Unified Multi-modal Diagnostic Framework with Reconstruction Pre-training and Heterogeneity-combat Tuning

Yupei Zhang, Li Pan, Qiushi Yang, Tan Li, and Zhen Chen

***Abstract*— Medical multi-modal pre-training has revealed promise in computer-aided diagnosis by leveraging large-scale unlabeled datasets. However, existing methods based on masked autoencoders mainly rely on data-level reconstruction tasks, but lack high-level semantic information. Furthermore, two significant heterogeneity challenges hinder the transfer of pre-trained knowledge to downstream tasks, *i.e.*, the distribution heterogeneity between pre-training data and downstream data, and the modality heterogeneity within downstream data. To address these challenges, we propose a Unified Medical Multi-modal Diagnostic (UMD) framework with tailored pre-training and downstream tuning strategies. Specifically, to enhance the representation abilities of vision and language encoders, we propose the Multi-level Reconstruction Pre-training (MR-Pretrain) strategy, including a feature-level and data-level reconstruction, which guides models to capture the semantic information from masked inputs of different modalities. Moreover, to tackle two kinds of heterogeneities during the downstream tuning, we present the heterogeneity-combat downstream tuning strategy, which consists of a Task-oriented Distribution Calibration (TD-Calib) and a Gradient-guided Modality Coordination (GM-Coord). In particular, TD-Calib fine-tunes the pre-trained model regarding the distribution of downstream datasets, and GM-Coord adjusts the gradient weights according to the dynamic optimization status of different modalities. Extensive experiments on five public medical datasets demonstrate the effectiveness of our UMD framework, which remarkably outperforms existing approaches on three kinds of downstream tasks.**

***Index Terms*— medical multi-modal diagnosis, reconstruction pre-training, downstream tuning**

I. Introduction

RECENTLY, deep learning techniques have shown advantages in computer-aided diagnosis, mainly relying on the knowledge of expert-annotated medical datasets [1], [2]. However, collecting high-quality annotations for medical data is time-consuming and costly, making it difficult to construct large-scale medical datasets, thereby restricting the performance of current diagnostic algorithms that benefit from the expert annotations [3]. Instead, a rational alternative is to exploit the knowledge of large amounts of unlabeled medical data effectively, which enhances the diagnostic performance and broadens the application scenarios of diagnostic algorithms. In particular, self-supervised pre-training [4] provides a *pretrain-finetune* paradigm that first performs pre-training on large-scale unlabeled data for superior representation learning, and then conducts the fine-tuning on a small amount of labeled data to adapt to downstream tasks.

Different from medical imaging with uni-modality, multi-modal medical data (*e.g.*, medical images and text descriptions) can improve diagnostic accuracy for various diseases by incorporating additional cross-modal knowledge independent of manual labeling [5]. Multi-modal pre-training [6] aims to encourage the model to capture semantic information in a self-supervised manner. By regularizing models to inter-modality and intra-modality, multi-modal pre-training works can be broadly categorized into two groups: contrastive learning-based [7], [8] and masked autoencoder-based methods [9]–[11]. Contrastive learning-based methods train models to differentiate between similar and dissimilar pairs of data samples. Note that the paradigm of pushing or pulling samples in contrastive learning requires extremely large batch sizes and suffers from low efficiency, and the model performances are highly affected by the tricky selection of positive pairs and negative pairs [12]. Meanwhile, masked autoencoder-based methods randomly remove a large proportion of the original data and encourage the model to reconstruct them [13]. This pre-training strategy efficiently achieves notable improvement over various downstream tasks, benefiting from more abundant supervision [14]. Although advancements have been achieved by masked autoencoder-based methods, they still have two major limitations worthy of improving.

The first limitation is the insufficient feature representation caused by using heuristic reconstruction targets, which may not fully capture the underlying structure of the data and result in insufficient pre-training. As illustrated in Fig. 1 (a), most of the masked autoencoder-based methods [15] simply employ original data (*e.g.*, image pixels and text tokens) as prediction targets. However, this strategy can lead to overfitting

This work was supported in part by the InnoHK program. (Yupei Zhang and Li Pan contributed equally to this work) (Corresponding author: Zhen Chen, e-mail: zhen.chen@cair-cas.org.hk.)

Y. Zhang is with Centre for Intelligent Multidimensional Data Analysis (CIMDA), Hong Kong SAR.

L. Pan is with Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong SAR.

Q. Yang is with Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR.

T. Li is with Department of Computer Science, The Hang Seng University of Hong Kong, Hong Kong SAR.

Z. Chen is with Centre for Artificial Intelligence and Robotics (CAIR), HKISI, Chinese Academy of Sciences, Hong Kong SAR.

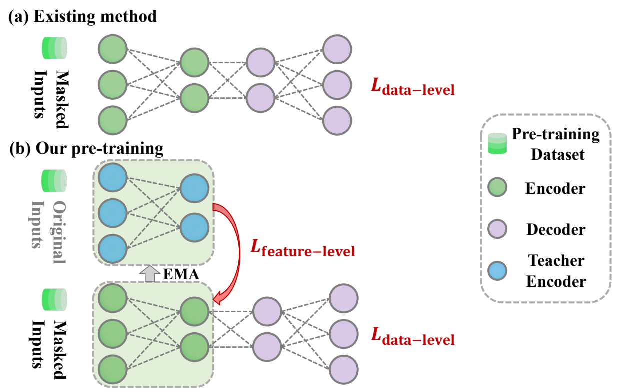


Fig. 1. The comparison of pre-training strategies. Different from the existing methods (a) that aim for data-level reconstruction, we design a novel multi-level reconstruction pre-training (b) that enhances the encoder to learn transferable semantic features by incorporating data-level and feature-level reconstruction.

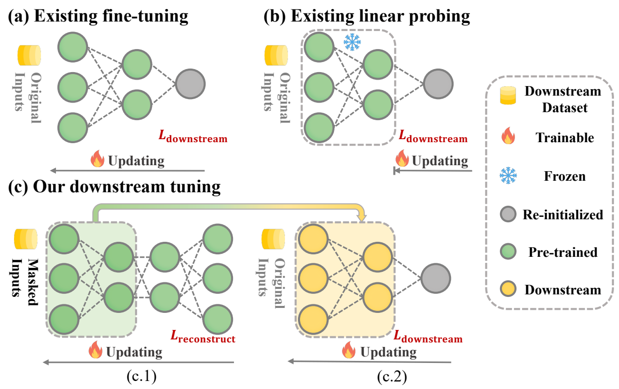


Fig. 2. The comparison of fine-tuning strategies. Different from existing methods (a) and (b) that directly fine-tune the entire network or the final linear layer, we design a novel heterogeneity-combat downstream tuning (c) that promotes the encoder to learn semantic features of downstream data with reconstruction and boosts various downstream tasks.

of local statistics and high-frequency details that may be less relevant for data interpretation, such as background, illumination, and noise disturbance [16]. To address the disadvantages of data-level reconstruction, recent studies attempt to use various manually extracted features as reconstruction targets. LocalMIM [17], MaskFeat [16], and FreMAE [13] incorporate handcrafted feature descriptors (*e.g.*, SIFT [18], HOG [19], and Fourier spectrum [13]) of the original data, to enhance the model's understanding on high-level features. Nevertheless, manually designed descriptors with specific strategies are sub-optimal and limit the model's generalization to other tasks and datasets. Rather than heuristically defining original data and handcrafted feature descriptors as reconstruction targets, we aim to further guide the model with masked inputs to reconstruct the high-level features of the original inputs. As shown in Fig. 1 (b), two networks extract features of masked and original data respectively, and the target features of original data are dynamically adjusted by the model to perform the feature-level reconstruction, and accordingly, the difficulty of reconstruction task is modulated. By incorporating this feature-level reconstruction, our method enhances semantic understanding and improves feature representation learning.

Another critical limitation is that existing pre-training works ignore the connection between the pre-training and fine-tuning stages, which hinders the knowledge transfer from pre-training to downstream tasks [20]. Herein, we formulate this challenge from two perspectives, *i.e.*, the distribution heterogeneity between pre-training and downstream data and the modality heterogeneity within multi-modal downstream data. On the one hand, pre-training optimizes models to be robust on pre-training datasets, while the ultimate pursuit of model performance is for downstream task scenarios with data distribution shift [20], [21]. To perform the downstream diagnostic tasks, current medical pre-training methods [22], as shown in Fig. 2 (a) and (b), directly fine-tune the entire network or the last linear layer of pre-trained model on the downstream datasets [6], [23]. However, simple fine-tuning may not be enough to bridge the distribution gap from pre-training to downstream datasets, resulting in even worse performance than specialized expert models on the target downstream task [24], [25]. On the other hand, during the fine-tuning phase, existing multi-modal pre-training methods [10], [26] jointly optimize the modules of different modalities. Nevertheless, since the rate of convergence varies for different modalities, this joint optimization strategy may lead the modules to the sub-optimum [27], where the potential of multi-modal data hasn't been fully exploited by multi-modal models [28]. To address the above issues, we adapt the model for the specific downstream dataset through the same reconstruction task as in the pre-training stage, as shown in Fig. 2 (c), to enhance the model's awareness toward the target distribution. Additionally, we present a dynamic gradient weighting mechanism for different modalities to ensure coordinated multi-modal training. By these means, our model can more effectively leverage pre-trained knowledge and adaptively balance the optimization of multiple modalities, ultimately improving performance on downstream medical diagnosis tasks.

In this work, we propose the Unified Medical Multi-modal Diagnostic (UMD) framework with Multi-level Reconstruction Pre-training (MR-Pretrain) and heterogeneity-combat downstream tuning strategies, which leverage the vast amounts of unlabelled medical data in pre-training, and bridge gaps in terms of distribution and modality. In addition to the data-level reconstruction supervision, we strengthen the constraints on the encoder by performing a novel feature-level reconstruction. Specifically, we feed the original data into a teacher encoder and the masked data into a student encoder, where the student is supervised with the features extracted by the teacher in the feature space. This design enhances the encoder's representation learning for high-level semantic information and thus improves the generalization ability of pre-training. Moreover, to combat the inter-dataset distribution heterogeneity and inter-modality optimization interference, we propose a task-oriented distribution calibration (TD-Calib) module and a gradient-guided modality coordination (GM-Coord) module. TD-Calib calibrates the model trained on the pre-training datasets with instances from the downstream datasets in a mask-and-reconstruct manner, and GM-Coord dynamically adjusts the gradient weights of different modalities for coordinated multi-modal tuning. By enhancing the model's understanding of high-level semantic information with masked inputs and bridging the gap between pre-training and fine-tuning, the proposed UMD framework achieves superior performance on three kinds of downstream tasks. Our contributions are four-fold:

- To perform a more accurate diagnosis using medical multi-modal data, we introduce UMD, a novel unified framework that incorporates data-level and feature-level reconstruction to improve representation learning and heterogeneity-combat downstream tuning to bridge gaps in terms of distribution and modality.
- To promote representation capabilities, we devise the MR-Pretrain strategy to enhance the multi-modal encoders with feature-level reconstruction in addition to data-level reconstruction. The MR-Pretrain feeds masked multi-modal samples into the student model to reconstruct target feature representations obtained from the teacher model with the original inputs, enabling the model to learn richer transferable multi-modal representations.
- To improve the transfer ability of the pre-trained model to various downstream tasks, we introduce the heterogeneity-combat downstream tuning, composed of two simple but effective modules, *i.e.*, TD-Calib and GM-Coord. As such, our UMD framework can effectively bridge the distribution gap between pre-train data and downstream data, and thoroughly unleash the potential of multi-modal data.
- We conduct experiments on three kinds of downstream tasks using five public multi-modal medical datasets. The results demonstrate the effectiveness of our UMD framework, which outperforms the state-of-the-art methods by a significant margin on all datasets.

II. Related work

A. Multi-Modal Pre-Training

Inspired by the great success achieved in uni-modal pre-training (*e.g.*, natural language processing and computer vision), such as BERT [29] and MAE [15], the multi-modal pre-training has gained increasing attention in recent years [30]. The multi-modal pre-training aims to learn universal transferable representations from large-scale unannotated multi-modal data. Generally, the inter-view and intra-view perspectives on the image and text lead to two main streams of pretext design for multi-modal pre-training, *i.e.*, contrastive learning [7], [8] and masked multi-modal modeling [9]–[11].

Contrastive learning trains models to maximize the similarity between positive pairs and minimize the similarity among negative pairs. Based on this simple idea, a large number of studies extend contrastive learning to perform self-supervised pre-training [7]. Despite its effectiveness in learning useful representations, contrastive learning suffers from two drawbacks. Firstly, it demands a significant number of negative samples, which can be resource-intensive [11]. Secondly, it relies on the complex manual definition of positive and negative sample pairs [16].

Masked autoencoder is another type of paradigm for vision and language pre-training, which masks a portion of the input data and learns to recover the removed content [15]. This mask-and-reconstruct strategy significantly reduces computational costs and encourages the model to learn data representations in a self-supervised manner [29]. Specifically, MAE [15] demonstrated the self-supervised learning capability of masked autoencoders in computer vision by adopting the mask-and-reconstruct pretext which masked image patches. Singh *et al.* [9] proposed unified pre-training schemes for the vision and language data by applying the mask-and-reconstruct to each modality. MaskVLM [10] improved the existing vision-and-language pre-training approaches by alternately masking one modality to enhance the cross-modality alignment. DeepMIM [31] boosted the masked image modeling by the deep supervision of intermediate features to drive the shallower layers to learn meaningful representations.

As discussed above, most existing masked autoencoders set the original inputs as reconstruction targets. However, this data-level reconstruction strategy may cause overfitting to the low-level local statistics and high-frequency details, which can impede the model from capturing high-level semantic features from the inputs [16]. To address this challenge, some recent studies attempt to improve feature-level supervision on the intermediate outputs of the encoders. For instance, MaskFeat [16] explicitly utilized the handcrafted image descriptors (*e.g.*, HOG) as reconstruction targets to enhance feature representations. Wang *et al.* [17] proposed a multi-scale reconstruction approach, which encourages the encoder to predict various handcrafted image descriptors at different layers. Yet, these methods bring strong assumptions on the reconstruction targets, which are biased and hard to be generalized due to the manual inputs. Different from previous works, our UMD framework first performs pre-training with feature-level reconstruction to enhance feature representation learning, and then promotes fine-tuning stage through the tailored heterogeneity-combat downstream tuning.

B. Medical Multi-Modal Pre-Training

The medical multi-modal pre-training aims to improve the performance of diagnostic models by leveraging large-scale unlabeled multi-modal medical datasets [6]. On the one hand, unlike general computer vision datasets, medical datasets are naturally multi-modal, containing diverse imaging types and text data (*e.g.*, diagnosis reports) [32], [33]. On the other hand, medical data requires manual annotations of human experts, which is time-consuming and costly [3]. Therefore, there is a high demand for developing a self-supervised multi-modal pre-training method that can utilize unannotated medical data to improve the performance of existing deep learning models.

To achieve this goal, recent studies have explored self-supervised pre-training on medical datasets. For example, Li *et al.* [23] validated the effectiveness of medical multi-modal pre-training by evaluating four pre-trained vision-and-language models on medical datasets. To improve the performance of visual question-answering models, Khare *et al.* [34] tokenized the medical images using convolutional neural networks to jointly pre-train both vision and language encoders under masked reconstruction modeling. Zhang *et al.* [35]

utilized contrastive learning to pre-train models on paired medical images and texts, and evaluated pre-trained models on three medical imaging tasks, *i.e.*, image classification, zero-shot image-image retrieval, and zero-shot text-image retrieval. Endo *et al.* [36] pre-trained a model on public datasets for gait movement forecasting, which can be further applied to clinical data to predict the severity of gait impairment for diagnosing Parkinson's disease. Moon *et al.* [37] presented a multi-modal attention masking approach to maximize generalization ability for both medical vision-language understanding tasks. Chen *et al.* [6] proposed a transformer-based pre-training model via multi-modal masked autoencoders, and achieved promising results on multiple multi-modal medical downstream tasks. On this basis, our UMD framework elaborately investigates the entire process from pre-training to fine-tuning, and leverages the characteristics of multi-modal medical data to facilitate the performance of medical diagnoses.

III. Methodology

A. Overview

Our UMD framework contains two stages, *i.e.*, MR-Pretrain in Fig. 3 to enhance the general feature representation and the heterogeneity-combat downstream tuning in Fig. III-B.2 to boost the fine-tuning performance on various downstream tasks. In the MR-Pretrain stage, besides the widely-applied data-level reconstruction [10], we propose the feature-level reconstruction by a dual-stream workflow to encourage transferable representation learning from high-level features. In the heterogeneity-combat downstream tuning stage, the TD-Calib bridges the distribution gap between the pre-training and downstream datasets, and the GM-Coord adjusts the gradient optimization of different modalities, thereby facilitating the performance of downstream tasks.

1) Model Architecture: **Multi-modal encoder** $\mathcal{E}$ comprises a student multi-modal encoder $\mathcal{E}(\theta)$ and a teacher multi-modal encoder $\mathcal{E}(\bar{\theta})$. They share the same network architecture, and each of them consists of a vision transformer (ViT)-based [7] vision encoder $\mathcal{E}^I$, a transformer-based [38] language encoder $\mathcal{E}^T$, and a cross-attention-based multi-modal fusion module $\mathcal{F}$ [6]. For the multi-modal fusion module, we use two N_m-layer transformer models. Each model includes a self-attention layer for intra-modality learning, a cross-attention layer for inter-modality learning, and a feed-forward layer. The weights of teacher model $\mathcal{E}(\bar{\theta})$ is updated by the exponential moving average (EMA) [39] of the weights from student model $\mathcal{E}(\theta)$.
Decoder $\mathcal{D}$ comprises a vision decoder $\mathcal{D}^I$ and a language decoder $\mathcal{D}^T$, and is designed to reconstruct the original image and text using the latent representations obtained through multi-modal fusion $\mathcal{F}$. Vision decoder $\mathcal{D}^I$ aims to reconstruct raw pixels that contain low-level textural information, while language decoder $\mathcal{D}^T$ is expected to recover the text tokens that represent high-level semantic information. To this end, we employ a transformer [40] as the vision decoder for the low-level reconstruction, and the multi-layer perceptron (MLP) is utilized as the language decoder.
Downstream task head $\mathcal{H}$ for visual question-answering and image-text classification tasks, is a fully connected neural network with a layer normalization [41] and a GELU activation [42]. For the image-text retrieval task, we utilize a linear layer as the head.

2) Dataflow: The input of our UMD framework consists of two streams: the image-text pair (I, T) and the corresponding masked pair $(I^{\text{mask}}, T^{\text{mask}})$. The $I \in \mathbb{R}^{H \times W \times C}$ and $T \in \mathbb{R}^{L}$ represent the image and text, where H and W are image resolution, C is the number of image channels, and L is the length of a text sample. Following M^3AE [6], we employ data sequentialization, linear projection embeddings, random masking, and position embeddings during the data preprocessing. A start-of-sequence token embedding and a special boundary token embedding are appended to the text sequence to indicate the beginning and end of the input text [40]. The vision $\mathcal{E}^I$ and language $\mathcal{E}^T$ encoder extract contextual representations of the image H^I and text H^T from the image-text pair $(I^{\text{mask}}, T^{\text{mask}})$. We further fuse H^I and H^T using the multi-modal fusion module $\mathcal{F}$, which produces multi-modal representations $Z^I = [z^I_{\text{CLS}}; z^I_1; z^I_2; ...; z^I_{n^I}]$ for vision and $Z^T = [z^T_{\text{CLS}}; z^T_1; z^T_2; ...; z^T_{n^T}; z^T_{\text{SEP}}]$ for language, where n^I and n^T are the numbers of image patches and text tokens, respectively. To perform data-level reconstruction of the MR-Pretrain stage and TD-Calib of the heterogeneity-combat downstream tuning stage, we concatenate the average embeddings of Z^I and Z^T, and feed the resulting vector into $\mathcal{D}$. For GM-Coord, we feed the concatenated embeddings into $\mathcal{H}$, for downstream tasks' prediction.

B. Multi-Level Reconstruction Pre-Training

To promote the representation capabilities of the multi-modal encoders, we propose a multi-level reconstruction method in Fig. 3, named MR-Pretrain, containing a novel feature-level reconstruction loss and a data-level reconstruction loss, encouraging the encoders to learn the high-level semantic representation from the unlabeled medical data.

1) Data-Level Reconstruction: The information density of language is poles apart from vision [6], [15], where the language is highly informative, and the vision is instead spatially redundant. To eliminate the redundancies of vision and language while enabling the model to acquire valuable features from both vision and language in data-level reconstruction, we randomly mask image I with a 75% ratio and mask text T with a 15% masking ratio. Then we recover the masked inputs using the remaining data. The masked image modeling (MIM) loss and masked language modeling (MLM) loss are defined as follows:

$$L_{\text{MIM}} = \frac{1}{N} \sum_{n=1}^{N} (Y_{\text{MIM}} - \mathcal{D}(\mathcal{E}\ (I_n^{\text{mask}}, T_n^{\text{mask}}; \theta))^2, \quad (1)$$

$$L_{\text{MLM}} = -\frac{1}{N} \sum_{n=1}^{N} \log P_{\text{MLM}}(Y_{\text{MLM}} \mid I_n^{\text{mask}}, T_n^{\text{mask}}), \quad (2)$$

where N is the total number of samples, Y_{MIM} is the raw pixel values of masked image patches, Y_{MLM} represents the labels of masked text tokens, and P_{MLM} represents the likelihood of each label given the input image-text pair. By supervising the model to reconstruct details from masked image-text pairs, this data-level reconstruction task encourages the model to perceive low-level textual information without manual labels.

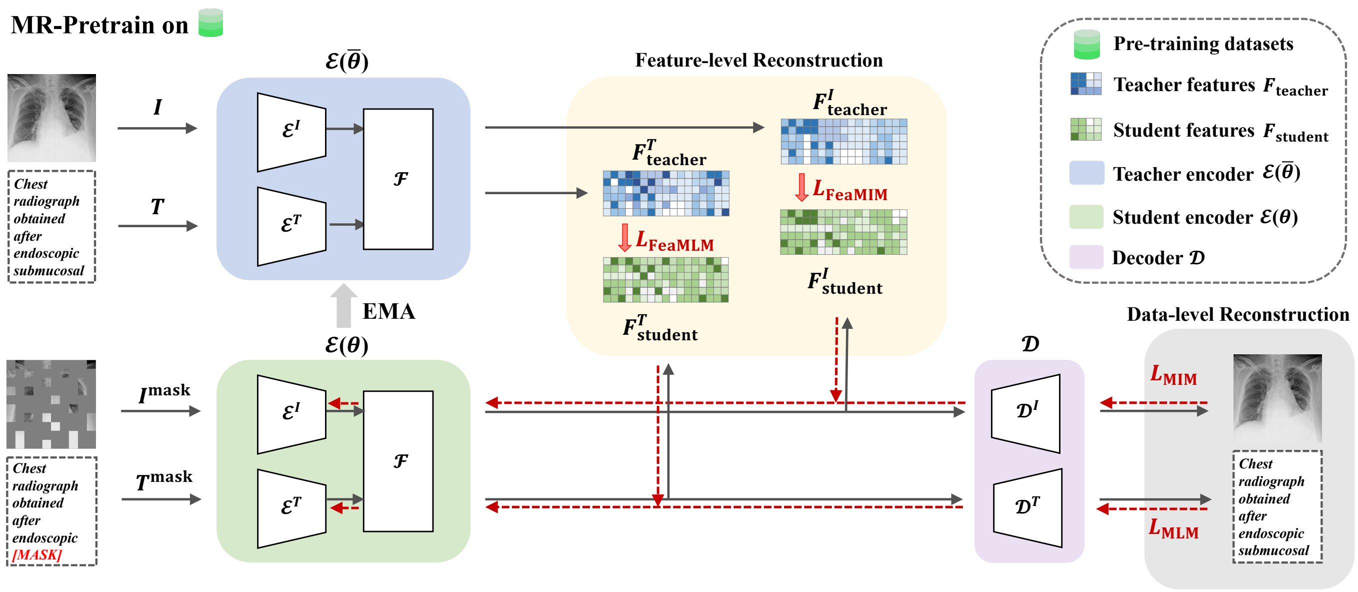

MR-Pretrain on
I
T
Chest radiograph obtained after endoscopic submucosal
ℰ(θ̄)
ℰ^I
ℰ^T
ℱ
EMA
ℰ(θ)
I^mask
T^mask
Chest radiograph obtained after endoscopic [MASK]
Feature-level Reconstruction
F^I teacher
L_FeaMIM
F^I student
F^T teacher
L_FeaMLM
F^T student
𝒟
𝒟^I
𝒟^T
Data-level Reconstruction
L_MIM
L_MLM
Chest radiograph obtained after endoscopic submucosal
Pre-training datasets
Teacher features F_teacher
Student features F_student
Teacher encoder ℰ(θ̄)
Student encoder ℰ(θ)
Decoder 𝒟

Fig. 3. Our MR-Pretrain exploits generalizable features from a large-scale unlabeled pre-training dataset in a dual-stream workflow. Besides the data-level reconstruction, we perform the feature-level reconstruction pretext task of features to encourage transferable representation learning.

2) Feature-Level Reconstruction: Data-level reconstruction reduces the demand for data labeling, but entirely relying on this method may cause the model to overfit to the fine details of the input, hindering its ability to learn higher-level representations [13], [17]. To address this problem, we propose the feature-level reconstruction that directly supervises the model in the feature space, encouraging the model to capture high-level semantic representations. Specifically, we employ a teacher model $\mathcal{E}(\bar{\theta})$ to extract the representations from original pairs (I, T) and a student model $\mathcal{E}(\theta)$ for the masked image-text pairs $(I^{\text{mask}}, T^{\text{mask}})$. The outputs of the multi-modal encoder model contain two vectors, Z^I for the image and Z^T for the text. Accordingly, the outputs of the student model and teacher model are defined as follows:

$$\begin{aligned} Z_n^I(\theta), Z_n^T(\theta) &= \mathcal{E}\,(I_n^{\text{mask}}, T_n^{\text{mask}}; \theta), \\ Z_n^I(\bar{\theta}), Z_n^T(\bar{\theta}) &= \mathcal{E}\,(I_n, T_n; \bar{\theta}), \end{aligned} \tag{3}$$

where n represents the index of the paired samples. Following [43], we apply two linear layers h^I and h^T as projection heads to map the vision feature Z_n^I and language feature Z_n^T to a lower-dimensional latent space. The feature reconstruction loss for the image L_{FeaMIM} and text L_{FeaMLM} can be formulated as follows:

$$\begin{aligned} L_{\text{FeaMIM}} &= \frac{1}{N}\sum_{n=1}^{N}(h^I(Z_n^I(\bar{\theta})) - \ h^I(Z_n^I(\theta)))^2, \\ L_{\text{FeaMLM}} &= \frac{1}{N}\sum_{n=1}^{N}(h^T(Z_n^T(\bar{\theta})) - \ h^T(Z_n^T(\theta)))^2, \end{aligned} \tag{4}$$

where N is the number of pre-training samples. Different from the data-level reconstruction [6], [10] that encourages the model to concentrate on the low-level details, the proposed multi-modal feature reconstruction L_{FeaMIM} and L_{FeaMLM} guides the model to capture high-level semantic information from masked inputs. The teacher model weights $\bar{\theta}$ are updated under the exponential moving average [39] of the student model weights θ, as follows:

$$\bar{\theta}_t \ = \ \lambda\bar{\theta}_{t-1} + (1 - \ \lambda)\theta_t, \tag{5}$$

where λ is the smoothing factor of teacher model $\bar{\theta}$ updating, and t indicates the current iteration number. Given the full view of the input modalities, the teacher encoder $\mathcal{E}(\bar{\theta})$ can provide global feature-level guidance for the student encoder $\mathcal{E}(\theta)$, which encourages the student encoder $\mathcal{E}(\theta)$ to capture high-level semantic information during the pre-training.

Furthermore, our MR-Pretrain incorporates the image-text matching (ITM) objective, a popular approach in vision-language understanding that aims to distinguish if a pair of image and text is matched. As such, our MR-Pretrain with ITM is effective for learning representations and improving the downstream performance [44]:

$$L_{\text{ITM}} = -\frac{1}{N}\sum_{n=1}^{N}\log P_{\text{ITM}}(Y_{\text{ITM}} \mid I_n, T_n), \tag{6}$$

where P_{ITM} is the probability distribution obtained by applying a softmax function to the ITM decoder that consists of a linear layer, and Y_{ITM} represents the binary label for the ITM task. The value of one indicates a matched image-text pair, while zero indicates a mismatched pair. By promoting a joint representation of image and text inputs, our model enriches the correlation information between the modalities, thereby boosting the downstream tasks.

3) MR-Pretrain Objective: The total MR-Pretrain objective L_{pretrain} consists of five losses, *i.e.*, L_{MIM}, L_{MLM}, L_{FeaMIM}, L_{FeaMLM}, and L_{ITM}, which is calculated as follows:

$$\begin{aligned} L_{\text{pretrain}} \ &= (1-\alpha)(L_{\text{MIM}} \ + \ L_{\text{MLM}}) \ + \\ &\quad \alpha(L_{\text{FeaMIM}} \ + \ L_{\text{FeaMLM}} \ ) \ + \ L_{\text{ITM}}, \end{aligned} \tag{7}$$

where α is a trade-off factor to balance the data-level and feature-level reconstruction. By optimizing Eq. (7), the model

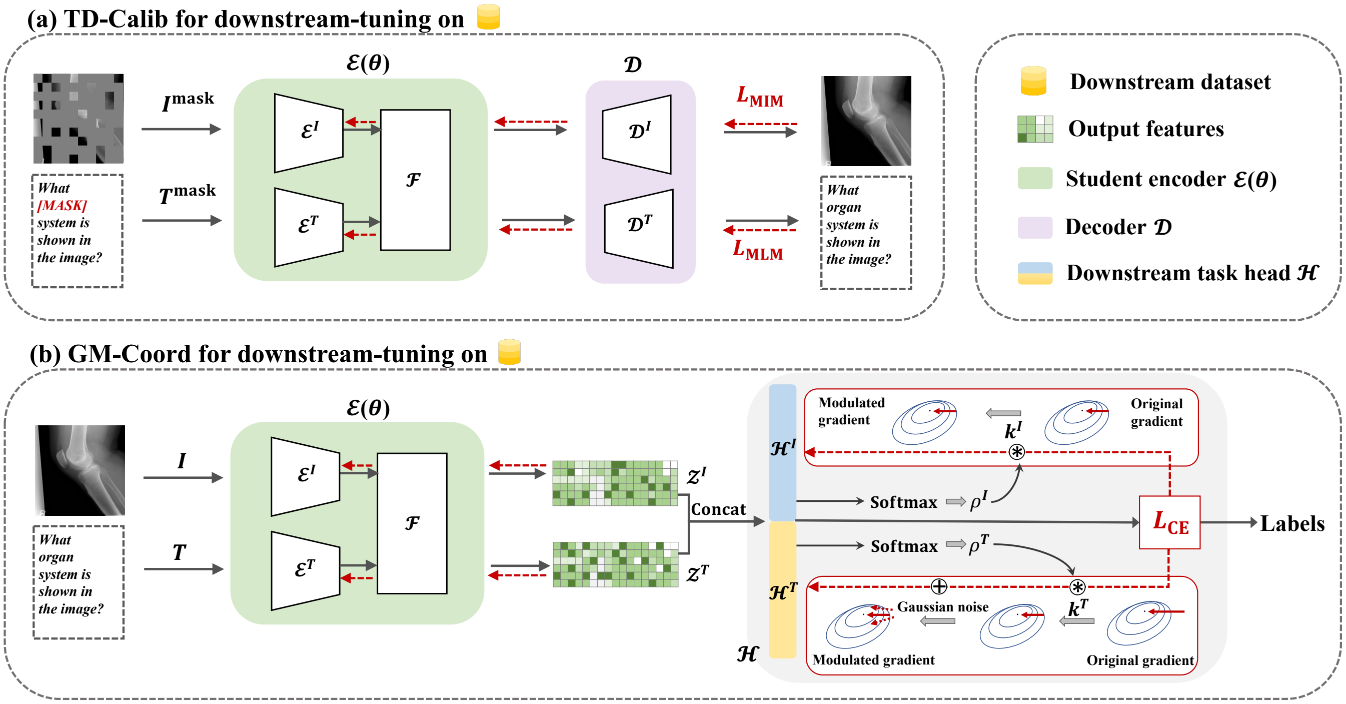


Fig. 4. Our heterogeneity-combat tuning facilitates medical diagnosis on downstream datasets. (a) The TD-Calib firstly calibrates the student multi-modal encoder to bridge the distribution gap, and then (b) the GM-Coord performs supervised fine-tuning to balance the modality optimization. For ease of understanding, we elaborate on the case of $\rho^T > 1$, where the gradient of the language modality should be modulated, as shown in (b).

is pre-trained to learn image and text representation at multiple levels. In this way, our MR-Pretrain can improve the model's transferable representation for diverse downstream tasks, by leveraging the unlabeled pre-training data comprehensively.

C. Heterogeneity-Combat Downstream Tuning

The current *pretrain-finetune* paradigm [4], [16] directly fine-tunes the pre-trained model on the downstream datasets as illustrated in Fig. 2. However, this straightforward fine-tuning approach neglects the heterogeneity between pre-training and downstream datasets, as well as the modality heterogeneity within downstream optimization, resulting in sub-optimal performance on the specific dataset. To tackle these two challenges, we propose the heterogeneity-combat downstream tuning in Fig. III-B.2, including TD-Calid to automatically calibrate the pre-trained encoder for a particular downstream dataset, and GM-Coord to dynamically balance the optimization of different modalities.

1) Task-Oriented Distribution Calibration: The distribution-heterogeneity inherently exists in pre-training and downstream data. A well-trained encoder carries abundant knowledge from the pre-training dataset, while how to efficiently transfer this pre-trained knowledge is an open-air question. Compared with the direct fine-tuning that is insufficient as the incoherent knowledge transfer from pre-training to downstream tuning, our TD-Calib module aims for a coherent transfer, which bridges the data distribution gap between the pre-training and the downstream tuning. To this end, the pre-trained model is further trained on the downstream datasets by reconstructing the masked multi-modal data, as shown in Fig. III-B.2 (a). This enables the model to adapt to the new distributions without explicit instruction of the ground truth, thus facilitating downstream objectives. Thus, we introduce the consistent pre-training objectives as TD-Calib training objectives, as follows:

$$\theta^*, \theta_1^*, \dots, \theta_S^* = \underset{\theta,\theta_1,\dots,\theta_S}{\arg\min} \sum_{s=1}^{S} L_s(Y_s, \mathcal{D}_s(\mathcal{E}(I^{\text{mask}}, T^{\text{mask}}; \theta_s))), \tag{8}$$

where Y_s represents the reconstruction targets of the masked image or text inputs, L_s is the training objectives of the TD-Calib module, S is the total number of training objectives that are empirically set as 4, and s is the index of each training objective. In contrast to directly fine-tuning, our model optimized with Eq. (8) not only adapts to the data distribution of the downstream domain, but also leverages the pre-trained knowledge to a greater extent.

2) Gradient-Guided Modality Coordination: Due to the modality heterogeneity, the optimization imbalance phenomenon exists in the joint training of multi-modal data, where the dominant modality suppresses the optimization of the other modalities during training. This phenomenon impacts the performance of medical multi-modal downstream tasks, such as visual question-answering and image-text classification tasks. To tackle this problem, we introduce the GM-Coord module to coordinate the optimization of each modality under the guidance of gradient changes, as illustrated in Fig. III-B.2 (b). During the GM-Coord, the contribution discrepancy among images and texts toward the learning objective is continuously monitored. The information is then utilized to adaptively modulate the gradients, thereby allocating more significant optimization updates to the suppressed modality.

For each multi-modal downstream task, the GM-Coord calculates the contribution C^I of vision modality and C^T of language modality. We split the task head $\mathcal{H} = [\mathcal{H}^I, \mathcal{H}^T]$ to separately measure the contribution of each modality, which is supervised by downstream task supervision. The calculation can be formulated with $u \in \{I, T\}$ for vision or language modality, as follows:

$$Z^I, Z^T = \mathcal{E}(I, T; \theta), \tag{9}$$

$$C^u = \text{softmax}(\mathcal{H}^u(Z^u))[j], \tag{10}$$

where C^u and $\mathcal{H}^u$ represent the contribution and task head for a specific modality respectively, and $[j]$ denotes a selection operator on the j-th class and j means the ground truth. In this way, C^I means the prediction score on the ground truth class for the vision modality and C^T means the prediction score on the ground truth class for the language modality. To quantify the optimization status of each modality, ρ^I and ρ^T are computed by the modality contribution, as follows:

$$\rho^I = \frac{\sum C^I}{\sum C^T}, \quad \rho^T = \frac{\sum C^T}{\sum C^I}. \tag{11}$$

Then, we modulate the gradient of the modality that is optimizing fast. Taking the language modality as an example, the coordination coefficient k^T is as follows:

$$k^T = \begin{cases} 1 - \tanh(\beta \cdot \rho^T), & \rho^T > \rho^I \\ 1, & \text{others} \end{cases} \tag{12}$$

where the factor β controls the degree of modulation and is set to 0.1. As such, the coordination coefficient $k^T < 1$ when the optimization rate of one modality is higher than another, resulting in a reduction in the optimization speed of the faster modality. Finally, we integrate the coefficient k^T into optimization together with Gaussian noise $\sigma(\theta) \sim \mathcal{N}(0, \Sigma^2_{\nabla\theta})$, where $\Sigma^2_{\nabla\theta}$ represents the variance of the parameters' gradient, to improve the generalization ability. The modulated gradient $\tilde{g}$ of GM-Coord is as follows:

$$\tilde{g}(\theta^T) \leftarrow k^T \tilde{g}(\theta^T) + \sigma(\theta^T). \tag{13}$$

Similarly, the gradients of vision modality $\tilde{g}(\theta^I)$ are also modulated following Eq. (12) and Eq. (13) if vision optimization is faster than the language optimization. As such, the modulated gradients lead to balanced optimization of multi-modal data, thereby facilitating the performance of downstream tasks.

D. Algorithm Pipeline

The training pipeline of our UMD framework is summarized in Algorithm 1, which includes the MR-Pretrain and heterogeneity-combat downstream tuning. We first perform the MR-Pretrain using Eq. (7) on unannotated data, and obtain a pre-trained model that can generate general feature representations. Then, we conduct the TD-Calib in heterogeneity-combat downstream tuning using Eq. (8), which promotes the pre-trained model's smooth adaptation to downstream datasets. Finally, we perform the optimization of GM-Coord together with downstream objectives, enabling the model to capture semantic features of multi-modal data. The source code is available at https://github.com/helenypzhang/UMD.

Algorithm 1 The pipeline of UMD

```
Input: Paired images and texts for Pre-training $\{I^P, T^P\}$ and for
    Downstream Tuning $\{I^D, T^D\}$; Model parameters $\Theta = \{\theta\} \cup \{\theta_s\}_{s=1}^S$;
    Random masking $M_I(\cdot)$ for images and $M_T(\cdot)$ for texts.
Output: The trained optimal parameters $\Theta^*$
    {MR-Pretrain}
while $\Theta$ doesn't reach convergence do
    for each $I$ and $T \in \{I^P, T^P\}$ do
        $I^{\text{mask}} \leftarrow M_I(I)$; $T^{\text{mask}} \leftarrow M_T(T)$
        Minimize $L_{\text{pretrain}}$ using Eq. (7)
    end for
end while
The $\mathcal{E} \circ \mathcal{D}$ is well-trained for MR-Pretrain stage.
    {TD-Calib}
while $\Theta$ doesn't reach convergence do
    for each $I$ and $T \in \{I^D, T^D\}$ do
        $I^{\text{mask}} \leftarrow M_I(I)$; $T^{\text{mask}} \leftarrow M_T(T)$
        Minimize $\sum_{s=1}^S L_s(Y_s, \mathcal{D}_s(\mathcal{E}(I^{\text{mask}}, T^{\text{mask}}; \theta))$
    end for
end while
The $\mathcal{E} \circ \mathcal{D}$ is well-trained for TD-Calib module.
    {GM-Coord}
while $\Theta$ doesn't reach convergence do
    for each $I$ and $T \in \{I^D, T^D\}$ do
        Minimize $L_{\text{CE}}(Y, \mathcal{H}(\mathcal{E}(I, T; \theta))$
        Calculate $C^u$, $\rho^u$, $k^u$, $\tilde{g}(\theta^u)$ in Eq. (10), (11), (12), (13)
    end for
end while
The $\mathcal{E} \circ \mathcal{H}$ is well-trained for GM-Coord module.
```

IV. Experiment

A. Dataset

We pre-train the model in our UMD framework using MedICaT [33] and ROCO [32] datasets and conduct the fine-tuning experiments on three downstream tasks, including three visual question-answering (VQA) datasets, one image-text retrieval dataset, and one image-text classification dataset.

1) Pre-Training Datasets: In our experimental setup, we conduct self-supervised pre-training on two datasets, *i.e.*, MedICaT [33] and ROCO [32] dataset.

MedICaT dataset [33] comprises more than 217,000 medical images and their corresponding captions and inline textual references. Following M^3AE [6], we randomly allocate 1,000 samples for test, 1,000 for validation, and the remaining data for training purposes.

ROCO dataset [32] contains more than 81,000 medical radiology images, encompassing a variety of imaging modalities such as Computed Tomography (CT), X-ray, ultrasound, fluoroscopy, angiography, mammography, positron emission tomography, and Magnetic Resonance Imaging (MRI). Each image is accompanied by a corresponding caption. We follow the dataset splits in ROCO [32], with over 65,000 radiology images to the training set, over 8,000 radiology images to the validation set, and over 8,000 radiology images to the test set.

2) Downstream Tuning Datasets: We evaluate the effectiveness of our pre-training approach by conducting experiments on the VQA, image-text retrieval tasks, and image-text classification, utilizing the official split of each dataset in downstream experiments. For the VQA task, we select three public datasets,

TABLE I
COMPARISON WITH STATE-OF-THE-ART ALGORITHMS ON THREE MEDICAL VQA DATASETS REGARDING ACCURACY METRIC. BEST AND SECOND RESULTS ARE HIGHLIGHTED WITH **BOLD** AND UNDERLINE.

Methods	VQA-RAD			SLAKE			VQA-Med-2019
	Open	Closed	Overall	Open	Closed	Overall	Overall
MFB [45]	14.50	74.30	50.60	72.20	75.00	73.30	-
SAN [46]	31.30	69.50	54.30	74.00	79.10	76.00	-
BAN [47]	37.40	72.10	58.30	74.60	79.10	76.30	-
MEVF-BAN [25]	49.20	77.20	66.10	77.80	79.80	78.60	77.86
CPRD-BAN [48]	52.50	77.90	67.80	79.50	83.40	81.10	-
MAE [15]	67.04	77.94	73.61	77.21	82.21	79.17	73.60
CLIP [7]	64.80	79.78	73.84	78.45	84.62	80.87	76.80
FLIP [11]	65.92	78.31	73.39	80.47	84.86	82.19	78.40
M^{3}AE [6]	67.23	83.46	77.01	80.31	87.82	83.25	79.87
UMD	**68.16**	**85.66**	**78.71**	**82.17**	**88.70**	**84.73**	**80.53**

i.e., VQA-RAD [49], SLAKE [50], and VQA-Med-2019 [51]. Moreover, the ROCO dataset [32] and MELINDA dataset [52] are utilized in image-text retrieval and image-text classification tasks, respectively.

VQA-RAD dataset [49] includes 315 images, consisting of 104 axial single-slice CTs or MRIs for head, 107 X-rays for chest, and 104 axial CTs for abdomen, each accompanied by corresponding captions. There are over 3.5K visual questions in VQA-RAD, including open-ended and closed-ended answer types. In particular, there are 3,064 question-answer pairs in training set, 451 question-answer pairs in validation set, and 451 question-answer pairs in test set.

SLAKE dataset [50] comprises 642 multi-modal images covering 12 diseases and 39 organs of the human body to ensure dataset diversity. The question-answer pairs are 14K. Both open-ended and closed-ended answer types are included in the SLAKE and VQA-RAD datasets, determined by whether the answer choices are limited or not. In particular, the dataset is divided into training, validation, and test sets with the ratio of 75%, 15% and 15%.

VQA-Med-2019 dataset [51] is composed of 4,200 radiological images and 15,292 question-answer pairs. The dataset is split into training set with 3,200 images, validation set with 500 images, and test set with 500 images.

ROCO dataset [32] is utilized on the image-text retrieval task. The image-text retrieval task comprises two subtasks: image-to-text retrieval and text-to-image retrieval. The former aims to retrieve the most relevant texts based on the given image, while the latter aims to retrieve the most relevant images based on the given text.

MELINDA dataset [52], which contains 2,833 figures paired with corresponding detailed sub-figures and sub-captions, is utilized on the image-text classification task. The dataset is split into train, validation, and test sets, with the ratio of 80%, 10% and 10%.

B. Implementation Details

Our experiments are implemented with the PyTorch Lightning library [55] on three NVIDIA A100 PCIe 40 GB GPUs. Details of each task are elaborated as follows.

1) Multi-Level Reconstruction Pre-Training: For MR-Pretrain, we train $\mathcal{E} \circ \mathcal{D}$ end-to-end. We start from the CLIP-ViT-B model [7] as the vision encoder, the RoBERTa-base [38] as the language encoder, with the multi-modal fusion module provided by $\mathrm{M}^3\mathrm{AE}$ [6]. The multi-modal module consists of 6 Transformer layers with a hidden state dimension of 768 and 12 heads. We use AdamW optimizer [56] to train the models for 100,000 steps, with a learning rate of 1×10^{-5} for the uni-modal encoders and 5×10^{-5} for the multi-modal fusion module. We set the warm-up ratio to 10%, with a linear learning rate scheduler after warm-up. To resize each image, we use a center-crop method with a size of 314×314. The trade-off factor α in MR-Pretain is set as 0.5. The smoothing factor λ of EMA is 0.995 for weight updating.

2) Heterogeneity-combat Downstream Tuning: For TD-Calib, we fine-tune the $\mathcal{E} \circ \mathcal{D}$ end-to-end. In order to bridge the data distribution gap between pre-training and fine-tuning, we conduct TD-Calib-guided downstream tuning. Specifically, we initialize the multi-modal encoder with the pre-trained weights, and feed images and texts to the model to further pre-train both $\mathcal{E}$ and $\mathcal{D}$. The masking ratio is set to 75% for images and 15% for texts. Moreover, the warm-up ratio is 10%, with a linear learning rate scheduler used after warm-up steps. We use AdamW as the optimizer with a weight decay of 0.01 for all downstream tasks. The initial learning rate for VQA-RAD, SLAKE, VQA-Med-2019, ROCO, and MELINDA is set to 1×10^{-5}, 5×10^{-6}, 5×10^{-6}, 1×10^{-5} and 1×10^{-5}, respectively, and linearly decay to zero during training.

For GM-Coord, we fine-tune $\mathcal{E} \circ \mathcal{H}$ end-to-end. Specifically, we further fine-tune the multi-modal encoder $\mathcal{E}$ optimized by the TD-Calib module under different downstream tasks, together with the downstream task-specific head $\mathcal{H}$. For each downstream task, we guarantee the fairness of the experiment by adopting the same $\mathcal{H}$ for different comparison methods. We utilize the AdamW optimizer with an initial learning rate of 5×10^{-6}, a warm-up ratio of 10%, and a linear decay during training for VQA-Med-2019, while for other downstream datasets, we use cosine decay. The weight decay is set to 0.01 for SLAKE and MELINDA datasets, and 0.1 for VQA-RAD, VQA-Med-2019 and ROCO datasets.

3) Evaluation Metric: To conduct a comprehensive evaluation, we analyze diverse performance metrics on different downstream tasks. We follow the previous study [6] to adopt

TABLE II
COMPARISON WITH STATE-OF-THE-ART ALGORITHMS ON MEDICAL IMAGE-TEXT RETRIEVAL TASK ON ROCO DATASET. BEST AND SECOND RESULTS ARE HIGHLIGHTED WITH **BOLD** AND $\underline{\text{UNDERLINE}}$.

Methods	Text-to-image retrieval			Image-to-text retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
ViT+BERT [53]	5.25	15.85	25.85	6.85	21.25	31.60
ViLT [54]	9.75	28.95	41.40	11.90	31.90	43.20
METER [53]	11.30	27.25	39.60	14.45	33.30	45.10
MAE [15]	4.35	17.96	28.96	4.95	18.31	28.06
CLIP [7]	14.41	39.67	54.68	17.61	42.92	57.98
FLIP [11]	17.66	46.62	61.03	17.46	45.57	61.53
M^{3}AE [6]	$\underline{22.20}$	$\underline{52.50}$	$\underline{66.65}$	$\underline{22.90}$	$\underline{51.05}$	$\underline{65.80}$
UMD	**23.21**	**54.28**	**67.88**	**24.39**	**54.27**	**68.97**

TABLE III
COMPARISON WITH STATE-OF-THE-ART ALGORITHMS ON MEDICAL IMAGE-TEXT CLASSIFICATION TASK ON MELINDA DATASET. BEST AND SECOND RESULTS ARE HIGHLIGHTED WITH **BOLD** AND $\underline{\text{UNDERLINE}}$.

Modalities	Methods	Accuracy
Image-only	ResNet-101 [57]	63.84
Text-only	LSTM [58]	59.20
	RoBERTa [38]	75.40
	SciBERT [59]	77.70
Multi-modal	NLF [52]	76.60
	SAN [46]	72.30
	ViLBERT [60]	$\underline{78.60}$
	MAE [15]	78.03
	CLIP [7]	77.16
	FLIP [11]	77.36
	M^{3}AE [6]	78.50
	UMD	**79.58**

the accuracy for the VQA and image-text classification tasks, and Recall@K with K=1, 5 and 10 for the image-text retrieval task, respectively. In VQA and image-text classification, the *Overall* term specifically refers to the micro-average accuracy of both open-ended and closed-ended questions. In addition, the Recall@K, commonly used in information retrieval tasks, is an evaluation metric that measures the proportion of relevant items that are retrieved in the top K results. In other words, the Recall@K measures how many of the relevant items are actually retrieved in the top K results. We conduct experiments for Recall@K with K=1, 5 and 10, which represent the proportion of relevant items retrieved in different predictions.

C. Downstream Experiments

1) Medical Visual Question-Answering: The VQA is a multi-modal task that requires both images and questions as input, and is expected to answer questions about medical images. The VQA questions can belong to either open or closed categories, where open-category questions require the model to generate a free-form answer, while closed-category questions require the model to select a predefined answer from a set of options. Medical VQA task is particularly useful in medical diagnosis and treatment planning, where doctors often rely on visual information to make informed decisions. As shown in Table I, our UMD framework outperforms state-of-the-art models on all VQA datasets, achieving the accuracy of 68.16%, 85.66%, and 78.71% for VQA-RAD, 82.17%, 88.70%, and 84.73% for SLAKE, and 80.53% for VQA-Med-2019. In terms of *Overall* performance, our UMD framework outperforms the second-best method (marked in underline) by 1.70%, 1.48%, and 0.66% on three VQA datasets. These improvements contribute to the transferable features learned by our tailored MR-Pretrain and the heterogeneity-combat downstream tuning stages.

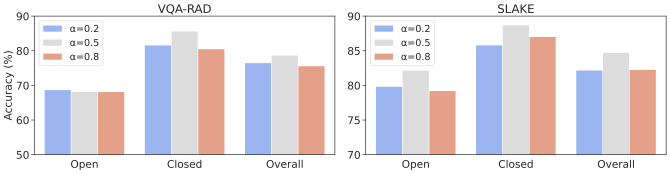


Fig. 5. Ablation study on the hyper-parameter α in MR-Pretrain. Our UMD framework achieves the best performance when α is set as 0.5.

We also conduct experiments using strong baselines of MAE, CLIP, and FLIP algorithms. The models are first pre-trained with different training objectives on the MedICaT and the ROCO datasets, and then fine-tuned with the same prediction head with the cross-entropy loss on VQA datasets. We adopt the same backbones and task heads to ensure fairness in the experiment. Compared with four strong baselines, two of which are masked autoencoder-based (*i.e.*, MAE and M^{3}AE), and the other two are contrastive learning-based pre-training methods (*i.e.*, CLIP and FLIP), our accuracy increases by 6.93%, 0.66%, 3.73%, and 2.13%, on the VQA-Med-2019 dataset, respectively. These results show that UMD is not only superior to masked autoencoder-based pre-training, but also outperforms other types of pre-training algorithms.

Hyper-parameters Analysis. We further conduct experiments on one of the most significant hyper-parameters, *i.e.*, α in Eq. (7), to investigate the trade-off between feature-level and data-level reconstruction in MR-Pretrain and TD-Calib. In our hyper-parameters study, we set α as 0.2, 0.5, and 0.8 both for VQA-RAD and SLAKE datasets. As illustrated in Fig. 5, our UMD framework achieves the best performance when α is set as 0.5, further demonstrating the rationality of the hyper-parameter setting in our UMD framework.

2) Medical Image-Text Retrieval: Image-text retrieval is a cross-modal task including medical image-to-text retrieval and medical text-to-image retrieval tasks, requiring the model to exploit useful information across modalities. The experimental results are presented in Table II. We perform comprehensive

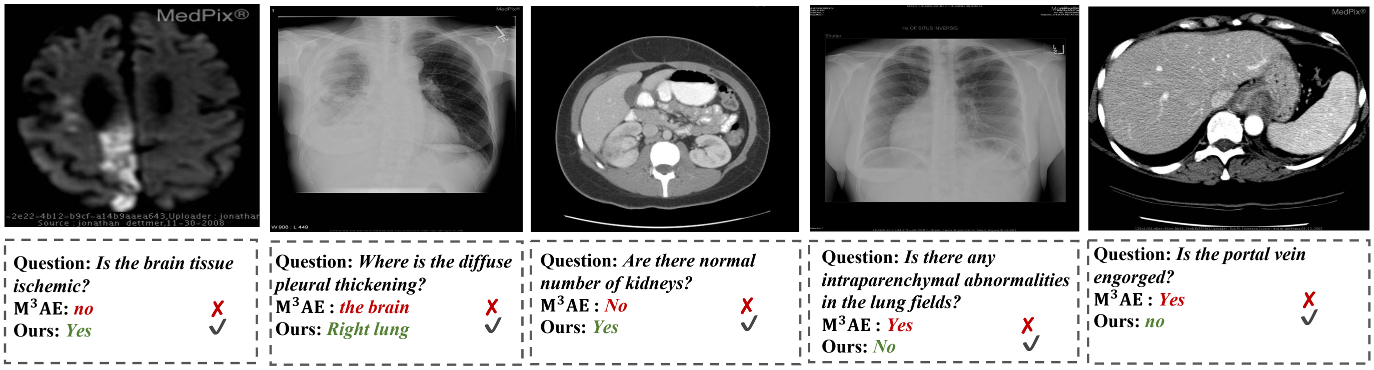

MedPix®
Question: Is the brain tissue ischemic?
M^{3}AE: no ✗
Ours: Yes ✓
Question: Where is the diffuse pleural thickening?
M^{3}AE : the brain ✗
Ours: Right lung ✓
Question: Are there normal number of kidneys?
M^{3}AE : No ✗
Ours: Yes ✓
Question: Is there any intraparenchymal abnormalities in the lung fields?
M^{3}AE : Yes ✗
Ours: No ✓
MedPix®
Question: Is the portal vein engorged?
M^{3}AE : Yes ✗
Ours: no ✓

Fig. 6. Visualization of medical VQA comparison on VQA-RAD dataset. Our UMD framework is capable of providing more accurate answers to medical questions of different difficulties.

TABLE IV
ABLATION STUDY OF UMD ON THREE MEDICAL VQA DATASETS.

	Pre-training	Fine-tuning		VQA-RAD			SLAKE			VQA-Med-2019
	MR-Pretrain	TD-Calib	GM-Coord	Open	Close	**Overall**	Open	Close	**Overall**	**Overall**
1				65.36	78.68	73.39	74.88	78.13	76.15	72.00
2	✓			**69.27**	82.72	77.38	81.86	86.06	83.51	77.07
3		✓		66.48	80.51	74.95	79.84	85.10	81.90	74.67
4			✓	68.16	80.51	75.61	77.36	86.54	80.96	76.00
5		✓	✓	67.60	81.99	76.27	80.00	86.54	82.56	77.87
6	✓	✓		68.16	84.19	77.83	81.40	87.74	83.88	79.47
7	✓		✓	**69.27**	83.82	78.05	**82.95**	86.54	84.35	80.23
8	✓	✓	✓	68.16	**85.66**	**78.71**	82.17	**88.70**	**84.73**	**80.53**

comparisons with state-of-the-art methods, including ViLT, METER, MAE, CLIP, FLIP, and M^{3}AE. The results show that our UMD framework achieves the best R@K (K=1,5 and 10) performances of 23.21%, 54.28% and 67.88% for text-to-image retrieval, and 24.39%, 54.27%, and 68.97% for image-to-text retrieval task. UMD surpasses the second-best one by 1.78% and 3.22% in terms of R@5 text-to-image and image-to-text retrieval tasks respectively. These experimental results show the effectiveness of our UMD framework on medical image-text retrieval.

3) Medical Image-Text Classification: Image-text classification aims to give a label to a medical image-text pair, which also belongs to the multi-modal task. By training a model that can classify medical images associated with the text descriptions, this task is beneficial in medical research and clinical scenarios. Besides the baselines with image-only data and text-only data, we perform the comparison with advanced multi-modal methods ViLBERT, CLIP, and FLIP in the general domain, and M^{3}AE in the medical multi-modal domain. As shown in Table III, our UMD framework achieves the best accuracy of 79.58% on the MELINDA dataset, outperforming the second-best ViLBERT by 0.98%. The remarkable advantage in image-text classification demonstrates the effectiveness of our UMD framework on medical multi-modal data.

D. Ablation Study

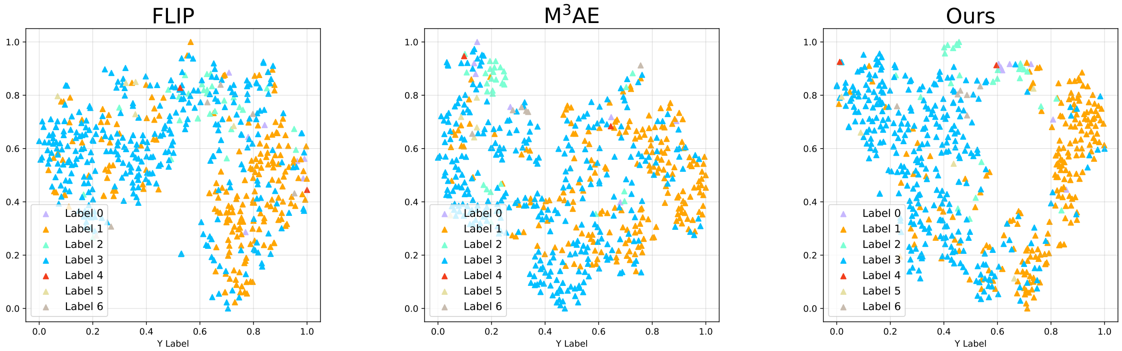
To quantitatively evaluate the effectiveness of our proposed components, *i.e.*, MR-Pretrain, TD-Calib and GM-Coord, we conduct ablation studies for each component on three VQA datasets and one image-text classification dataset. The ablation results are illustrated in Table IV and Table V.

TABLE V
ABLATION STUDY OF UMD ON MELINDA DATASET.

	Pre-training	Fine-tuning		Accuracy	AUC	Sensitivity	Specificity
	MR-Pretrain	TD-Calib	GM-Coord				
1				66.09	78.20	27.16	90.86
2	✓			78.55	85.48	34.59	94.16
3		✓		74.91	86.98	32.23	92.94
4			✓	69.72	81.74	31.11	91.70
5		✓	✓	75.26	84.28	33.83	93.49
6	✓	✓		78.72	89.13	34.16	94.14
7	✓		✓	79.24	89.39	34.59	94.27
8	✓	✓	✓	**79.58**	**91.52**	**36.00**	**94.56**

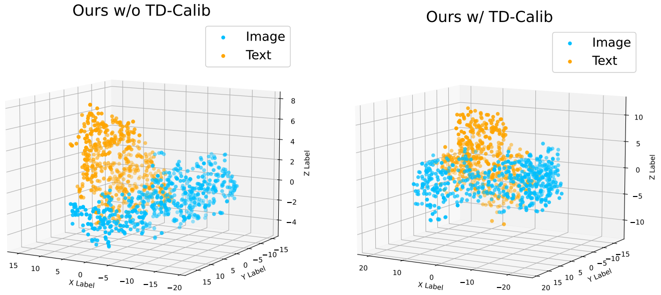
- Line 1: The baseline simply trains the multi-modal encoder and downstream task head from scratch, without relying on pre-trained models or the proposed components. This baseline serves as a performance lower bound for VQA and medical image-text classification.
- Line 2-4: The proposed components (*i.e.*, MR-Pretrain, TD-Calib without pre-training, and GM-Coord without pre-training) are individually added on the basis of the baseline (Line 1). These records can validate the independent impact of these three components.
- Line 5-8: The possible combination of the proposed three components. These records are crucial for the ablation study of MR-Pretrain, TD-Calib and GM-Coord.

For the VQA tasks in Table IV, the MR-Pretrain model (Line 2) achieves the *Overall* accuracy of 77.38%, 83.51%, and 77.07% for the VQA-RAD, SLAKE and VQA-Med-2019 datasets, respectively, which exhibits 3.99% for VQA-RAD, 7.36% for SLAKE, and 5.07% for VQA-Med-2019 increase compared with the baseline (Line 1). These improvements



(a) 2D Clustering of FLIP (b) 2D Clustering of $\mathrm{M}^3\mathrm{AE}$ (c) 2D Clustering of Ours (UMD)

Fig. 7. Visualization of feature representation using (a) FLIP (b) $\mathbf{M^3AE}$, and (b) our UMD on the MELINDA dataset. Our UMD demonstrates a clearer clustering of data, which is beneficial for multi-modal classification.



(a) 3D Clustering without TD-Calib (b) 3D Clustering with TD-Calib

Fig. 8. Visualization of image and text embeddings in our UMD framework (a) without or (b) with TD-Calib on MELINDA dataset. The TD-Calib in (b) makes the fusion of multi-modal more adequate.

can be attributed to the transferable weights learned by the multi-modal encoder model. Moreover, the *Overall* accuracy increase (Lines 2-4) compared with the baseline (Line 1) in all types of questions on three VQA datasets verifies the effectiveness of the three proposed components (*i.e.*, MR-Pretrain, TD-Calib, GM-Coord). Furthermore, the *Overall* performance of the model with two components (Lines 5-7) is better than the results of the model with one component (Lines 2-4), which confirms the complementary enhancement of the proposed components. In addition, the complete UMD framework (Line 8) achieves the best *Overall* performance, validating the effectiveness of our UMD framework.

For the medical image-text classification task, we perform the ablation study of UMD to investigate the combination of pre-training and fine-tuning techniques with various metrics on the MELINDA datasets, as shown in Table V. Similar to the conclusion in Table IV, the results in Lines 2-4 in Table V outperform the baseline (Line 1), which demonstrates each of our designs is rational. Especially, the MR-Pretrain improves by 12.46% compared with the baseline (Line 1), showing the effectiveness of our pre-training method. Furthermore, compared with models with a single design (Lines 2-4), models of pairwise combination (Lines 5-7) deliver higher performance, which demonstrates the complementarity of the proposed three components. Finally, when all three proposed components are applied (Line 8), our UMD framework achieves the best performance. The ablation study verifies the effectiveness of our MR-Pretrain, TD-Calib and GM-Coord modules.

E. Qualitative Analysis

For a qualitative comparison, we further present 5 VQA test samples from the VQA-RAD dataset to provide predicted results of $\mathrm{M}^3\mathrm{AE}$ and our UMD framework, as shown in Fig. 6. Compared with $\mathrm{M}^3\mathrm{AE}$, UMD can understand diagnosis-related information better and predict more accurate answers, which can benefit clinical diagnosis more effectively.

Furthermore, we visualize the t-SNE features [61] of randomly sampled cases in the MELINDA dataset, as depicted in Fig. 8, where the blue and orange points represent image features Z^I and text features Z^T, respectively. By comparing (a) and (b) in Fig. 8, we observe that the fusion of two modalities is more adequate in the case with TD-Calib. This finding highlights another benefit of TD-Calib by enhancing modality fusion, which explains from another perspective why TD-Calib can contribute to various multi-modal downstream tasks. Additionally, we perform the 2D clustering of FLIP, $\mathrm{M}^3\mathrm{AE}$, and our UMD framework, and visualize the results in Fig. 7. The different colors represent different categories in the MELINDA dataset. The comparison between (a), (b), and (c) in Fig. 7 indicates that our UMD framework can separate the categories more distinctly.

V. Conclusion

In this work, we propose the Unified Medical Multi-modal Diagnostic (UMD) framework, which utilizes unlabeled multi-modal medical datasets to enhance the representation learning of deep learning models in a self-supervised manner. Specifically, we devise a novel MR-Pretrain strategy, which guides models to capture semantic information from masked inputs of various modalities through feature-level and data-level reconstruction. Moreover, to tackle the distribution heterogeneity between pre-training and downstream data and the modality heterogeneity within downstream datasets, we present a heterogeneity-combat downstream tuning strategy, including the TD-Calib and the GM-Coord. In particular, the TD-Calib fine-tunes the pre-trained model based on the distribution of the downstream datasets, while GM-Coord adjusts the gradient weights according to the dynamic optimization status of different modalities. Extensive experiments on five public medical datasets demonstrate the effectiveness of our UMD framework, which outperforms state-of-the-arts on three kinds of downstream tasks by a remarkable margin.

References

[1] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, K. Moore, R. S. Mannel, H. Liu, B. Zheng, and Y. Qiu, "Recent advances and clinical applications of deep learning in medical image analysis," *Med. Image Anal.*, vol. 79, p. 102444, 2022.

[2] X. Xie, J. Niu, X. Liu, Z. Chen, S. Tang, and S. Yu, "A survey on incorporating domain knowledge into deep learning for medical image analysis," *Med. Image Anal.*, vol. 69, p. 101985, 2021.

[3] S. Wang, C. Li, R. Wang, Z. Liu, M. Wang, H. Tan, Y. Wu, X. Liu, H. Sun, R. Yang *et al.*, "Annotation-efficient deep learning for automatic medical image segmentation," *Nat. Commun.*, vol. 12, no. 1, p. 5915, 2021.

[4] S.-C. Huang, A. Pareek, M. Jensen, M. P. Lungren, S. Yeung, and A. S. Chaudhari, "Self-supervised learning for medical image classification: a systematic review and implementation guidelines," *NPJ Digital Medicine*, vol. 6, no. 1, p. 74, 2023.

[5] Á. S. Hervella, J. Rouco, J. Novo, and M. Ortega, "Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis," *Expert Syst. Appl.*, vol. 185, p. 115598, 2021.

[6] Z. Chen, Y. Du, J. Hu, Y. Liu, G. Li, X. Wan, and T.-H. Chang, "Multi-modal masked autoencoders for medical vision-and-language pre-training," in *MICCAI*. Springer, 2022, pp. 679–689.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *ICML*, 2021, pp. 8748–8763.

[8] J. Yang, J. Duan, S. Tran, Y. Xu, S. Chanda, L. Chen, B. Zeng, T. Chilimbi, and J. Huang, "Vision-language pre-training with triple contrastive learning," in *CVPR*, 2022, pp. 15 671–15 680.

[9] A. Singh, R. Hu, V. Goswami, G. Couairon, W. Galuba, M. Rohrbach, and D. Kiela, "Flava: A foundational language and vision alignment model," in *CVPR*, 2022, pp. 15 638–15 650.

[10] G. Kwon, Z. Cai, A. Ravichandran, E. Bas, R. Bhotika, and S. Soatto, "Masked vision and language modeling for multi-modal representation learning," *arXiv preprint arXiv:2208.02131*, 2022.

[11] Y. Li, H. Fan, R. Hu, C. Feichtenhofer, and K. He, "Scaling language-image pre-training via masking," in *CVPR*, 2023, pp. 23 390–23 400.

[12] G. Wang, K. Wang, G. Wang, P. H. Torr, and L. Lin, "Solving inefficiency of self-supervised representation learning," in *ICCV*, 2021, pp. 9505–9515.

[13] W. Wang, J. Wang, C. Chen, J. Jiao, L. Sun, Y. Cai, S. Song, and J. Li, "Fremae: Fourier transform meets masked autoencoders for medical image segmentation," *arXiv preprint arXiv:2304.10864*, 2023.

[14] Z. Qing, S. Zhang, Z. Huang, X. Wang, Y. Wang, Y. Lv, C. Gao, and N. Sang, "Mar: Masked autoencoders for efficient action recognition," *IEEE Transactions on Multimedia*, 2023.

[15] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *CVPR*, 2022, pp. 16 000–16 009.

[16] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *CVPR*, 2022, pp. 14 668–14 678.

[17] H. Wang, Y. Tang, Y. Wang, J. Guo, Z.-H. Deng, and K. Han, "Masked image modeling with local multi-scale reconstruction," in *CVPR*, 2023, pp. 2122–2131.

[18] D. G. Lowe, "Object recognition from local scale-invariant features," in *ICCV*, vol. 2, 1999, pp. 1150–1157.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, vol. 1, 2005, pp. 886–893.

[20] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.

[21] Y. Gandelsman, Y. Sun, X. Chen, and A. Efros, "Test-time training with masked autoencoders," *NeurIPS*, vol. 35, pp. 29 374–29 385, 2022.

[22] A. Kumar, A. Raghunathan, R. Jones, T. Ma, and P. Liang, "Fine-tuning can distort pretrained features and underperform out-of-distribution," *arXiv preprint arXiv:2202.10054*, 2022.

[23] Y. Li, H. Wang, and Y. Luo, "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports," in *BIBM*. IEEE, 2020, pp. 1999–2004.

[24] R. Chang, Y.-X. Wang, and E. Ertekin, "Towards overcoming data scarcity in materials science: unifying models and datasets with a mixture of experts framework," *npj Computational Materials*, vol. 8, no. 1, p. 242, 2022.

[25] B. D. Nguyen, T.-T. Do, B. X. Nguyen, T. Do, E. Tjiputra, and Q. D. Tran, "Overcoming data limitation in medical visual question answering," in *MICCAI*. Springer, 2019, pp. 522–530.

[26] W. Su, X. Zhu, C. Tao, L. Lu, B. Li, G. Huang, Y. Qiao, X. Wang, J. Zhou, and J. Dai, "Towards all-in-one pre-training via maximizing multi-modal mutual information," in *CVPR*, 2023, pp. 15 888–15 899.

[27] W. Wang, D. Tran, and M. Feiszli, "What makes training multi-modal classification networks hard?" in *CVPR*, 2020, pp. 12 695–12 705.

[28] X. Peng, Y. Wei, A. Deng, D. Wang, and D. Hu, "Balanced multimodal learning via on-the-fly gradient modulation," in *CVPR*, 2022, pp. 8238–8247.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[30] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," *arXiv preprint arXiv:2301.12597*, 2023.

[31] S. Ren, F. Wei, S. Albanie, Z. Zhang, and H. Hu, "Deepmim: Deep supervision for masked image modeling," *arXiv preprint arXiv:2303.08817*, 2023.

[32] O. Pelka, S. Koitka, J. Rückert, F. Nensa, and C. M. Friedrich, "Radiology objects in context (roco): a multimodal image dataset," in *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis*. Springer, 2018, pp. 180–189.

[33] S. Subramanian, L. L. Wang, S. Mehta, B. Bogin, M. van Zuylen, S. Parasa, S. Singh, M. Gardner, and H. Hajishirzi, "Medicat: A dataset of medical images, captions, and textual references," *arXiv preprint arXiv:2010.06000*, 2020.

[34] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "Mmbert: Multimodal bert pretraining for improved medical vqa," in *ISBI*. IEEE, 2021, pp. 1033–1036.

[35] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *MLHC*, 2022, pp. 2–25.

[36] M. Endo, K. L. Poston, E. V. Sullivan, L. Fei-Fei, K. M. Pohl, and E. Adeli, "Gaitforemer: Self-supervised pre-training of transformers via human motion forecasting for few-shot gait impairment severity estimation," in *MICCAI*. Springer, 2022, pp. 130–139.

[37] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE J Biomed Health Inform*, vol. 26, no. 12, pp. 6070–6080, 2022.

[38] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[39] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *NeurIPS*, vol. 30, 2017.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[41] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

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

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*, 2020, pp. 1597–1607.

[44] T. Liu, Z. Wu, W. Xiong, J. Chen, and Y.-G. Jiang, "Unified multimodal pre-training and prompt-based tuning for vision-language understanding and generation," *arXiv preprint arXiv:2112.05587*, 2021.

[45] Z. Yu, J. Yu, J. Fan, and D. Tao, "Multi-modal factorized bilinear pooling with co-attention learning for visual question answering," in *ICCV*, 2017, pp. 1821–1830.

[46] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, “Stacked attention networks for image question answering,” in *CVPR*, 2016, pp. 21–29.

[47] J.-H. Kim, J. Jun, and B.-T. Zhang, “Bilinear attention networks,” *NeurIPS*, vol. 31, 2018.

[48] B. Liu, L.-M. Zhan, and X.-M. Wu, “Contrastive pre-training and representation distillation for medical visual question answering based on radiology images,” in *MICCAI*. Springer, 2021, pp. 210–220.

[49] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman, “A dataset of clinically generated visual questions and answers about radiology images,” *Sci. Data*, vol. 5, no. 1, pp. 1–10, 2018.

[50] B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu, “Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering,” in *ISBI*. IEEE, 2021, pp. 1650–1654.

[51] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, “Vqa-med: Overview of the medical visual question answering task at imageclef 2019,” in *CLEF*, 2019.

[52] T.-L. Wu, S. Singh, S. Paul, G. Burns, and N. Peng, “Melinda: A multimodal dataset for biomedical experiment method classification,” in *AAAI*, vol. 35, no. 16, 2021, pp. 14 076–14 084.

[53] Z.-Y. Dou, Y. Xu, Z. Gan, J. Wang, S. Wang, L. Wang, C. Zhu, P. Zhang, L. Yuan, N. Peng *et al.*, “An empirical study of training end-to-end vision-and-language transformers,” in *CVPR*, 2022, pp. 18 166–18 176.

[54] W. Kim, B. Son, and I. Kim, “Vilt: Vision-and-language transformer without convolution or region supervision,” in *ICML*, 2021, pp. 5583–5594.

[55] W. A. Falcon, “Pytorch lightning,” *GitHub*, vol. 3, 2019.

[56] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” *arXiv preprint arXiv:1711.05101*, 2017.

[57] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.

[58] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[59] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019.

[60] J. Lu, D. Batra, D. Parikh, and S. Lee, “Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks,” *NeurIPS*, vol. 32, 2019.

[61] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.” *J. Mach. Learn. Res.*, vol. 9, no. 11, 2008.