loss.

To reduce the stochastic variation in the results, multiple (n = 3) models were trained in every experiment and metrics were averaged. Balanced accuracy is adopted as the main metric to account for class imbalance in the dataset. Adam (Kingma 2014) optimizer is used with a learning rate of $1e^{-4}$. All models are trained across 4 x NVIDIA A100-SXM4-40GB GPUs using a batch size of 128, with fully distributed parallel processing. Memory consumption (in GB) and throughput (in FPS) are measured on a single GPU of the same specifications.

Results

Table 1 shows the quantitative results. The proposed method obtains a balanced accuracy of 0.83, outperforming the baselines. Both the publicly available medical VLMs (BiomedCLIP and MedFlamingo) encounter limitations on this task. These challenges likely stem from two key factors: (a) the VLMs were trained on a diverse array of medical imag-

Table 2: Ablation studies: Model is retrained using leave-one-out strategy, keeping everything else the same. pp stands for percentage points.

Method	Accuracy
Proposed	0.83
(-) Synthetic scar augmentation	0.73 (-10 pp)
(-) LV orientation normalization	0.77 (-6 pp)
(-) Caption loss	0.80 (-3 pp)

Table 3: Ablation study: Effect of different encoders on performance.

Method	Accuracy
Proposed - Task specific encoder	**0.83**
Task agnostic encoder - CMR FM (Jacob et al. 2024)	0.80
Imagenet pretrained encoder	0.75

ing data, with cardiac MR constituting only a small subset, and LGE sequences representing an even smaller fraction of this subset; (b) LGE detection is an inherently challenging task that necessitates specialized clinical domain knowledge and the ability to analyze subtle, fine-grained features within highly localized regions of the images. The image-only classifier trained on this dataset demonstrates higher performance but lags behind the proposed method by 6 pp.

Ablation Studies Table 2 illustrates the impact of omitting different components of the proposed method. Among the components, synthetic scar augmentation has the most significant impact on performance, followed by the normalization of LV orientation, and lastly, the caption loss.

Table 3 presents the impact of different initialization strategies for the vision encoder. As previously described, the proposed model uses a vision encoder pre-trained on the related task of myocardium segmentation in LGE images. Here, we explore two alternative initializations:

a) Task agnostic encoder: This is a unimodal foundation model pre-trained on 36 million CMR images (Jacob et al. 2024). It was trained in a self-supervised manner, without any labelled data, hence is agnostic to any specific task. The model consisted of a ViT-S architecture, and was pretrained across many diverse sequences of CMR, such as cine, LGE, and mapping.

b) Imagenet pretrained encoder: This uses the same DenseNet-UNet architecture of the proposed model, but with the publicly available ImageNet trained weights.

Both initialization choices provide practical alternatives for when training data and/or labels for a directly related task is unavailable. Our results indicate that the task-agnostic CMR foundation model (FM), pre-trained in a self-supervised manner, outperforms the ImageNet pre-trained model; however, it still falls short compared to the model trained on a closely related task.

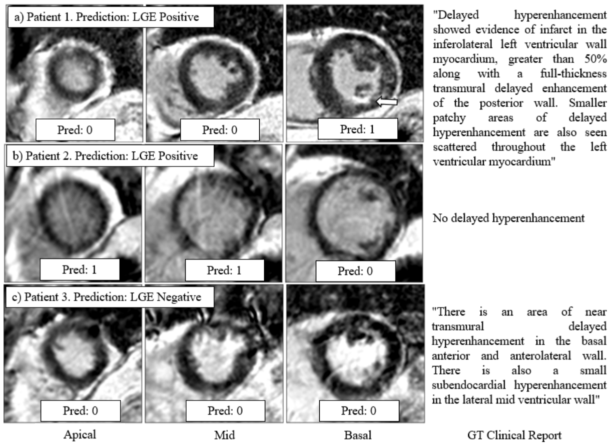
a) Patient 1. Prediction: LGE Positive
Pred: 0
Pred: 0
Pred: 1
"Delayed hyperenhancement showed evidence of infarct in the inferolateral left ventricular wall myocardium, greater than 50% along with a full-thickness transmural delayed enhancement of the posterior wall. Smaller patchy areas of delayed hyperenhancement are also seen scattered throughout the left ventricular myocardium"
b) Patient 2. Prediction: LGE Positive
Pred: 1
Pred: 1
Pred: 0
No delayed hyperenhancement
c) Patient 3. Prediction: LGE Negative
Pred: 0
Pred: 0
Pred: 0
"There is an area of near transmural delayed hyperenhancement in the basal anterior and anterolateral wall. There is also a small subendocardial hyperenhancement in the lateral mid ventricular wall"
Apical
Mid
Basal
GT Clinical Report

Figure 4: Qualitative results on real clinical data. Model predictions for each image, with the patient-level GT text

Qualitative Results

Fig 4 visualizes the results for 3 patients. Patient 1 shows a true positive detection. Note that though the GT text describes LGE presence across the whole heart, the model is able to produce predictions for individual slices, which aids interpretability. Patient 2 presents a false positive, with the network predicting the LGE presence in apical and mid-ventricular slices. This could be because of the presence of streak artifacts in these two images, degrading image quality and possibly confounding the model. Patient 3 presents a false negative case, with no LGE detected by the model, despite the GT indicating LGE in basal slices. Visual inspection did not reveal LGE in these slices; however, further review of other LGE images in the study confirmed the presence of basal scarring at the level of the LV outflow tract. This highlights a method limitation, as selecting only three input images may omit critical details, reducing the model's efficacy. This is also observed in Patient 1, where the selected images do not reflect all of the described scar.

Conclusions

Text-based training using clinical reports offers an alternative to obtaining hard labels in the clinical domain, which might be expensive and challenging. However, typically used methods such as CLIP, require large amounts of pre-training data, and further finetuning stages for downstream tasks. We present a method that incorporates domain knowledge to enable CLIP based training for small datasets. We create synthetic image-text pairs to augment the training set, use anatomical information to normalize the orientation of the image, use additional caption loss to enable fine-grained supervision and use related-task pretraining to improve the accuracy for the task. We demonstrate the feasibility of text-based training for specific tasks on small datasets, followed by zero-shot inference without any further finetuning stages.

Disclaimer. The concepts and information presented in this paper/presentation are based on research results that are not commercially available. Future commercial availability cannot be guaranteed.

References

Amado, L. C.; Gerber, B. L.; Gupta, S. N.; Rettmann, D. W.; Szarf, G.; Schock, R.; Nasir, K.; Kraitchman, D. L.; and Lima, J. A. 2004. Accurate and objective infarct sizing by contrast-enhanced magnetic resonance imaging in a canine myocardial infarction model. *Journal of the American College of Cardiology*, 44(12): 2383–2389.

Ba, J. L. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Cao, A.-Q.; Jaritz, M.; Guillaumin, M.; de Charette, R.; and Bazzani, L. 2024. LatteCLIP: Unsupervised CLIP Fine-Tuning via LMM-Synthetic Texts. *arXiv preprint arXiv:2410.08211*.

Cao, Q.; Najibi, M.; and Mehta, S. 2024. CtrlSynth: Controllable Image Text Synthesis for Data-Efficient Multimodal Learning. *arXiv preprint arXiv:2410.11963*.

Chen, H.; Zhao, B.; Yue, G.; Liu, W.; Lv, C.; Wang, R.; and Zhou, F. 2024. Clip-Medfake: Synthetic Data Augmentation With AI-Generated Content for Improved Medical Image Classification. In *2024 IEEE International Conference on Image Processing (ICIP)*, 3854–3860. IEEE.

Flett, A. S.; Hasleton, J.; Cook, C.; Hausenloy, D.; Quarta, G.; Ariti, C.; Muthurangu, V.; and Moon, J. C. 2011. Evaluation of techniques for the quantification of myocardial scar of differing etiology using cardiac magnetic resonance. *JACC: cardiovascular imaging*, 4(2): 150–156.

Girum, K. B.; Skandarani, Y.; Hussain, R.; Grayeli, A. B.; Créhange, G.; and Lalande, A. 2021. Automatic myocardial infarction evaluation from delayed-enhancement cardiac MRI using deep convolutional networks. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, 378–384. Springer.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heiberg, E.; Engblom, H.; Carlsson, M.; Erlinge, D.; Atar, D.; Aletras, A. H.; and Arheden, H. 2022. Infarct quantification with cardiovascular magnetic resonance using" standard deviation from remote" is unreliable: validation in multicentre multi-vendor data. *Journal of Cardiovascular Magnetic Resonance*, 24(1): 53.

Hendrycks, D.; and Gimpel, K. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Hsu, L.-Y.; Natanzon, A.; Kellman, P.; Hirsch, G. A.; Aletras, A. H.; and Arai, A. E. 2006. Quantitative myocardial infarction on delayed enhancement MRI. Part I: Animal validation of an automated feature analysis and combined thresholding infarct sizing algorithm. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 23(3): 298–308.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Ikezogwo, W.; Seyfioglu, S.; Ghezloo, F.; Geva, D.; Sheikh Mohammed, F.; Anand, P. K.; Krishna, R.; and Shapiro, L. 2024. Quilt-1m: One million image-text pairs for histopathology. *Advances in neural information processing systems*, 36.

Jacob, A. J.; Borgohain, I.; Chitiboi, T.; Sharma, P.; Comaniciu, D.; and Rueckert, D. 2024. Towards a vision foundation model for comprehensive assessment of Cardiac MRI. *arXiv preprint arXiv:2410.01665*.

Jenista, E. R.; Wendell, D. C.; Azevedo, C. F.; Klem, I.; Judd, R. M.; Kim, R. J.; and Kim, H. W. 2023. Revisiting how we perform late gadolinium enhancement CMR: insights gleaned over 25 years of clinical practice. *Journal of Cardiovascular Magnetic Resonance*, 25(1): 18.

Kim, Y.-C.; Chung, Y.; and Choe, Y. H. 2024. Deep learning for classification of late gadolinium enhancement lesions based on the 16-segment left ventricular model. *Physica Medica*, 117: 103193.

Kingma, D. P. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lalande, A.; Chen, Z.; Decourselle, T.; Qayyum, A.; Pommier, T.; Lorgis, L.; de la Rosa, E.; Cochet, A.; Cottin, Y.; Ginhac, D.; et al. 2020. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac MRI. *Data*, 5(4): 89.

Lalande, A.; Chen, Z.; Pommier, T.; Decourselle, T.; Qayyum, A.; Salomon, M.; Ginhac, D.; Skandarani, Y.; Boucher, A.; Brahim, K.; et al. 2022. Deep learning methods for automatic evaluation of delayed enhancement-MRI. The results of the EMIDEC challenge. *Medical Image Analysis*, 79: 102428.

Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; and Gao, J. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Lin, J.; and Gong, S. 2023. Gridclip: One-stage object detection by grid-level clip representation learning. *arXiv preprint arXiv:2303.09252*.

Luo, H.; Bao, J.; Wu, Y.; He, X.; and Li, T. 2023. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *International Conference on Machine Learning*, 23033–23044. PMLR.

Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E. P.; and Rajpurkar, P. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, 353–367. PMLR.

Muehlberg, F.; Arnhold, K.; Fritschi, S.; Funk, S.; Prothmann, M.; Kermer, J.; Zange, L.; von Knobelsdorff-Brenkenhoff, F.; and Schulz-Menger, J. 2018. Comparison of fast multi-slice and standard segmented techniques for detection of late gadolinium enhancement in ischemic and non-ischemic cardiomyopathy – a prospective clinical cardiovascular magnetic resonance trial. *Journal of Cardiovascular Magnetic Resonance*, 20(1): 13.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Shakeri, F.; Huang, Y.; Silva-Rodríguez, J.; Bahig, H.; Tang, A.; Dolz, J.; and Ben Ayed, I. 2024. Few-Shot Adaptation of Medical Vision-Language Models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 553–563. Springer.

Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Hou, L.; Clark, K.; Pfohl, S.; Cole-Lewis, H.; Neal, D.; et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Wang, Z.; Wu, Z.; Agarwal, D.; and Sun, J. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Yang, S.; and Wang, X. 2021. A hybrid network for automatic myocardial infarction segmentation in delayed enhancement-mri. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, 351–358. Springer.

Yu, J.; Wang, Z.; Vasudevan, V.; Yeung, L.; Seyedhosseini, M.; and Wu, Y. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; et al. 2023a. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2(3): 6.

Zhang, S.; Xu, Y.; Usuyama, N.; Xu, H.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; et al. 2023b. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*.

Zhang, X.; Xu, M.; Qiu, D.; Yan, R.; Lang, N.; and Zhou, X. 2024. Mediclip: Adapting clip for few-shot medical image anomaly detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 458–468. Springer.

Zhang, Y. 2021. Cascaded convolutional neural network for automatic myocardial infarction segmentation from delayed-enhancement cardiac MRI. In *Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4, 2020, Revised Selected Papers 11*, 328–333. Springer.

Zhang, Y.; Jiang, H.; Miura, Y.; Manning, C. D.; and Langlotz, C. P. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, 2–25. PMLR.

Zhao, Z.; Liu, Y.; Wu, H.; Wang, M.; Li, Y.; Wang, S.; Teng, L.; Liu, D.; Cui, Z.; Wang, Q.; et al. 2023. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022a. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16816–16825.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022b. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022c. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.