

MVCM: Enhancing Multi-View and Cross-Modality Alignment for Medical Visual Question Answering and Medical Image-Text Retrieval

Yuanhao Zou
University of Michigan, Ann Arbor
yuanhaoz@umich.edu

Zhaozheng Yin
Stony Brook University
zyin@cs.stonybrook.edu

Abstract

Recent advancements in medical vision-language tasks, such as Medical Visual Question Answering (Med-VQA) and Medical Image-Text Retrieval (Med-ITR), aim to jointly learn from images and texts. However, two main issues persist in the field: the neglect of multi-view medical images and incomplete cross-modality understanding. Current studies often treat each image-text pair as independent instances (i.e., at the instance-level), neglecting the comprehensive contextual information available from multi-view images of the same study. Although some methods have explored refined alignments, combining the alignment of global representation with the token-wise alignment of local representations, they often utilize only a uni-modality encoder (e.g., visual encoder) for downstream applications, lacking a comprehensive cross-modality understanding. To address these issues, this paper introduces a framework MVCM that supports Multi-View and Cross-Modality alignment for Med-VQA and Med-ITR tasks. Our proposed method fully utilizes multi-view images in radiology datasets and aligns them at the study-level. We also employ various pretext tasks to support cross-modality alignment. We fine-tune the proposed model on downstream tasks Med-VQA and Med-ITR, outperforming state-of-the-art methods across multiple datasets. The code is available at <https://github.com/AlexCold/MVCM>.

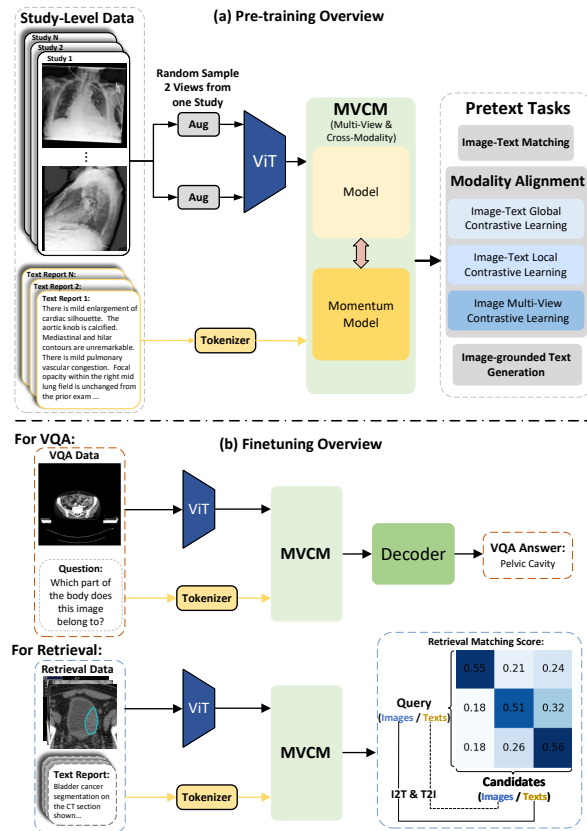


Figure 1. Overview of MVCM for pre-training and fine-tuning.

1. Introduction

Deep learning has significantly advanced image understanding, particularly when large-scale labeled datasets are available for training [19, 45, 46]. However, for some tasks like visual question answering (VQA) and image text retrieval (ITR) that require multi-modal understanding, especially vision and language understanding, obtaining well-annotated datasets is labor-intensive [4]. One way to address this challenge is Vision-Language Pre-training (VLP, [32–34, 44]), which strategically pre-trains on unlabeled

datasets via pretext tasks before fine-tuning for specific downstream tasks. VLP learns the multi-modal knowledge by two main types of pretext tasks, generative way [12, 22, 40] and contrastive way [17, 44]. Generative way attempts to reconstruct the masked information based on the remaining information. Contrastive way, tries to use contrastive learning to maximize the mutual information between images and texts. The two types of VLPs are also combined [31–33] for comprehensive understanding. Recently, VLP is applied to medical domains, yielding Medi-

cal Vision-Language Pre-training (**Med-VLP**).

Med-VLP studies [9, 25, 48, 53] have limitations in fully utilizing the medical image-report datasets. For example, a single radiology report corresponds to a complete *study* that includes multiple image views. However, existing Med-VLP studies commonly treat each image-text pair as a unique sample during training, so the model is trained at *instance-level* (e.g., image-text pairs are constructed based on single-view image and its corresponding text report, without including other views). Some studies like [48] attempt to align image-text information at the disease level. However, they still rely on instance-level datasets where each image-text pair consists of only frontal view image. Therefore, the single-view based methods lead to inefficient utilization of valuable medical image-report data and incomplete medical cross-modality understanding.

Besides, to achieve cross-modality understanding and address more complex downstream tasks such as Medical Visual Question Answering (**Med-VQA**) and Medical Image-Text Retrieval (**Med-ITR**), several studies [7, 8] attempt to fuse multi-modal information via various pretext tasks. Meanwhile, inspired by [24, 31], other researches like [35, 36] not only implement modality fusion but also explore the use of Momentum Distillation (**MoD**) across pretext tasks. Nevertheless, despite their attempts to apply MoD to the contrastive pretext tasks, they only focus on aligning the global representations between image and text, while ignoring other refined alignments. Meanwhile, although studies like [25, 41, 48] explore refined local-level alignment between image tokens and text tokens, they ultimately utilize only the visual encoder in downstream tasks, failing to achieve comprehensive cross-modal understanding for tasks like Med-VQA and Med-ITR. Consequently, the full potential of cross-modality alignment still remains under-explored.

To solve the problems of fully utilizing the multi-view images in the dataset, and explore the cross-modality alignment, we propose a novel and systematic approach for Med-VLP, named **MVCM**: a framework that supports **Multi-View** alignment and comprehensive **Cross-Modality** alignment for Medical Vision Language Pre-training. Our contributions can be summarized as follows:

- We propose to utilize the multi-view images in the medical image-report dataset at study-level. Specifically, we re-organize the dataset per study, keeping all multi-view images of one study. When one study contains multiple views, we randomly sample two images of different views and align them in a pretext task. When the study in a dataset corresponds to only one image, different views of one image can be generated via strong random augmentations [5, 21]. The related pretext task is proposed as Image Multi View Contrastive learning (IMVC), which explores the inner relationship within the visual modal-

ity. Our approach fully utilizes the medical image-report dataset, enabling our model to learn medical knowledge from multiple views.

- We propose to use multiple refined pretext tasks for cross-modality alignment. We not only use Image-Text Global Contrastive Learning (ITGC) to align the global image-text representations but also align the representations locally via Image-Text Local Contrastive Learning (ITLC). Besides, we utilize pretext tasks that require a cross-modality understanding of image and text, like Image-grounded Text Generation (ITG) and Image Text Matching (ITM).
- Our MVCM achieves state-of-the-art performance on various downstream tasks that require a joint understanding of images and texts, such as VQA and Image-Text Retrieval. We first pre-train MVCM (Fig. 1(a)) on multiple datasets jointly including MIMIC-CXR [28], ROCO [43] and MedCaT [47]. Then we fine-tune MVCM (Fig. 1(b)) on downstream datasets like VQA-RAD [30], SLAKE [37] and PathVQA[23] datasets for VQA task, as well as CheXpert 5×200 [25, 26] and ROCO datasets for Image-Text Retrieval task.

2. Related Works

2.1. Medical Vision Language Pre-training

Recently, jointly pre-training vision and language for medical data has been explored. ConVIRT [53] and MedCLIP [50] perform global alignment via contrastive learning. CPCR [38] combines conditional reasoning and contrastive learning. GLoRIA [25], LoVT [41], MGCA [48] and PRIOR [9] try to combine the local alignment plus global alignment together. To explore generative pretext tasks, [51, 55] focus on tasks like Masked Image Modeling (MIM), while M3AE [7] combines Masked Language Modeling (MLM) and MIM, fusing multi-modality information via cross-attention to reconstruct the masked information. Further studies like MRM[54] and CMITM [3] combine contrastive and generative pretext tasks, focusing on the global alignment. PRIOR [9] attempts to additionally learn from prototype representation. However, most of these models only utilize the uni-modal encoders for downstream tasks (e.g., Classification and Segmentation), constraining their cross-modality understanding.

2.2. Learning from Multi-View Medical Images

There are many works on multi-view images for classification [1, 2], segmentation [39] and text-report generation [49, 52]. For the chest X-rays dataset, CheXpert [26] and MIMIC [28] are widely used for Med-VLP’s pre-training. However, most previous studies use chest X-ray images for disease classification and report generation, considering the frontal and lateral images from the same patient as two in-

dependent instances [27], while some studies like [49, 52] argue that lateral images contain complementary information to frontal images that benefit text report generation. LIMITR [11] utilizes both the frontal and lateral views of one patient for aligning with text reports. To the best of our knowledge, there is still no study that explores the alignment within multi-view images for Med-VLP.

2.3. Cross-Modality Understanding for Med-VLP

To explore the fusion module for cross-modality understanding and explore complex downstream tasks like Visual Question Answering (VQA), M3AE [7] tries to use cross-attention to integrate image and text information via Vision Transformer. ARL [8] and LaPA [18] use knowledge extracted outside the dataset. Inspired by works from [24, 31], M2I2 [36] uses Momentum Distillation (MoD) for efficient supervision while training, also combining contrastive and generative pretext tasks. Following M2I2, MUMC [35] not only uses MoD, but also tries to align representations of multi-modality and uni-modality and achieves better VQA ability. Though these methods apply MoD to the cross-modality alignment, they only focus on pretext tasks like Image-Text Contrastive learning, without exploring the local alignment of image and text or the alignment of multi-view images to cross-modality understanding.

3. Methodology

We first introduce our pre-training model’s architecture (Sec. 3.1), then we introduce the pre-training objectives (Sec. 3.2), and how Momentum Distillation works with the cross-modality alignment (Sec. 3.3). We then introduce the downstream tasks for our pre-trained model (Sec. 3.4).

3.1. Pre-training Model

3.1.1 Overview

Our pre-training model (Fig. 1(a)) tries to employ the dataset at study-level, which contains multi-view images and matched text reports. Given a training set $\mathcal{D} = \{(X_1, R_1), \dots, (X_n, R_n), \dots, (X_N, R_N)\}$ that represents N studies of images and text reports. We use $X_n = \{\mathbf{x}_n^1, \mathbf{x}_n^2, \dots, \mathbf{x}_n^{s_n}\}$, $s_n \geq 1$ to represent the n -th study that has s_n amount of image views, and R_n to represent the corresponding text report. During each training epoch, two views of a study are randomly sampled, denoted as $\mathbf{x}_n^{(1)}, \mathbf{x}_n^{(2)}$, which are then encoded by a visual encoder (e.g., Vision Transformer [14]), yielding two features $V_n^{(1)}, V_n^{(2)}$. In addition, the text report R_n are directly encoded by the tokenizer, yielding W_n .

Our MVCM, denoted as f in Fig. 2(a), consists of visual branch and textual branch (i.e., (f_v, f'_v) and (f_t, f'_t)), which are both able to process uni-modality and cross-modality

information separately. For uni-modality, MVCM takes $V_n^{(1)}, V_n^{(2)}, W_n$ as input, and output features via f_v, f_t :

$$\begin{aligned} (I_n^{(1)}, I_n^{(2)}, T_n) &= f(V_n^{(1)}, V_n^{(2)}, W_n) \\ &= (f_v(V_n^{(1)}), f_v(V_n^{(2)}), f_t(W_n)), \end{aligned} \quad (1)$$

where $I_n^{(1)}, I_n^{(2)} \in \mathbb{R}^{p \times d}$, $T_n \in \mathbb{R}^{w_n \times d}$. $I_n^{(1)}$ and $I_n^{(2)}$ represent image features that contain p d -dimensional tokens $\{z_n^1, z_n^2, \dots, z_n^p\}$, while T_n represents text feature with $w_n + 1$ tokens $\{t_n^{cls}, t_n^1, t_n^2, \dots, t_n^{w_n}\}$. t_n^{cls} is the [CLS] token of T_n . Uni-modality features $I_n^{(1)}, I_n^{(2)}, T_n$ are utilized for modality alignment (i.e., Image Multi-View Contrastive Learning (IMVC), Image-Text Global Contrastive Learning (ITGC) and Image-Text Local Contrastive Learning (ITLC)). The cross-modality information in MVCM is processed via f'_v, f'_t :

$$(I'_n, T'_n) = (f'_v(V_n, W_n), f'_t(W_n, V_n)), \quad (2)$$

where $V_n \in \{V_n^{(1)}, V_n^{(2)}\}$ is one of the multi-view images selected randomly, I'_n and T'_n are of the same shape as the I_n and T_n . They are subsequently used for Image-grounded Text Generation (ITG) and Image-Text Matching (ITM), which require cross-modality information.

3.1.2 Architecture of MVCM

As shown in Fig. 2(a), for Modality Alignment (i.e., ITGC, ITLC, and IMVC), we use f_t and f_v , which keep uni-modality information independently (Eq. (1)), while for ITG and ITM we use f'_t and f'_v , where cross-modality representations are bidirectional (Eq. (2)). To achieve this, the sub-modules within each branch use different Self-Attention (SA) layers, denoted with different colors in Fig. 2(a). Visual branch utilizes a query $Q \in \mathbb{R}^{p \times d}$ with p learnable tokens to succinctly represent image information. In the Cross-Attention (CA) layer, f_v performs cross-attention between Q and paired input image features $V_n^{(1)}$ and $V_n^{(2)}$, whereas f'_v conducts cross-attention between Q and V_n . Meanwhile, textual branch f_t and f'_t take W_n as input in Self-Attention layers. Finally, after the respective feed-forward (FF) layers of the visual branch and textual branch, our MVCM outputs $I_n^{(1)}, I_n^{(2)}, T_n$ as depicted in Eq. (1), and outputs I'_n, T'_n as Eq. (2).

MVCM also has a momentum model $f_{\mathcal{M}}$, which is of the identical structure as the main model f and uses a momentum strategy [24, 31] to update its weight. The outputs of $f_{\mathcal{M}}$ are denoted as $(I_{\mathcal{M},n}^{(1)}, I_{\mathcal{M},n}^{(2)}, T_{\mathcal{M},n}, I'_{\mathcal{M},n}, T'_{\mathcal{M},n})$, which are involved in computing the momentum distillation loss of ITG, IMVC, ITGC.

3.2. Pre-training Objectives

As shown in Fig. 2(a), we train MVCM via five objectives. We use Image-Text Global Contrastive Learning (ITGC),

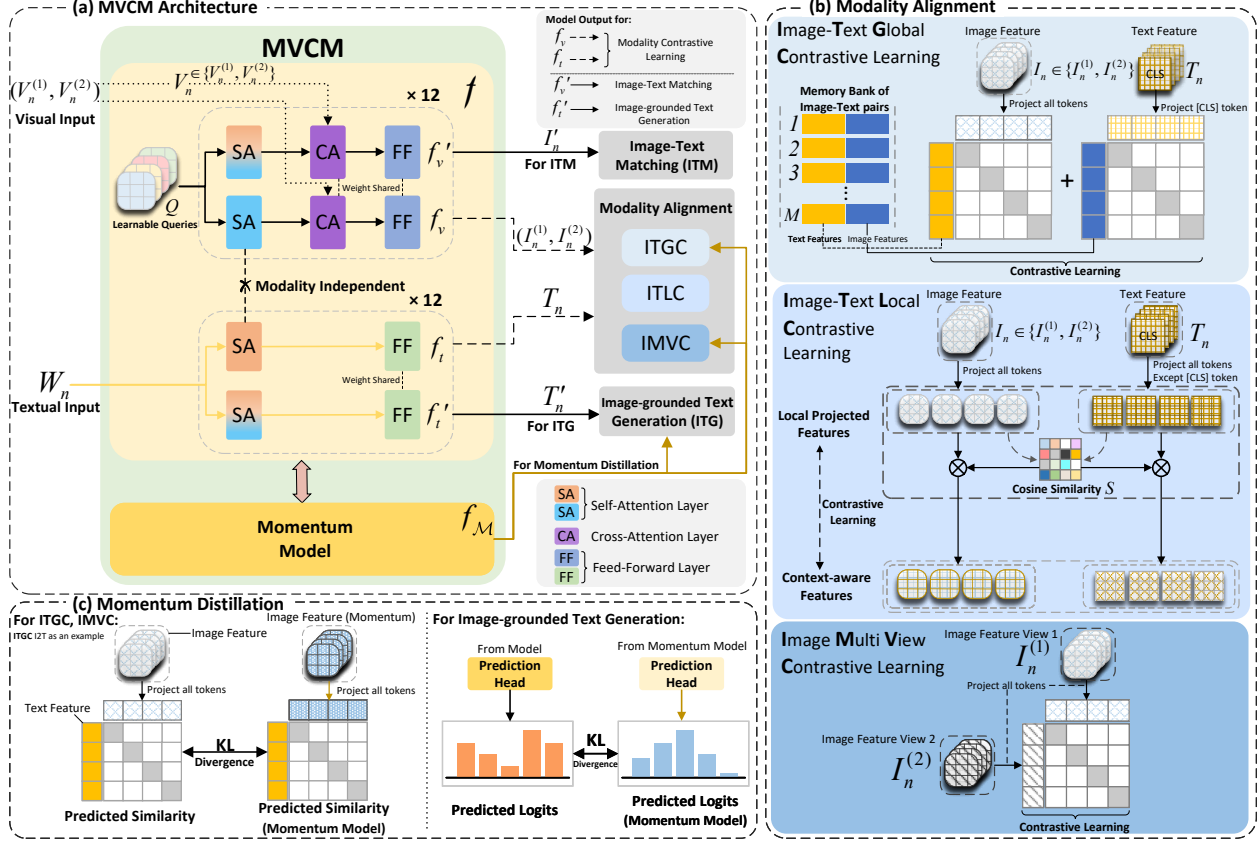


Figure 2. (a) The architecture of MVCMM for pre-training. (b) Modality Alignment with three pretext tasks. (c) The overview of Momentum Distillation, where Kullback-Leibler (KL) divergence loss is computed between the model and its momentum model.

Image-Text Local Contrastive Learning (ITLC) and Image Multi-View Contrastive Learning (IMVC) to train Modality Alignment. Besides, we utilize Image-grounded Text Generation (ITG) and Image-Text Matching (ITM) to enable pretext tasks with cross-modality information.

3.2.1 Image-Text Global Contrastive Learning

ITGC learns to predict whether an image-text pair is positive or negative. As shown in Fig. 2(b), given an input image-text pair (I_n, T_n) from MVCMM, we compute the InfoNCE Loss [42] between its features. Besides, to expand the range of contrastive learning, we maintain the most recent M pairs of projected image-text features in a Memory Bank, denoting them as $\{(I_1^q, T_1^q), (I_2^q, T_2^q), \dots, (I_M^q, T_M^q)\}$. We first compute the similarity between global image feature (I_n) and text features (t_n^{cls}) using:

$$s(I_n, t_n^{cls}) = \text{Max}(\text{Proj}_v(I_n)\text{Proj}_t(t_n^{cls})), \quad (3)$$

where $\text{Max}(\cdot)$ selects the highest value from all components, Proj_v and Proj_t is a d to d' projection for global

image features and text features. Then, the InfoNCE Loss is calculated using the features from the Memory Bank:

$$\begin{aligned} \mathcal{L}^{i2t}(I_n) &= -\log \frac{\exp(s(I_n, t_n^{cls})/\tau_1)}{\sum_{k=1}^M \exp(s(I_n, t_k^{cls,q})/\tau_1)}, \\ \mathcal{L}^{t2i}(T_n) &= -\log \frac{\exp(s(t_n^{cls}, I_n)/\tau_1)}{\sum_{k=1}^M \exp(s(t_n^{cls}, I_k^q)/\tau_1)}, \end{aligned} \quad (4)$$

where τ_1 is a temperature parameter, and $t_k^{cls,q}$ represents the [CLS] token of T_k in the Memory Bank. Afterwards, we calculate \mathcal{L}_{ITGC} as:

$$\mathcal{L}_{ITGC} = \frac{1}{2N} \sum_{n=1}^N [\mathcal{L}^{i2t}(I_n) + \mathcal{L}^{t2i}(T_n)]. \quad (5)$$

3.2.2 Image-Text Local Contrastive Learning

ITLC aligns image features and text features at the token-wise level. Specifically, given (I_n, T_n) from MVCMM, ITLC aligns the image tokens $\{z_n^1, z_n^2, \dots, z_n^p\}$ in the image feature I_n , with the word tokens $\{t_n^1, t_n^2, \dots, t_n^w\}$ in the text feature T_n except the [CLS] token t_n^{cls} .

We first project I_n and T_n into lower-dimensional features with dimension d'' using function g , obtaining $\tilde{I}_n = g(I_n)$, $\tilde{T}_n = g(T_n)$, where $\tilde{I}_n \in \mathbb{R}^{p \times d''}$, $\tilde{T}_n \in \mathbb{R}^{w_n \times d''}$. The tokens in the projected features are marked as $\{\tilde{z}_n^1, \tilde{z}_n^2, \dots, \tilde{z}_n^p\}$ and $\{\tilde{t}_n^1, \tilde{t}_n^2, \dots, \tilde{t}_n^{w_n}\}$. The cosine similarity between image and text is calculated as $S = \tilde{I}_n(\tilde{T}_n)^\top$, $S \in \mathbb{R}^{p \times w_n}$. Let $S_{ij} = \text{sim}(\tilde{z}_n^i, \tilde{t}_n^j)$ denotes the similarity between the i -th image token and the j -th word token, thus the softmax-normalized cosine similarity is:

$$a_{ij}^{t2i} = \frac{\exp(S_{ij}/\tau_2)}{\sum_{k=1}^p \exp(S_{kj}/\tau_2)}, \quad (6)$$

where τ_2 is a temperature parameter. The context-aware image feature based on the j -th word token is as follows:

$$c_j^{t2i} = \sum_{k=1}^p a_{kj}^{t2i} \tilde{z}_n^k. \quad (7)$$

We then propose to use c_j^{t2i} to compute the $P_j^{t2i}(n)$, which stands for the InfoNCE Loss between localized text feature \tilde{t}_n^j and context-aware localized image feature c_j^{t2i} . Given the j -th word token, $P_j^{t2i}(n)$ is computed as:

$$P_j^{t2i}(n) = -\frac{1}{2} \left(\log \frac{\exp(\text{sim}(\tilde{t}_n^j, c_j^{t2i})/\tau_3)}{\sum_{k=1}^{w_n} \exp(\text{sim}(\tilde{t}_n^j, c_k^{t2i})/\tau_3)} + \log \frac{\exp(\text{sim}(c_j^{t2i}, \tilde{t}_n^j)/\tau_3)}{\sum_{k=1}^{w_n} \exp(\text{sim}(c_j^{t2i}, \tilde{t}_n^k)/\tau_3)} \right), \quad (8)$$

where τ_3 is a temperature parameter similar to τ_2 . We compute the localized text feature loss $\mathcal{L}_{\text{ITLC}}^{t2i}$ as:

$$\mathcal{L}_{\text{ITLC}}^{t2i} = \frac{1}{N} \sum_{n=1}^N \left[\frac{1}{w_n} \sum_{j=1}^{w_n} r_j^{t2i} P_j^{t2i}(n) \right], \quad (9)$$

where r_j^{t2i} is an assigned weight for the loss of each word token \tilde{t}_n^j . Specifically, r_j^{t2i} is set to the averaged attention weight from \tilde{t}_n^j to \tilde{t}_n^{cls} generated in the last layer of f_t .

Similarly, the $\mathcal{L}_{\text{ITLC}}^{i2t}$ can be calculated as:

$$\mathcal{L}_{\text{ITLC}}^{i2t} = \frac{1}{Np} \sum_{n=1}^N \sum_{i=1}^p r_i^{i2t} P_i^{i2t}(n), \quad (10)$$

where $P_i^{i2t}(n)$ stands for the InfoNCE Loss computed similarly as Eq. (8), and r_i^{i2t} is an assigned weight for image token \tilde{z}_n^i . We use the averaged attention weight from p image tokens to \tilde{z}_n^i in the last layer of f_v . Ultimately, using Eq. (10) and Eq. (9), we have:

$$\mathcal{L}_{\text{ITLC}} = \frac{1}{2} (\mathcal{L}_{\text{ITLC}}^{t2i} + \mathcal{L}_{\text{ITLC}}^{i2t}). \quad (11)$$

3.2.3 Image Multi-View Contrastive Learning

IMVC enables MVCM to align two different views of images that belong to one study. Calculating IMVC is similar to ITGC. Given two different views of images $I_n^{(1)}, I_n^{(2)}$, we first compute the similarity between the projected image features and text features using

$$s'(I_n^{(1)}, I_n^{(2)}) = \text{Max}(\text{Proj}_v(I_n^{(1)})(\text{Proj}_t(I_n^{(2)}))^\top), \quad (12)$$

then we compute the InfoNCE loss within one batch with B samples $\{(I_1^{(1)}, I_1^{(2)}), (I_2^{(1)}, I_2^{(2)}), \dots, (I_B^{(1)}, I_B^{(2)})\}$:

$$l^{i1,i2}(I_n^{(1)}) = -\log \frac{\exp(s'(I_n^{(1)}, I_n^{(2)})/\tau_4)}{\sum_{k=1}^B \exp(s'(I_n^{(1)}, I_k^{(2)})/\tau_4)}, \quad (13)$$

$$l^{i2,i1}(I_n^{(2)}) = -\log \frac{\exp(s'(I_n^{(2)}, I_n^{(1)})/\tau_4)}{\sum_{k=1}^B \exp(s'(I_n^{(2)}, I_k^{(1)})/\tau_4)},$$

where τ_4 is a temperature parameter. Afterwards, we have:

$$\mathcal{L}_{\text{IMVC}} = \frac{1}{2N} \sum_{n=1}^N [l^{i1,i2}(I_n^{(1)}) + l^{i2,i1}(I_n^{(2)})]. \quad (14)$$

3.2.4 Image-grounded Text Generation

ITG trains our MVCM to generate the masked texts. For Modality Alignment, the multi-modality information is mutually independent, while for ITG, we need fused information. Since the architecture of MVCM only allows interactions between the image features extracted by query tokens and the text tokens, we utilize a self-attention masking strategy, passing the cross-modality information bidirectionally in self-attention layers. To compute \mathcal{L}_{ITG} , we decode the T'_n obtained from Eq. (2), yielding the generated text, and then we compare it with the W_n via a CrossEntropy loss.

3.2.5 Image-Text Matching

ITM predicts the matching logits between images and texts, via a linear prediction head. Specifically, the self-attention mask is similar as in ITG, and by using Eq. (2), we obtain fused feature I'_n . Afterwards, we apply I'_n to a two-class linear classifier, and compute the cross-entropy loss between the predicted label and true label, yielding \mathcal{L}_{ITM} .

3.3. Momentum Distillation

Momentum Distillation (MoD, [24]) is used for efficient training across various pretext tasks. Specifically, we implement MoD as a form of online self-distillation [31], which allows unpaired images and texts to have pseudo-labels generated by the model itself (Fig. 2 (c)).

Similar to the Eq. (4), we have $l_{\mathcal{M}}^{i2t}(I_{\mathcal{M},n}), l_{\mathcal{M}}^{t2i}(T_{\mathcal{M},n})$ for the momentum model, where $I_{\mathcal{M},n} \in \{I_{\mathcal{M},n}^{(1)}, I_{\mathcal{M},n}^{(2)}\}$

for simplicity. The modified ITGC loss (\mathcal{L}'_{ITGC}) is:

$$\begin{aligned} \mathcal{L}'_{ITGC} &= (1 - \alpha)\mathcal{L}_{ITGC} + \alpha\mathcal{L}_{ITGC}^{\text{MoD}}, \\ \mathcal{L}_{ITGC}^{\text{MoD}} &= \frac{1}{2N} \sum_{n=1}^N [\text{KL}(\mathbf{t}_{\mathcal{M}}^{i2t}(I_{\mathcal{M},n}) \parallel \mathbf{t}^{i2t}(I_n)) \\ &\quad + \text{KL}(\mathbf{t}_{\mathcal{M}}^{t2i}(T_{\mathcal{M},n}) \parallel \mathbf{t}^{t2i}(T_n))], \end{aligned} \quad (15)$$

where α is a coefficient for MoD. For IMVC, we calculate the $\mathbf{t}_{\mathcal{M}}^{i1,i2}(I_{\mathcal{M},n}^{(1)})$, $\mathbf{t}_{\mathcal{M}}^{i2,i1}(I_{\mathcal{M},n}^{(2)})$ similarly as in Eq. (13). Then we have the modified IMVC loss (\mathcal{L}'_{IMVC}) as:

$$\begin{aligned} \mathcal{L}'_{IMVC} &= (1 - \alpha)\mathcal{L}_{IMVC} + \alpha\mathcal{L}_{IMVC}^{\text{MoD}}, \\ \mathcal{L}_{IMVC}^{\text{MoD}} &= \frac{1}{2N} \sum_{n=1}^N [\text{KL}(\mathbf{t}_{\mathcal{M}}^{i1,i2}(I_{\mathcal{M},n}^{(1)}) \parallel \mathbf{t}^{i1,i2}(I_n^{(1)})) \\ &\quad + \text{KL}(\mathbf{t}_{\mathcal{M}}^{i2,i1}(I_{\mathcal{M},n}^{(2)}) \parallel \mathbf{t}^{i2,i1}(I_n^{(2)}))]. \end{aligned} \quad (16)$$

For ITG, given T'_n , we let $\mathbf{p}(T'_n)$ denotes the model's predicted logits for feature T'_n , and $\mathbf{p}_{\mathcal{M}}(T'_{\mathcal{M},n})$ denotes the momentum model's. Then we have:

$$\begin{aligned} \mathcal{L}'_{ITG} &= (1 - \alpha)\mathcal{L}_{ITG} + \alpha\mathcal{L}_{ITG}^{\text{MoD}}, \\ \mathcal{L}_{ITG}^{\text{MoD}} &= \frac{1}{2N} \sum_{n=1}^N \text{KL}(\mathbf{p}(T'_n) \parallel \mathbf{p}_{\mathcal{M}}(T'_{\mathcal{M},n})). \end{aligned} \quad (17)$$

Ultimately, we have the overall loss:

$$\begin{aligned} \mathcal{L} &= \lambda_1\mathcal{L}'_{ITGC} + \lambda_2\mathcal{L}_{ITLC} + \lambda_3\mathcal{L}'_{IMVC} \\ &\quad + \lambda_4\mathcal{L}'_{ITG} + \lambda_5\mathcal{L}_{ITM}, \end{aligned} \quad (18)$$

where λ_1 to λ_5 are coefficients for loss functions.

We do not apply MoD to ITLC, as the tokens of localized features and the corresponding context-aware features already have exact correspondence, which makes it more appropriate to use one-hot labels. Additionally, ITM is a pretext task that predicts binary results (*i.e.*, matched or not matched) per image-text pair, which does not consider the similarity of every image and text in one batch. Therefore, MoD is not applied to ITLC or ITM during training.

3.4. MVCM for Downstream Tasks

We apply MVCM to various downstream tasks like Visual Question Answering and Image-Text Retrieval.

3.4.1 Medical Visual Question Answering

Med-VQA requires a model to predict the answer given an image and a question. Common methods consider VQA a multi-answer classification task [6]. Instead, we consider VQA as an answer generation problem, using an auto-regressive decoder [10, 31], shown as the decoder in Fig. 1(b). For training, the decoder is trained with conditional language-modeling loss. To fairly compare with other methods, the decoder will generate answers closest to one of the candidate answers for evaluation.

3.4.2 Medical Image-Text Retrieval

For Med-ITR, we compute similarity matrices between images and texts to facilitate both Image-to-Text (I2T) and Text-to-Image (T2I) retrieval. In the I2T scenario, each image is treated as a query, and the texts are considered as the candidates. Conversely, for T2I, each text serves as a query, with images as the candidates. The specific logits between the image and text are obtained by adding up the predicted score of Image-Text Matching and the similarity of Image-Text Global Contrastive Learning.

4. Experiment

4.1. Datasets

For pre-training MVCM, we use three datasets: MIMIC-CXR [28] is the largest publicly available radiology dataset that consists of 377110 X-ray images and 227827 reports; ROCO [43] is a radiology dataset that has over 81,000 radiology images; and MedCaT [47] contains over 217,000 image-caption pairs.

For Med-VQA, we use: VQA-RAD [30] has 315 radiology images with 3064 question-answer pairs; SLAKE [37] has 14,028 pairs of samples; and PathVQA[23] contains 32,799 image-text pairs. All these datasets contain *Open* (form-free answers) and *Closed* (*yes/no* answers) types of questions.

For Med-ITR, we utilize CheXpert 5×200 [26] and ROCO [43] datasets. Since CheXpert 5×200 has not publicly released its reports, we follow the pre-processing method of PRIOR [9] and GLoRIA [25], randomly selecting 1000 reports from MIMIC-CXR with 200 samples for each of 5 corresponding diseases. For the ROCO dataset, we use the same test split following [7].

4.2. Multi-View Data Generation

We pre-process the pre-training datasets, to fully utilize study-level data and learn from the multi-view images. MIMIC-CXR [28] stores images and reports at study-level, each study of which has a Study-ID that corresponds to one report and several images of different views. Unlike instance-level, we reorganize the dataset by the unique Study-ID for each sample. For multi-view generation for all datasets, two distinct strategies were employed based on the number of views per sample. For samples with only one view, we apply strong random augmentations used in SimCLR [5] and MoCo [21]. For samples with multiple image views, we employed a bootstrapping approach to randomly select two views, which were then subjected to the same random augmentation process. The random selection from one study allows an equal chance for each view. The processed MIMIC-CXR dataset comprises 216,306 samples, with an average of 1.667 different views per sample. Besides, since ROCO [43] and MedCaT [47] are not origi-

Method	VQA-RAD			SLAKE			PathVQA		
	Open	Closed	Overall	Open	Closed	Overall	Open	Closed	Overall
MMQ [13]	53.70	75.80	67.00	-	-	-	13.40	84.0	48.80
ARL [8]	65.10	85.96	77.50	79.70	89.30	84.00	-	-	-
M3AE [7]	67.23	83.46	77.01	80.31	87.82	83.25	-	-	-
CPCR [38]	60.50	80.40	72.50	80.50	84.10	81.90	-	-	-
PubMedCLIP [16]	60.10	80.00	72.10	78.40	82.50	80.10	-	-	-
MUMC [35]	71.50	84.20	79.20	-	-	84.90	<u>39.00</u>	90.40	65.10
M2I2 [36]	61.80	81.60	73.70	74.70	<u>91.10</u>	<u>81.20</u>	36.30	88.00	62.20
PeFoMed [20]	62.60	<u>87.10</u>	77.40	77.80	88.70	82.10	35.70	91.30	<u>63.60</u>
LaPA [18]	<u>68.72</u>	86.40	<u>79.38</u>	<u>82.17</u>	88.70	84.73	-	-	-
MVCM (Ours)	68.50	90.84	80.93	83.71	91.83	86.89	39.83	<u>90.49</u>	65.26

Table 1. Comparison with existing methods on various VQA datasets for accuracy. *Open* represents the form-free question set while *Closed* represents the question set with *yes/no* as the answer. The best and second-best results are **bolded** and underlined, respectively.

Method	Image to Text			Text to Image			Class-Based Retrieval			
	Prec@1	Prec@5	Prec@10	Prec@1	Prec@5	Prec@10	Accuracy	F1-Score	Precision	Recall
GloRIA [25]	34.40	31.18	28.90	34.10	<u>35.08</u>	34.10	21.70	22.32	31.20	31.10
ConVIRT [53]	28.20	30.30	26.50	33.80	32.00	32.10	23.74	21.50	25.00	24.20
PRIOR [9]	<u>36.00</u>	34.40	33.82	<u>36.50</u>	35.06	<u>34.60</u>	<u>35.90</u>	<u>34.15</u>	<u>39.27</u>	<u>35.90</u>
LIMITR [11]	-	37.20*	<u>35.90*</u>	-	-	-	-	-	-	-
MGCA [48]	-	29.30*	<u>27.60*</u>	-	-	-	-	-	-	-
MVCM (Ours)	36.52	<u>36.53</u>	36.30	42.70	38.46	37.54	44.40	41.92	48.42	44.40

Table 2. Comparison with existing methods on CheXpert 5x200 dataset for Med-ITR. Precision (%) at the top 1, 5, and 10 candidates are reported. Class-Based Retrieval (*i.e.*, disease labels as texts) is also evaluated using F1-score, Precision, and Recall. * denotes the reported results are from other papers due to unavailability to evaluate ourselves.

Method	Image to Text			Text to Image		
	R@1	R@5	R@10	R@1	R@5	R@10
ViLT [29]	11.90	31.90	43.20	9.75	28.95	41.40
METER [15]	14.45	33.30	45.10	11.30	27.25	39.60
M3AE [7]	19.10	45.60	61.20	19.05	47.75	61.35
ARL [8]	<u>23.45</u>	<u>50.60</u>	<u>62.05</u>	23.50	49.05	63.00
MVCM (Ours)	24.20	55.40	71.85	<u>22.95</u>	54.92	73.30

Table 3. Comparison with existing methods for Med-ITR on ROCO dataset. Recall (%) is reported.

nally organized at study-level, we apply the method for single view case to them. We finally have a total of 498,783 samples with multi-view images for pre-training.

4.3. Implementation Details

We apply ViT/B-16 [14] for visual encoder and bert-base-uncased [12] for textual encoder. All images are resized to 288x288 for the pre-training and retrieval task, while set to 384x384 for fine-tuning the VQA task. We set the number of learnable query tokens $p = 32$. The temperature parameters $\tau_1, \tau_2, \tau_3, \tau_4$ are all set to 0.07. For the lower dimensional embedding, we set $d' = 256, d'' = 128$. For the Memory Bank, we set $M = 65536$, and we update it as a queue, removing the oldest batch and appending the current batch. For Momentum Distillation, we set α to 0.04 in the first epoch. As for the loss function coefficients, due to

the various compositions and high computing cost for pre-training, we set $\lambda_1 = \lambda_4 = \lambda_5 = \lambda_2 = \lambda_3 = 1$.

Our MVCM is pre-trained with batch size of 64 for 40 epochs and then fine-tuned on VQA tasks with a batch size of 64 for 60 epochs. The learning rate for pre-training is set to $1e-4$ with a minimum of $2e-6$, while for fine-tuning it is set to $5e-5$ with the same minimum. Besides, we use the Adam optimizer with a weight decay of 0.02. For computing resources, we train our model on 4 NVIDIA A100 GPUs, the total time for pre-training is about 5 days, while the fine-tuning time is around 10 hours depending on the specific Med-VQA datasets. For the Med-ITR task, our inference time is around 2.3s for one query on the CheXpert 5x200 dataset using a single GPU.

4.4. Comparison with State-of-the-Art

For Med-VQA (Tab. 1), our model outperforms other methods in both the *Close* subset and *Overall* on VQA-RAD dataset. For SLAKE dataset, MVCM achieves the best scores in all types, surpassing the second-best method by 1.99% for *overall* accuracy. For PathVQA, our model also achieves the highest score on both *Open* and the *overall*. These results underscore our MVCM’s superior ability of cross-modality understanding for VQA.

For Med-ITR, our MVCM outperforms all existing

#	ITM	ITG	ITGC	IMVC	ITLC	CheXpert 5×200			RAD-VQA		
						Prec@1 (I2T)	Prec@1 (T2I)	Acc	Open	Closed	Overall
1	-	-	-	-	-	-	-	-	59.32	72.58	67.32
2	✓	-	✓	-	-	25.60	32.45	35.24	63.81	79.09	73.03
3	✓	✓	✓	-	-	26.36	33.74	36.55	64.93	79.51	73.72
4	✓	\mathcal{M}	\mathcal{M}	-	-	31.71	36.74	38.92	65.81	83.09	76.23
5	✓	\mathcal{M}	\mathcal{M}	\mathcal{M}	-	33.27	37.33	41.52	67.33	86.71	79.02
6	✓	\mathcal{M}	\mathcal{M}	\mathcal{M}	\mathcal{M}	33.53	37.21	37.21	66.82	83.94	77.15
7	✓	\mathcal{M}	\mathcal{M}	\mathcal{M}	✓	36.52	42.70	44.40	68.50	90.84	80.93

Table 4. Impact of different components on the CheXpert 5×200 and RAD-VQA datasets. We evaluate various combinations of Image-Text Global Contrastive Learning (ITGC), Image-Grounded Text Generation (ITG), Image Multi-View Contrastive Learning (IMVC), Image-Text Local Contrastive Learning (ITLC). Checkmark (✓) indicates using the component without Momentum Distillation (MoD), while \mathcal{M} indicates using it with MoD.

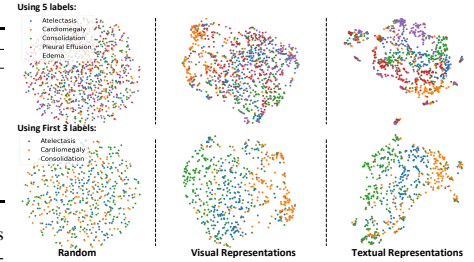


Figure 3. UMAP visualization of MVCM’s representations on the CheXpert 5×200 dataset.

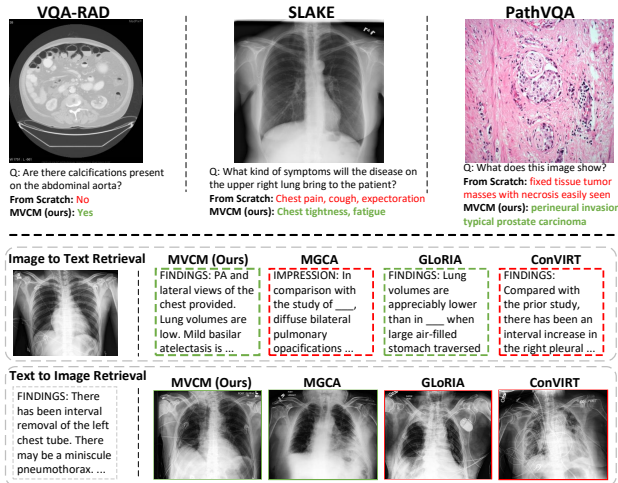


Figure 4. Comparison cases for Med-VQA (top) and Med-ITR (bottom, on the CheXpert 5×200 dataset). Green ones denote correct answers or retrievals, while red ones denotes failed cases. We compare with MGCA [48], GLoRIA [25], and ConVIRT [53].

methods on the CheXpert 5×200 dataset (Tab. 2). Additionally, MVCM achieves the generally best results on ROCO dataset (as shown in Tab. 3), with only one metric at the second best. In the recall of T2I with 10 candidates, our model greatly surpasses the second-best method by 10.30%. The performance of MVCM on the Med-ITR tasks demonstrates its superior capability in aligning cross-modal information.

4.5. Ablation Study

In Tab. 4, we evaluate the impacts of various components of MVCM to prove the effectiveness of our method using two representative datasets (CheXpert 5×200 and RAD-VQA). Our baseline is the model trained from scratch (#1). To validate the significance of multiple cross-modality alignments, we first compare the baseline with #2, showing that **ITGC** and **ITM** significantly enhance the results. Then we observe that the incorporation of **ITG** (#3) results in further improvements. Besides, we find applying **MoD** to the pretext tasks facilitates the model’s ability by comparing #4 with #3. Notably, using multi-view alignment (**IMVC**, #5)

greatly improves the model’s performance compared with #4, boosting the overall accuracy to 41.52% and 79.02% separately. We also notice that applying MoD to ITLC (#6) results in a decline across all metrics while applying **ITLC without MoD** (#7) boosts the model’s performance, which is consistent with our previous analysis at the end of Sec. 3.3.

4.6. Qualitative Results

We provide some representative cases for Med-VQA tasks in the top part of Fig. 4. For *Closed* questions (e.g., the left one), our model can identify the target (i.e., *calcifications*) and then predict the right answer. For *Open* questions (e.g., the middle and the right ones), our model can recognize and depict the characteristics of the disease in certain regions (e.g., *upper right*), and then accurately answer them.

The visualization of image and text representations for Med-ITR (Fig. 3) shows that our model’s representations cluster according to distinct classes, especially when we use the first three classes that have significant differences (i.e., *Atelectasis*, *Cardiomegaly*, and *Consolidation*). For the bottom part of Fig. 4, both results of I2T and T2I retrievals indicate that our model can correctly retrieve the related representations of the same classes based on the query.

5. Conclusion

In this paper, we fully utilize the multi-view images in the medical datasets. We consider each study with multi-view images and the corresponding reports as one sample, and a multi-view alignment is proposed to align these multi-view representations within one study. Besides, to advance a comprehensive cross-modality understanding, our method employs various refined pretext tasks focusing on both global and local alignments between image-text representations. Additional tasks (e.g., Image-Text Matching and Image-grounded Text Generation) and momentum distillation are used to further augment this understanding. Our method is applied to various downstream tasks like Visual Question Answering and Image Text Retrieval. Extensive experiments on these downstream tasks demonstrate the effectiveness of our method.

References

- [1] Alan Joseph Bekker, Moran Shalhon, Hayit Greenspan, and Jacob Goldberger. Multi-view probabilistic classification of breast microcalcifications. *IEEE Transactions on medical imaging*, 35(2):645–653, 2015. 2
- [2] Gustavo Carneiro, Jacinto Nascimento, and Andrew P Bradley. Deep learning models for classifying mammogram exams containing unregistered multi-view images and segmentation maps of lesions. *Deep learning for medical image analysis*, pages 321–339, 2017. 2
- [3] Cheng Chen, Aoxiao Zhong, Dufan Wu, Jie Luo, and Quanzheng Li. Contrastive masked image-text modeling for medical visual representation learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 493–503. Springer, 2023. 2
- [4] Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019. 1
- [5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2, 6
- [6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 6
- [7] Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 679–689. Springer, 2022. 2, 3, 6, 7
- [8] Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5152–5161, 2022. 2, 3, 7
- [9] Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21361–21371, 2023. 2, 6, 7
- [10] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 6
- [11] Gefen Dawidowicz, Elad Hirsch, and Ayellet Tal. Limitr: Leveraging local information for medical image-text representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21165–21173, 2023. 3, 7
- [12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 7
- [13] Tuong Do, Binh X Nguyen, Erman Tjiputra, Minh Tran, Quang D Tran, and Anh Nguyen. Multiple meta-model quantifying for medical visual question answering. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*, pages 64–74. Springer, 2021. 7
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 7
- [15] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 7
- [16] Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1181–1193, 2023. 7
- [17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent—a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020. 1
- [18] Tiancheng Gu, Kaicheng Yang, Dongnan Liu, and Weidong Cai. Lapa: Latent prompt assist model for medical visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4971–4980, 2024. 3, 7
- [19] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016. 1
- [20] Jinlong He, Pengfei Li, Gang Liu, Zixu Zhao, and Shenjun Zhong. Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. *arXiv preprint arXiv:2401.02797*, 2024. 7
- [21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2, 6
- [22] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 1
- [23] Xuehai He. Towards visual question answering on pathology images. In *Proceedings of the 59th annual meeting of*

the association for computational linguistics and the 11th international joint conference on natural language processing, 2021. 2, 6, 1

- [24] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 2, 3, 5
- [25] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 2, 6, 7, 8
- [26] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 2, 6
- [27] Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017. 3
- [28] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 2, 6
- [29] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 7
- [30] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2, 6
- [31] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1, 2, 3, 5, 6
- [32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1
- [33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1
- [34] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1
- [35] Pengfei Li, Gang Liu, Jinlong He, Zixu Zhao, and Shenjun Zhong. Masked vision and language pre-training with unimodal and multimodal contrastive losses for medical visual question answering. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 374–383. Springer, 2023. 2, 3, 7, 1
- [36] Pengfei Li, Gang Liu, Lin Tan, Jinying Liao, and Shenjun Zhong. Self-supervised vision-language pretraining for medical visual question answering. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. 2, 3, 7
- [37] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 2, 6
- [38] Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE transactions on medical imaging*, 42(5):1532–1545, 2022. 2, 7
- [39] Di Liu, Yunhe Gao, Qilong Zhangli, Ligong Han, Xiaoxiao He, Zhaoyang Xia, Song Wen, Qi Chang, Zhennan Yan, Mu Zhou, et al. Transfusion: multi-view divergent fusion for medical image segmentation with transformers. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 485–495. Springer, 2022. 2
- [40] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Viltbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 1
- [41] Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. Joint learning of localized representations from medical images and reports. In *European Conference on Computer Vision*, pages 685–701. Springer, 2022. 2
- [42] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 4
- [43] Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph 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: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer, 2018. 2, 6
- [44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1
- [45] Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of

- the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018. [1](#)
- [46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. [1](#)
- [47] Sanjay Subramanian, Lucy Lu Wang, Sachin Mehta, Ben Bogin, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. Medcat: A dataset of medical images, captions, and textual references. *arXiv preprint arXiv:2010.06000*, 2020. [2](#), [6](#)
- [48] Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *Advances in Neural Information Processing Systems*, 35:33536–33549, 2022. [2](#), [7](#), [8](#)
- [49] Ruizhi Wang, Xiangtao Wang, Jie Zhou, Thomas Lukasiewicz, and Zhenghua Xu. C²m-dot: Cross-modal consistent multi-view medical report generation with domain transfer network. *arXiv preprint arXiv:2310.05355*, 2023. [2](#), [3](#)
- [50] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. [2](#)
- [51] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3588–3600, 2023. [2](#)
- [52] Jianbo Yuan, Haofu Liao, Rui Luo, and Jiebo Luo. Automatic radiology report generation based on multi-view image fusion and medical concept enrichment. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, pages 721–729. Springer, 2019. [2](#), [3](#)
- [53] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR, 2022. [2](#), [7](#), [8](#)
- [54] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023. [2](#)
- [55] Lei Zhou, Huidong Liu, Joseph Bae, Junjun He, Dimitris Samaras, and Prateek Prasanna. Self pre-training with masked autoencoders for medical image classification and segmentation. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–6, 2023. [2](#)