MedUnifier: Unifying Vision-and-Language Pre-training on Medical Data with Vision Generation Task using Discrete Visual Representations

Ziyang Zhang[1,2], Yang Yu[3], Yucheng Chen[1,4], Xulei Yang[3,*], Si Yong Yeo[1,4,*]
[1]MedVisAI Lab, [2]ECE, Northwestern University, [3]Institute for Infocomm Research (I[2]R),
A*STAR, Singapore, [4]Lee Kong Chian School of Medicine, Nanyang Technological University
* Corresponding authors

Abstract

*Despite significant progress in Vision-Language Pre-training (VLP), current approaches predominantly emphasize feature extraction and cross-modal comprehension, with limited attention to generating or transforming visual content. This gap hinders the model's ability to synthesize coherent and novel visual representations from textual prompts, thereby reducing the effectiveness of multi-modal learning. In this work, we propose **MedUnifier**, a unified VLP framework tailored for medical data. MedUnifier seamlessly integrates text-grounded image generation capabilities with multi-modal learning strategies, including image-text contrastive alignment, image-text matching and image-grounded text generation. Unlike traditional methods that reply on continuous visual representations, our approach employs visual vector quantization, which not only facilitates a more cohesive learning strategy for cross-modal understanding but also enhances multi-modal generation quality by effectively leveraging discrete representations. Our framework's effectiveness is evidenced by the experiments on established benchmarks, including uni-modal tasks, cross-modal tasks, and multi-modal tasks, where it achieves state-of-the-art performance across various tasks. MedUnifier also offers a highly adaptable tool for a wide range of language and vision tasks in healthcare, marking advancement toward the development of a generalizable AI model for medical applications.*

1. Introduction

The rapid growth of medical imaging datasets has accelerated the development of deep-learning models to enhance clinical decision-making processes. However, annotating these extensive datasets requires specialized expertise, making large-scale annotation unfeasible. To overcome this limitation, one effective approach is to leverage associated medical reports containing detailed diagnostic descriptions provided by radiologists [52]. In recent years, deep learning models that utilize multi-modal data as inputs have drawn more attention, driven by advancements in attention mechanisms or transformer-based models [27, 40, 70].

Accordingly, vision-and-language pre-training (VLP) models have been developed, many drawing inspiration from the foundational CLIP model [48]. These models primarily leverage a dual-encoder approach, consisting of an image encoder and a text encoder, to independently extract uni-modal features. They aim to maximize cosine similarity between paired data via contrastive learning. Researchers have further enhanced these models by incorporating domain-specific knowledge and making targeted adjustments to the original CLIP, resulting in label-efficient adaptations [26, 62–64, 69, 71]. In addition, fusion-based encoders have attracted considerable attention. These fusion models utilize self-attention or co-attention mechanisms to achieve early integration of visual and textual modalities [3, 34, 35]. This joint processing enables the learning of multi-modal representations that are crucial for tasks requiring complex multi-modal reasoning, such as medical visual question answering. For fusion models, cross-modal matching with hard sampling strategies is employed to strengthen correlations between matched data. Image-grounded text understanding with masked language modelling (MLM), originally developed for the BERT [13], is also deployed to enhance multi-model interaction. PTUnifier [10], as an example, focused on multi-modal understanding yet did not possess generative ability during pre-training, thus necessitating the use of an additional language decoder and fine-tuning for language generative tasks. Meanwhile, image-grounded text generation with the causal language modelling (CLM) is often applied to facilitate vision-grounded language generation tasks. However, we observe that current VLP approaches often overlook the generation of visual content, limiting the model's capacity to produce coherent and novel visual representations based on textual or multi-modal prompts, thus reducing the potential of multi-modal learning. Although recent studies have

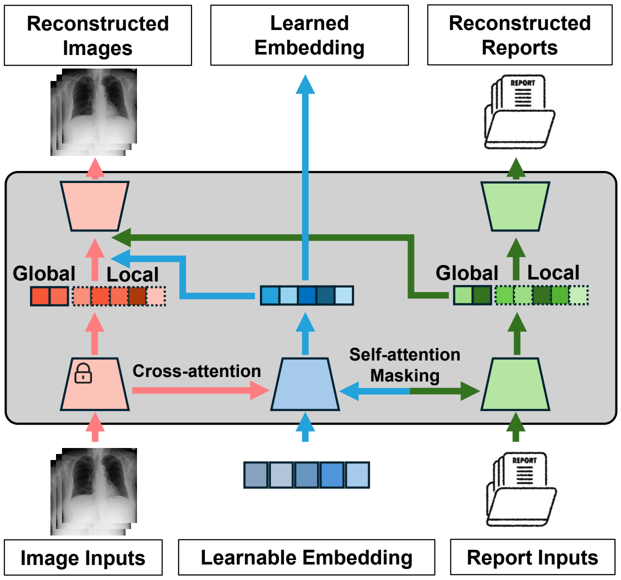

Figure 1. Our MedUnifier framework incorporates learnable embeddings to enable multi-modal interactions. The red components focus on the initial extraction of visual features and the reconstruction of medical images. The green elements are dedicated to the modelling and interpretation of medical reports. Meanwhile, the blue components apply a range of attention-masking strategies to achieve a comprehensive fusion of image and text representations.

integrated masked image modelling (MIM) into the VLP framework [7, 72], this approach does not fully enable the generation of comprehensive visual content nor capture detailed visual information effectively.

In this paper, we present **MedUnifier**, a unified VLP framework for medical data (Figure 1), designed to seamlessly integrate text-grounded image generation with advanced multi-modal learning strategies, including image-text contrastive alignment, image-text matching, and image-grounded text generation. Our approach begins with a Transformer model featuring learnable embeddings, inspired by BLIP-2 [36], a 12-layer transformer encoder as the trainable component, paired with a pre-trained visual encoder that embeds preliminary visual features. To extend its functionality for vision generation, we introduce a novel learning objective, termed text-grounded image generation (**TIG**) loss. This objective leverages vector-quantization to facilitate discrete visual representation learning [56], guiding vision generation using textual data. Additionally, we devise a novel latent adapter to connect the base model with the image generation module, enabling end-to-end co-training with three other learning objectives: image-text contrastive (**ITC**), image-text matching (**ITM**), and image-grounded text generation (**ITG**) losses. To our knowledge, **MedUnifier** is the first model to adapt learnable embeddings to the medical domain, bridging the gap between existing VLP paradigms and text-grounded image generation to enhance multi-modal alignment. The main contribution of this study is listed below:

- We introduce **MedUnifier**, a novel Med-VLP framework that unifies the current VLP paradigm with a language-guided visual generation task, marking a significant step toward an all-in-one VLP model that seamlessly integrates visual and linguistic information.
- We designed a novel TIG module to capture fine-grained details by recovering pixel-level information from hierarchical multi-modal representations, enabling the model to identify subtle visual details, commonly available in medical data (e.g. small nodules, slight opacities, etc).
- We perform a series of experiments on various downstream studies using Chest X-rays, showcasing performance enhancements over existing methods across uni-modality, cross-modality, and multi-modality tasks.
- We also demonstrate the model's adaptability in generating realistic medical images and reports, highlighting its unique capability to augment out-of-distribution datasets.

2. Related Work

2.1. Vision-and-language Pre-training (VLP)

Vision-language models (VLMs) have attracted considerable attention due to their powerful ability to integrate visual and textual data, significantly enhancing image captioning, visual question answering, and cross-modal retrieval. The predominant paradigms for VLP can be broadly classified into two main categories. The first one focuses on learning uni-modal encoders for text and images [26, 29, 62–64, 69, 71], respectively. However, this dual-encoder architecture limits the capacity to establish intricate interactions between text and image. Another line of work predominantly focuses on fusion encoder-based structure [3, 34, 55, 60] to facilitate meaningful interactions between the two modalities. In the medical setting, Chen *et al*. [10] proposed an effective framework to unify dual-encode style and fusion-encoder. However, these methods do not take the generation of visual information into consideration, and lack exploration of detailed vision content. In this study, we adopt and extend a fusion encoder-based framework to better align visual and textual features by incorporating vector quantization to enable the learning of discrete visual representations, thus facilitating effective vision generation guided by pertinent textual information. Compared to existing studies, our work aligns various modalities and simultaneously creates generic and versatile representations by leveraging the complementary strengths of various losses synergistically, therefore alleviating the additional pre-training stages for expert image tokenizer and iterative denoising.

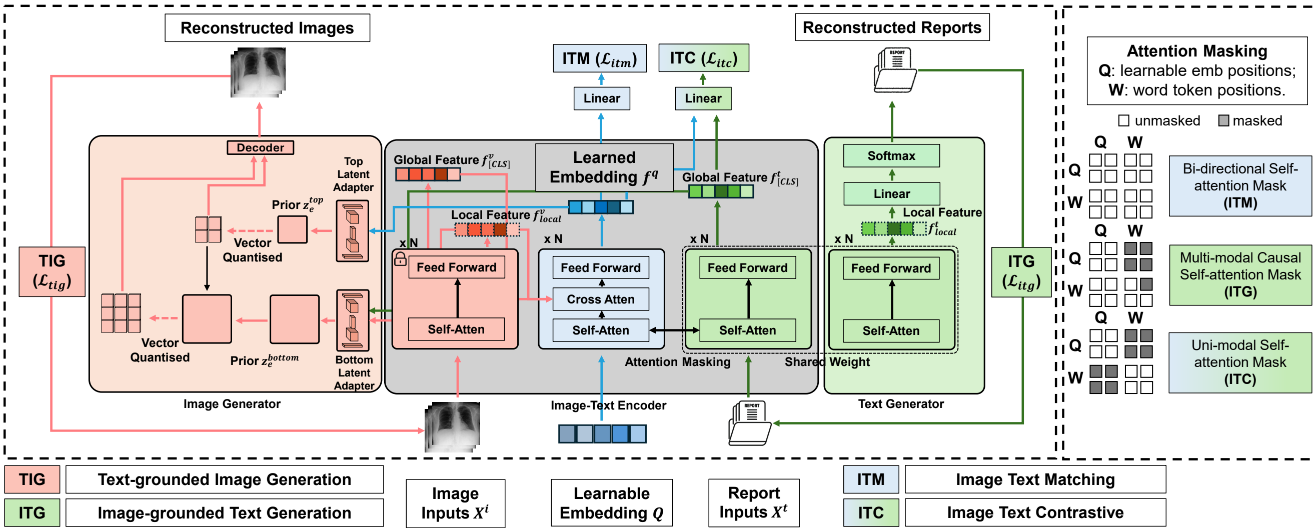

Figure 2. **Left**: model architecture consists of an image-text encoder, a text generator, and an image generator to extract the most relevant visual and textual representations by optimizing four distinctive loss functions (ITM, ITC, ITG, TIG). **Right**: self-attention masking strategies for different learning objectives. **Bottom**: detailed learning objectives. Integrating visual and textual information enables deep fusion through cross-modal interaction and allows each modality to be processed independently for uni-modal generation.

2.2. Text to Image (T2I) Generation

Text-to-image (T2I) tasks aim to generate an image according to a given textual description. Generative adversarial networks (GANs) [31, 39, 51, 68] and auto-regressive (AR) transformers [16, 19, 49, 65] were widely recognized for their exceptional performance and popularity. Advancements in variational auto-encoders (VAEs) and vector quantized VAEs (VQ-VAEs) have further improved T2I generation by introducing a discrete latent space for more stable generation [15, 20, 45, 49, 65]. LLM-CXR [33] incorporated pre-trained VQ-GAN into powerful LLM, predicting dual-modal tokens. Moreover, diffusion models [14, 25] have recently taken the leading position in T2I generation tasks. By adapting the advanced diffusion model, MedM2G [67] emphasized high-quality content generation. Despite their effectiveness, diffusion models are notably resource-intensive, often necessitating thousands of iterative steps for denoising, leading to significantly slower speeds. It is hard for diffusion models to get evident visual features due to modelling data distribution implicitly [54]. Consequently, for our T2I tasks, we turn to VQ-VAEs to learn more robust representations, thereby enhancing the quality and efficiency of medical image generation effectively.

3. Method

In this section, we introduce our **MedUnifier** framework for aggregating four key learning objects on Med-VLP. We first formulate the problem to be solved in §3.1 . Then we describe the process of extraction and fusion of multi-modal features using the base model in §3.3 with model architecture presented in §3.2. Lastly, we illustrate the integration of our proposed text-grounded image generation module and connection with the base model in §3.4.

3.1. Problem definition

We formulate the Med-VLP problem with inspiration from the previous studies [9, 10]. Formally, given a set of medical images $X^I \in \{x_1^i, x_2^i, \dots, x_n^i\}$ with corresponding clinical reports $X^T \in \{x_1^t, x_2^t, \dots, x_n^t\}$, the entire pre-training objective function can be defined as

$$\mathcal{L}_{total} = \sum_{m=1}^{M} \lambda_m \mathcal{L}_m(\mathcal{H}_m(\mathcal{F}(X^I, X^T))) \qquad (1)$$

where $\mathcal{F}$ represents backbone taking the paired $[x^i, x^t]$ as input. $\mathcal{H}_m$ stands for task-specific modules for further encoding visual and textual features. $\mathcal{L}_m$ and λ_m are different loss functions and their weights for the overall loss calculation with the total number of loss functions being M.

3.2. Model architecture

The proposed model mainly consists of three components, an image-text encoder, a text generator, and an image generator with cross-attention layer, masking strategies and vector discretization for extraction on the visual and textual representations, as shown in Figure 2.

Image-text encoder: We adopted BERT-styled Transformer [37] as image-text encoder network for fusing multi-modal information in depth. The image-text encoder is constructed by 12 transformer blocks. Its input contains a set of learnable embeddings and clinical reports tokenized by words [13]. Learnable embeddings and textual input interact with each other through self-attention layers [57] using different masking strategies for various pre-training tasks. We further processed input images into a set of patch embeddings by using a pre-trained, frozen Vision Transformer (ViT) from [18], attaining informative visual features while saving computational costs during training. Moreover, the initial visual embeddings engage with the image-text encoder network through cross-attention layers, which are integrated within each transformer block. At the final layer of the last transformer block, we incorporated distinct, task-specific heads tailored to address various tasks.

Text generator: To generate accurate and coherent medical reports, we duplicated the text encoder of the image-text encoder as a language-generative network with shared weights (coloured in green). A decoding head was further added on the top to map each word token embedding to the vocabulary dictionary.

Image generator: The primary objective of VLP is to learn representations that are transferable across tasks. While the most advanced vision generative models, such as GANs and DMs, excel at generating high-quality and diverse visual content, they typically capture data distributions implicitly, which these models make it challenging to explicitly access intermediate visual representations, limiting their utility in VLP applications. Furthermore medical data often display similar visual or textual patterns and hence encompass nuanced yet crucial medical pattern, which is often overlooked by the existing VLM approaches. To address these constraints, we integrated a vector-quantized variational auto-encoder (VQ-VAE) [56] within a cross-modal interactive fusion framework to generate high-quality synthetic visual content and improve performance across multiple downstream tasks.

3.3. Fusion of visual and textual features

Here, we outline the information flow within the image-text encoder, as well as describe three types of pre-training tasks (ITC, ITM, ITG) in detail.

Given that $x^i \in \mathbb{R}^{C\times H\times W}$, we divided the entire image into L_v patches with spatial size (h, w) through convolutional operation and added learnable positional encodings:

$$\boldsymbol{X}^i = [\boldsymbol{p}_{[CLS]}, \boldsymbol{p}_1, \boldsymbol{p}_2, \dots, \boldsymbol{p}_{L_v}] + \boldsymbol{E}^v_{pos} \tag{2}$$

We prepended $\boldsymbol{p}_{[CLS]}$ for aggregation of visual information. Then these patch embeddings get passed through a standard pre-trained ViT-g, denoted as E_I, to attain preliminary visual embeddings $\boldsymbol{f}^v \in \mathbb{R}^{(L_v+1)\times d_v}$:

$$E_I(\boldsymbol{X}^i) = \boldsymbol{f}^v = [\boldsymbol{f}^v_{[CLS]}, \boldsymbol{f}^v_{local}] \tag{3}$$

$$= [\boldsymbol{f}^v_{[CLS]}, \boldsymbol{f}^v_1, \boldsymbol{f}^v_2, \dots, \boldsymbol{f}^v_{L_v}] \tag{4}$$

where $\boldsymbol{f}^v_{[CLS]}$ is global visual feature, $\boldsymbol{f}^v_{local} \in \mathbb{R}^{L_v\times d_v}$ represent local visual features.

For the textual input, we followed BERT [13] to tokenize the input text to word embeddings, adding learnable positional encodings:

$$\boldsymbol{X}^t = [\boldsymbol{w}_{[SPE]}, \boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_{L_t}] + \boldsymbol{E}^t_{pos} \tag{5}$$

We prepended a special token $\boldsymbol{w}_{[SPE]}$, utilizing different $[SPE]$ tokens for identifying differing tasks.

In order to enable the interaction between word embeddings and preliminary visual embeddings, we constructed a set of learnable embeddings, denoted as $\boldsymbol{Q} = [\boldsymbol{q}_1, \boldsymbol{q}_2, \dots, \boldsymbol{q}_{L_q}], \boldsymbol{Q} \in \mathbb{R}^{L_q\times d_q}$. We unified word embeddings and learnable embeddings of the same feature dimension, e.g. $d_t = d_q$. Then we concatenated $\boldsymbol{Q}$ and $\boldsymbol{X}^t$ to form the input of the image-text encoder, denoted as E_Q, encoding it to get output embeddings:

$$E_Q([\boldsymbol{Q}, \boldsymbol{X}^t]) = [\boldsymbol{f}^q, \boldsymbol{f}^t] \tag{6}$$

$$= [\boldsymbol{f}^q, \boldsymbol{f}^t_{[SPE]}, \boldsymbol{f}^t_{local}] \tag{7}$$

We employed a cross-attention mechanism [1] to facilitate interaction between learnable embeddings and preliminary visual embeddings. This design enables $\boldsymbol{f}^q$ to function as the final visual representation. To ensure clarity over the fusion of learnable embeddings and textual representations, we implemented distinct masking strategies within the self-attention layers (Figure 2, right panel).

Image-text contrastive learning (ITC) The task aims to align visual and textual representations by maximizing their mutual information through a contrastive approach. To accomplish this, we replaced $\boldsymbol{w}[SPE]$ with $\boldsymbol{w}[CLS]$ to facilitate global textual representations denoted as $\boldsymbol{f}^t_{[CLS]} \in \mathbb{R}^{d_t}$. Furthermore, we implemented uni-modal masking (Figure 2, right panel) to enable learnable embeddings $\boldsymbol{Q}$ and textual embeddings $\boldsymbol{X}^t$ to attend exclusively to themselves. $\boldsymbol{f}^q$ and $\boldsymbol{f}^t_{[CLS]}$ are then linearly projected to representations as:

$$\boldsymbol{g}^q = \mathcal{H}^q_{itc}(\boldsymbol{f}^q) \tag{8}$$

$$\boldsymbol{g}^t = \mathcal{H}^t_{itc}(\boldsymbol{f}^t_{[CLS]}) \tag{9}$$

where $\mathcal{H}^q_{itc}, \mathcal{H}^t_{itc}$ are ITC heads. We computed the pairwise similarity between each visual and textual representation $\boldsymbol{g}^q$

and $\boldsymbol{g}^t$ and chose the highest one as the image-text similarity to calculate bi-directional contrastive loss:

$$\mathcal{L}_{itc}^{(q|t)} = \frac{1}{N}\sum_{k=1}^{N} -\log\left(\frac{\exp(\max < \boldsymbol{g}_k^q, \boldsymbol{g}_k^t > /\tau)}{\sum_{n=1}^{N}\exp(\max < \boldsymbol{g}_k^q, \boldsymbol{g}_n^t > /\tau)}\right) \tag{10}$$

$$\mathcal{L}_{itc}^{(t|q)} = \frac{1}{N}\sum_{k=1}^{N} -\log\left(\frac{\exp(\max < \boldsymbol{g}_k^q, \boldsymbol{g}_k^t > /\tau)}{\sum_{n=1}^{N}\exp(\max < \boldsymbol{g}_n^q, \boldsymbol{g}_k^t > /\tau)}\right) \tag{11}$$

where $\tau \in \mathbb{R}$ is a scaling temperature parameter initialized to 0.07, N is mini-batch size and $<\cdot,\cdot>$ represents the cosine similarity. The overall ITC loss is defined as:

$$\mathcal{L}_{itc} = \frac{1}{2}(\mathcal{L}_{itc}^{(q|t)} + \mathcal{L}_{itc}^{(t|q)}) \tag{12}$$

We here expanded visual representation space from a conventional single vector to a set of vectors e.g. $\boldsymbol{f}^q \in \mathbb{R}^{L_q \times d_q}$ which is different from [26, 63].

Image-text matching (ITM) This task aims to learn a precise alignment by classifying image-text pairs as either positive or negative. We implemented a bi-directional mask (Figure 2, right panel) that allows all learnable embeddings and word token embeddings to attend to one another. The resulting output of learned embeddings denoted as $\boldsymbol{f}^q$, capture enriched multi-modal information. These tokens were then fed into a two-class linear classifier, $\mathcal{H}_{itm}$, where the outputs are averaged across learned embeddings to generate a logit and compute the Image-Text Matching (ITM) loss:

$$\mathcal{L}_{itm} = \frac{1}{N}\sum_{k=1}^{N} -\log(p(Y_k|\hat{Y}_k)) \tag{13}$$

$$\hat{Y} = \frac{1}{L_q}\sum_{i=1}^{L_q} \mathcal{H}_{itm}(\boldsymbol{f}_i^q), \tag{14}$$

Y represents ground truth labels within mini-batch by hard negative samples mining, as stated in [34].

Image-grounded text generation (ITG) This task is to generate text conditioned on paired images. To achieve a coherent and precise generation of medical reports within a unified VLP framework, we chose CLM [5, 47] where each word token attends only to preceding tokens, following a GPT-style language model architecture [5]. Inspired by UniLM [17], we implemented a multi-modal causal self-attention mask (Figure 2, right panel). We replaced the special token $\boldsymbol{w}[SPE]$ with $\boldsymbol{w}[DEC]$ to signal a decoding task. We also introduced a word prediction head, denoted as $\mathcal{H}_{itg}$. This learning objective is formalized as:

$$\mathcal{L}_{itg} = \frac{1}{NL_t}\sum_{k=1}^{N}\sum_{i=1}^{L_t} -\log(p_i) \tag{15}$$

$$p_i = Softmax(\mathcal{H}_{itg}(\boldsymbol{f}_{local}^t)) \tag{16}$$

$$= p(\boldsymbol{w}_i|\boldsymbol{Q}, \ldots, \boldsymbol{w}_{i-1}) \tag{17}$$

3.4. Text-grounded image generation (TIG)

We also introduce an innovative and efficient module designed for the text-grounded image generation task, with integration into the previously discussed image-text encoder and text generator for a versatile Med-VLP model.

Rethinking VQ-VAE The VQ-VAE [56] offers a significant advantage over other generative models due to its ability to explicitly learn discrete visual representations. This feature aligns closely with the image-text encoder, which similarly learns discrete representations from a dictionary of learnable embeddings. Motivated by this, we chose to adopt VQ-VAE as the image generator in our framework, which forms a different unified Med-VLP framework from other studies [7, 10, 72]. Additionally, the image-text encoder generates two distinct types of visual features: one for abstract visual representations, denoted as $\boldsymbol{f}^q$, and another for fine-grained, local visual embeddings, $\boldsymbol{f}_{local}^v$. Consequently, the image-text encoder can be viewed as a powerful multi-modal encoder, akin to the image encoder in conventional VAEs [24, 56]. Inspired by the work of [50], we developed hierarchical vector quantizers and image decoders. Figure 2 further illustrates the proposed text-grounded image generator. Our TIG module is specifically designed to capture fine-grained details through recovering pixel-level information from hierarchical multi-modal representations, enabling the model to identify subtle visual details.

Bridging the gap The features $\boldsymbol{f}^q \in \mathbb{R}^{L_q \times d_q}$ contains textual implications derived from the co-training image-text encoder, which we regard as the top latent representation, denoted $\boldsymbol{z}^{top} \in \mathbb{R}^{L_q \times d_q}$. In contrast, $\boldsymbol{f}_{local}^v \in \mathbb{R}^{L_v \times d_v}$ does not include textual information. Therefore, we concatenate $\boldsymbol{f}_{local}^v$ with the aggregated textual representation, $\boldsymbol{f}_{[CLS]}^t \in \mathbb{R}^{1 \times d_t}$, along the feature dimension, resulting in the multi-modal bottom latent representation $\boldsymbol{z}^{bot} \in \mathbb{R}^{L_v \times (d_v + d_t)}$.

At the top level, we devise the latent adapter, denoted as $\mathcal{Z}_{top}$, for transforming $\boldsymbol{z}^{top}$ into spatial feature map $\boldsymbol{z}_e^{top}$:

$$\boldsymbol{z}_e^{top} = \mathcal{Z}_{top}(\boldsymbol{z}^{top}) \tag{18}$$

where $\mathcal{Z}_{top}$ consisted of a nonlinear transformation, spatial positional encoding summer [61] and a residual block [22]. The illustration can be found in appendix 7. We reshaped $\boldsymbol{z}^{top}$ and then pass it through $\mathcal{Z}_{top}$, reaping $\boldsymbol{z}_e^{top} \in$

$\mathbb{R}^{d_{top} \times h_{top} \times w_{top}}$. Followed by a vector quantization layer with a latent embedding space e^{top}, we gain discrete feature map $\boldsymbol{z}_q^{top} \in \mathbb{R}^{d_{top} \times h_{top} \times w_{top}}$:

$$\boldsymbol{z}_q^{top} = quantizer_{top}(\boldsymbol{z}_e^{top}) \tag{19}$$

At the bottom level, in accordance with the top level, a latent adapter and vector quantizer with a latent embedding space e^{bottom} are deployed to gain a discrete feature map:

$$\boldsymbol{z}_e^{bot} = \mathcal{Z}_{bot}(\boldsymbol{z}^{bot}) \tag{20}$$

$$\boldsymbol{z}_q^{bot} = quantizer_{bot}(\boldsymbol{z}_e^{bot}, \boldsymbol{z}_q^{top}) \tag{21}$$

where $\boldsymbol{z}^{bot}$ are viewed as having spatial size $h_{bot} \times w_{bot} = \frac{H}{h} \times \frac{W}{w}$. Producing $\boldsymbol{z}_q^{bottom}$ is conditioning on $\boldsymbol{z}_q^{top}$.

We built hierarchical decoders $\mathcal{D}$ to recover raw images from discrete multi-modal representations:

$$\hat{x}^i = \mathcal{D}(\boldsymbol{z}_q^{top}, \boldsymbol{z}_q^{bot}) \tag{22}$$

The text-grounded image generation (TIG) loss is formulated as:

$$\begin{aligned}\mathcal{L}_{tig} = \frac{1}{N}\sum_{k=1}^{N} -\log p(x_k^i | \boldsymbol{z}_q^{top}, \boldsymbol{z}_q^{bot}) \\ + \left\|\mathrm{sg}[\boldsymbol{z}_e^{top}] - e^{top}\right\|_2 + \beta_1 \left\|\mathrm{sg}[e^{top}] - \boldsymbol{z}_e^{top}\right\|_2 \\ + \left\|\mathrm{sg}[\boldsymbol{z}_e^{bot}] - e^{bot}\right\|_2 + \beta_2 \left\|\mathrm{sg}[e^{bot}] - \boldsymbol{z}_e^{bot}\right\|_2\end{aligned} \tag{23}$$

where the negative logarithmic term can be written as mean square error (MSE) $\left\|x_k^i - \hat{x}_k^i\right\|_2$, sg$[\cdot]$ is gradient stop operation. Hyper-parameters β_1, β_2 are both set to be 0.5.

3.5. Total learning objectives

We provide a comprehensive summary of all the learning objectives and present the ultimate loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{itc} + \lambda_2 \mathcal{L}_{itm} + \lambda_3 \mathcal{L}_{itg} + \lambda_4 \mathcal{L}_{tig} \tag{24}$$

Four weights λ were set to 1 in experiments. These weights were determined by ablation study (see appendix 11).

4. Experiments

We perform the pre-training on the current largest multi-modal medical dataset, MIMIC-CXR v2.0.0 [30] and evaluate our model on various downstream tasks, followed by an ablation study for probing purpose, which shows the proposed method's superiority.

4.1. Implementation details

We employed a BERT model as the primary network for the image-text encoder and utilized ViT-g as the pre-trained ViT. The input image resolution was set to 224 × 224, with a maximum text length of 95 tokens, and 32 learnable embeddings. The top and bottom codebook size were both set to 512 with feature dimension of 768. For optimization, we applied the AdamW optimizer [43] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, and a weight decay of 0.05. A cosine learning rate decay schedule was used, with a peak learning rate of 1e-4. We incorporated a warm-up phase for the initial 5% of training steps, starting with a learning rate of 1e-5. The pre-training process was conducted on four NVIDIA A100 GPUs.The implementation details are further elaborated in appendix 8.

4.2. Medical Vision-and-Language Benchmark

To assess the effectiveness of our method, we conducted experiments across three types of tasks: uni-modal, cross-modal, and multi-modal. We conducted experiments on MedUnifier with/without TIG and compared with previous studies as our main experimental results. All benchmark datasets used in experiments pertain specifically to radiology with detailed descriptions in the appendix 9. For baselines, we took reference from papers and implemented them using official codes: ConVIRT[69], GLoRIA[26], BioViL[4], LoVT [44], MedCLIP[62], MGCA [58], PRIOR [11], REFERS [71], CXR-CLIP [64], ViLT[2], R2Gen [9], PTUnifier [10], DCL [38], MOTOR [41], UniXGer[32], RoentGen[6], LLM-CXR[33].

Uni-modal tasks assess the learned visual representations for image modality by applying uni-modal masking within classification scope on datasets such as RSNA Pneumonia [53], SIIM-ACR [66], and COVIDx [59]. To examine the model's data efficiency, we fine-tune it using different proportions of the training data (1%, 10%, or 100%).

Cross-modal tasks require models to align vision and language modalities. We conduct experiments across three tasks: image-to-text retrieval (ITR), text-to-image retrieval (TIR), and zero-shot image classification (ZS). For ITR and TIR, we report cross-modal information retrieval metrics, the mean Average Precision at K (mAP@K), using the MIMIC 5x200 dataset. In the ZS task, we employ MIMIC 5x200, CheXpert 5x200 [11, 26], and draw 500 positive and 500 negative samples from the full RSNA Pneumonia dataset [53] for evaluation purpose.

Multi-modal tasks generate uni-modal content through multi-modal interaction. We carry out two kinds of experiments including image-grounded text generation (ITG) and text-grounded image generation (TIG). For ITG, we adopt the MIMIC-CXR held-out test set to evaluate the quality of generated reports. Standard natural language generation (NLG) criteria are used to assess the performance, including BLEUn [46], METEOR [12], and ROUGE-L [42]. For TIG, following [50], we first train two Pixelsnail models [8] to model multi-modal priors $\boldsymbol{z}_q^{top}, \boldsymbol{z}_q^{bot}$. Then we sample

Methods	RSNA (AUC)			SIIM (AUC)			COVIDx (ACC)		
	1%	10%	100%	1%	10%	100%	1%	10%	100%
ConVIRT[69]	84.2	86.9	88.7	84.2	85.7	91.5	72.5	82.5	92.0
GLoRIA[26]	84.1	86.8	89.1	85.1	88.5	92.1	66.5	80.5	88.0
BioViL[4]	82.0	85.4	88.6	79.8	81.6	90.5	-	-	-
LoVT [44]	85.1	86.5	89.3	85.5	88.5	92.2	-	-	-
MGCA [58]	85.8	87.7	89.2	86.1	89.6	92.0	74.8	84.8	92.3
PRIOR [11]	85.7	87.1	89.2	87.2	89.1	92.3	-	-	-
MedUnifer (w/o TIG)	86.3	87.1	88.1	87.6	88.7	92.3	75.3	87.5	92.8
MedUnifer (w TIG)	**87.6**	**88.8**	**91.7**	**87.9**	**92.5**	**94.8**	**76.8**	**88.3**	**93.5**

Table 1. Fine-tuned image classification results on RSNA, SIIM and COVIDx with 1%, 10%, 100% training data. Area under ROC curve (AUROC [%]) are reported for RSNA and SIIM datasets, and accuracy (ACC [%]) is reported for COVIDx dataset. The best and second-best results are highlighted in bold and underlined, respectively. Our method achieves the best performance across all datasets.

Methods	Image-Text Retrieval (ITR)			Text-Image Retrieval (TIR)		
	mAP@1	mAP@5	mAP@10	mAP@1	mAP@5	mAP@10
ConVIRT[69]	46.5	53.9	53.8	20.0	45.4	35.5
GLoRIA[26]	46.7	56.4	55.0	51.8	59.5	58.9
BioViL[4]	47.3	57.7	55.6	54.6	64.3	62.8
MedCLIP [62]	47.6	58.0	55.9	56.3	69.9	66.7
MGCA [58]	47.1	57.4	55.4	53.1	61.9	61.1
REFERS [71]	52.4	59.9	58.6	60.6	**71.9**	**69.0**
CXR-CLIP [64]	51.8	61.2	58.5	60.2	69.2	64.6
MedUnifier (w/o TIG)	57.4	65.4	60.3	59.6	68.3	63.4
MedUnifier (w TIG)	**60.7**	**66.6**	**61.7**	**63.1**	70.8	64.4

Table 2. Cross-modal retrieval results on MIMIC-CXR 5x200 dataset. The top K (1, 5, 10) mean Average Precision metrics are reported. Our method achieves the best performance for ITR tasks.

Methods	MIMIC 5x200 ACC	CheXpert 5x200 ACC	RSNA ACC
ConVIRT[69]	43.8	35.2	77.4
GLoRIA[26]	47.5	**45.0**	68.3
BioViL[4]	48.5	42.2	77.1
MedCLIP[62]	47.1	41.1	81.8
MGCA [58]	48.0	40.9	76.2
REFERS [71]	49.5	41.8	78.0
CXR-CLIP [64]	49.7	35.9	76.9
MedUnifer (w/o TIG)	44.8	40.8	**85.0**
MedUnifer (w TIG)	**50.4**	43.5	82.0

Table 3. Zero-shot image classification results on MIMIC 5x200, CheXpert 5x200 and RSNA datasets. Our method achieves the best performance for MIMIC 5x200 and RSNA datasets. Note that GLoRIA is trained on the CheXpert dataset.

latent encodings from both priors and generate new medical images by decoding latent encodings. For quantitative analysis, we present FID [23] scores.

4.3. Results and Analyses

To validate the effectiveness of MedUnifier, we conduct experiments on the above vision-and-language benchmark. The results of the main experiments are presented in Table 1,2, 3, 4 5 and Figure 6. We observe several noteworthy findings in our results. First, our model outperforms prior studies on uni-modal tasks across various downstream datasets, as shown in Table 1. This improvement suggests that integrating TIG significantly enhances the model's ability to learn more transferable visual representations. Second, for cross-modal retrieval, MedUnifier model achieves the highest performance, demonstrating a superior ability to understand and integrate cross-modal content compared to other models (see Table 2). The obtained results further highlight the better performance of MedUnifier, suggesting that our approach supports the necessary complementary semantic data for cross-modal retrieval. We also observe that excluding the TIG module slightly decreases MedUnifier's effectiveness in text-image retrieval tasks, which may be a result of an over-reliance on ITG that leads to an imbalance in vision-language fusion. In addition, MedUnifier demonstrates better performance on zero-shot classification tasks for both the MIMIC 5x200 and RSNA datasets in Table 3. However, GLoRIA outperforms our model on the CheXpert 5x200 dataset, likely due to its pre-training on the full CheXpert dataset with accompanying medical reports. Nonetheless, the results indicate the effectiveness of employing prompt ensembles within the proposed method, leading to enhanced overall performance improvements. Table 4 demonstrates that our models, both with and without the TIG module, surpass previous methods for image-grounded medical report generation. The Med-VLP framework also gains substantial advantages from incorporating causal language modelling, which together contributes to its improved performance. In Figure 3, we provide a comparison of the generated report and ground truth report. Finally, we conduct both quantitative and qualitative analyses on text-grounded image generation tasks in Table 5 and Figure 6 (see appendix). It highlights that MedUnifier with TIG achieve comparable performance for medical vision

Methods	MIMIC-CXR test set					
	BL-1	BL-2	BL-3	BL-4	MTR	RG-L
DCL[38]	-	-	-	10.9	15.0	28.4
MOTOR[41]	-	21.3	-	**15.6**	19.3	<u>31.4</u>
ViLT[2]	-	21.3	-	9.2	-	29.6
R2Gen [9]	35.3	21.8	14.5	10.3	14.2	27.7
PTUnifier [10]	-	-	-	10.7	-	-
MedUnifier (w/o TIG)	<u>40.1</u>	<u>24.4</u>	<u>16.0</u>	11.3	<u>18.8</u>	29.6
MedUnifier (w TIG)	**41.5**	**26.3**	**18.1**	<u>12.9</u>	**22.7**	**34.0**

Table 4. The performance of all baselines and our method on the test set of MIMIC-CXR dataset for Natural Language Generation (NLG) metrics. BL-n denotes BLEU score using up to n-grams; MTR and RG-L denote METEOR and ROUGE-L, respectively. Our method achieves the best performance across all metrics.

generation tasks with less model's complexity (one-stage pre-training) then existing models [6, 32, 33], all of which require additional pre-training for image tokenizer or iterative denoising. From a direct visual inspection, the reconstructed visual samples are nearly indistinguishable from authentic radiographs. Moreover, synthetic samples generated from multi-modal priors demonstrate high diversity, highlighting their potential to augment out-of-distribution medical data effectively.

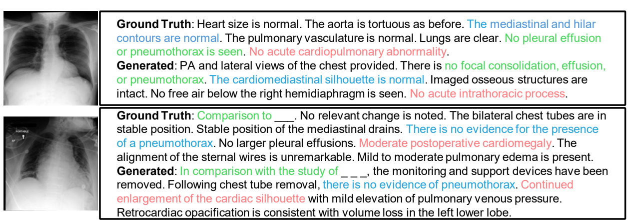

Figure 3. Comparison of ground truth and generated radiology reports reveals strong semantic alignment. In the top figure, both reports describe normal heart size, no pneumothorax or pleural effusion, and a normal cardiomediastinal silhouette, with the generated text adding details on osseous structures/intrathoracic processes. In the bottom figure, both reports align on pneumothorax and cardiomegaly. The same colours denote matched content between the generated sequences and the ground truth report.

4.4. Ablation study

To demonstrate the efficacy of our proposed method, we perform an ablation study (Table 6) across various learning objectives. The results indicate that using only the ITC loss (ID 1) yields the lowest performance. As ITM and ITG objectives are incrementally incorporated, performance gradually improves (IDs 2 and 3), highlighting how refined cross-modal alignment and multi-modal language modelling enhance the model's overall capabilities. Interestingly, the model with TIG (ID 4) surpasses the one with ITG (ID 3). We attribute this phenomenon to the relative difficulty

Methods	FID ↓	Methods	FID ↓
UniXGer[32]	78.2	Validation set	17.2
RoentGen[6]	<u>42.4</u>	Reconstruction	27.2
LLM-CXR[33]	**22.8**	MedUnifier	46.2

Table 5. FID score [23] on reconstructed and synthetic images, using the MIMIC-CXR pre-training dataset as the reference dataset. The FID score using MIMIC-CXR validation is also provided for a fair comparison. Reconstructed images are generated by passing the training set through our MedUnifier. Synthetic images are produced from sampled and decoded latent encodings using the trained Pixelsnail models and VAE decoder.

ID	Learning Objectives				ITR	Zero-shot cls	Fine-tuned cls		
	ITC	ITM	ITG	TIG	mAP@1	MIMIC 5x200 (ACC)	RSNA (AUC)	SIIM (AUC)	COVID (ACC)
1	✓				53.7	41.4	87.0	89.4	90.8
2	✓	✓			55.2	44.3	87.1	89.8	91.5
3	✓	✓	✓		57.4	44.8	88.1	92.3	92.8
4	✓	✓		✓	58.5	46.2	89.1	92.6	91.3
5	✓	✓	✓	✓	**60.7**	**50.4**	**91.7**	**94.8**	**93.5**

Table 6. Ablation studies on the different modules. The best performance is achieved using all objectives. Details see appendix 11.

of generating pixel-level image representations guided by text, as compared to generating word-level representations guided by visual input, which leads the model to learn more abstract, well-aligned representations. Ultimately, the integration of all objective types (ID 5) enables the model to achieve optimal performance, underscoring the viability of incorporating vision generation into existing frameworks.

5. Conclusion

In this paper, we introduce a novel and unified Med-VLP model, MedUnifier, which optimizes four distinct learning objectives simultaneously. By leveraging learnable embeddings and encoding raw images through a pre-trained Vision Transformer (ViT), MedUnifier circumvents the need to learn visual embeddings from scratch. Additionally, a VQ-VAE-based text-grounded image generation task is further incorporated into the Med-VLP framework to enhance its representation learning capacity. It reconstructs pixel-level visual details from both image and report, facilitating fine-grained visual understanding commonly available in medical data (subtle visual details e.g. small nodules, slight opacities, etc.) and efficient use of multi-modal representations through hierarchical latent adapters of dynamically adjusting the abstraction for each mode. Our proposed method effectively complements existing Med-VLP frameworks and achieves state-of-the-art performance. Our work also has significant implications for enhancing the VLP development for radiological applications.

6. Acknowledgement

This project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG25/24 and RS16/23), as well as the Lee Kong Chian School of Medicine - Ministry of Education Start-Up Grant. This research is also partially supported by the computational resources and staff contributions of the Quest high-performance computing facility at Northwestern University, jointly funded by the Office of the Provost, the Office for Research, and Northwestern University Information Technology.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 4

[2] Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023. 6, 8

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 1, 2

[4] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 6, 7, 1

[5] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 5

[6] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 6, 8

[7] 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, 5

[8] Xi Chen, Nikhil Mishra, Mostafa Rohaninejad, and Pieter Abbeel. Pixelsnail: An improved autoregressive generative model. In *International conference on machine learning*, pages 864–872. PMLR, 2018. 6, 1

[9] Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. *arXiv preprint arXiv:2010.16056*, 2020. 3, 6, 8

[10] Zhihong Chen, Shizhe Diao, Benyou Wang, Guanbin Li, and Xiang Wan. Towards unifying medical vision-and-language pre-training via soft prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23403–23413, 2023. 1, 2, 3, 5, 6, 8

[11] 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*, pages 21361–21371, 2023. 6, 7, 1, 2

[12] Michael Denkowski and Alon Lavie. Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems. In *Proceedings of the sixth workshop on statistical machine translation*, pages 85–91, 2011. 6

[13] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 4

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3

[15] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. Cogview: Mastering text-to-image generation via transformers. *Advances in neural information processing systems*, 34:19822–19835, 2021. 3

[16] Ming Ding, Wendi Zheng, Wenyi Hong, and Jie Tang. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *Advances in Neural Information Processing Systems*, 35:16890–16902, 2022. 3

[17] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32, 2019. 5

[18] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19358–19369, 2023. 4, 1

[19] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022. 3

[20] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706, 2022. 3

[21] Sylvain Gugger, Lysandre Debut, Thomas Wolf, Philipp Schmid, Zachary Mueller, Sourab Mangrulkar, Marc Sun, and Benjamin Bossan. Accelerate: Training and inference at scale made simple, efficient and adaptable. https:

//github.com/huggingface/accelerate, 2022. 1

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[23] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7, 8

[24] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017. 5

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3

[26] 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. 1, 2, 5, 6, 7

[27] Jaeyoung Huh, Sangjoon Park, Jeong Eun Lee, and Jong Chul Ye. Improving medical speech-to-text accuracy using vision-language pre-training models. *IEEE Journal of Biomedical and Health Informatics*, 2023. 1

[28] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana 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. 1

[29] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[30] 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. 6, 1

[31] Minguk Kang, Jun-Yan Zhu, Richard Zhang, Jaesik Park, Eli Shechtman, Sylvain Paris, and Taesung Park. Scaling up gans for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10124–10134, 2023. 3

[32] Hyungyung Lee, Da Young Lee, Wonjae Kim, Jin-Hwa Kim, Tackeun Kim, Jihang Kim, Leonard Sunwoo, and Edward Choi. Vision-language generative model for view-specific chest x-ray generation. *arXiv preprint arXiv:2302.12172*, 2023. 6, 8

[33] Suhyeon Lee, Won Jun Kim, Jinho Chang, and Jong Chul Ye. Llm-cxr: Instruction-finetuned llm for cxr image understanding and generation. *arXiv preprint arXiv:2305.11490*, 2023. 3, 6, 8

[34] 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, 5

[35] 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

[36] 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. 2

[37] 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 *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3

[38] Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3334–3343, 2023. 6, 8

[39] Zhiheng Li, Martin Renqiang Min, Kai Li, and Chenliang Xu. Stylet2i: Toward compositional and high-fidelity text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18197–18207, 2022. 3

[40] Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: language meets vision transformer in medical image segmentation. *IEEE transactions on medical imaging*, 2023. 1

[41] Bingqian Lin, Zicong Chen, Mingjie Li, Haokun Lin, Hang Xu, Yi Zhu, Jianzhuang Liu, Wenjia Cai, Lei Yang, Shen Zhao, et al. Towards medical artificial general intelligence via knowledge-enhanced multimodal pretraining. *arXiv preprint arXiv:2304.14204*, 2023. 6, 8

[42] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 6

[43] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6, 1

[44] 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. 6, 7

[45] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 6

[47] Alec Radford. Improving language understanding by generative pre-training. 2018. 5

[48] 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

[49] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 3

[50] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019. 5, 6

[51] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016. 3

[52] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: artificial intelligence*, 2(3):e190043, 2020. 1

[53] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 6, 1

[54] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[55] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019. 2

[56] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2, 4, 5

[57] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 4

[58] 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. 6, 7, 2

[59] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: a tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):19549, 2020. 6, 2

[60] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[61] Zelun Wang and Jyh-Charn Liu. Translating math formula images to latex sequences using deep neural networks with sequence-level training, 2019. 5

[62] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022*, 2022. 1, 2, 6, 7

[63] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21372–21383, 2023. 5

[64] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer, 2023. 1, 2, 6, 7

[65] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 3

[66] Anna Zawacki, Carol Wu, George Shih, Julia Elliott, Mikhail Fomitchev, Mohannad Hussain, ParasLakhani, Phil Culliton, and Shunxing Bao. Siim-acr pneumothorax segmentation. https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation, 2019. Kaggle. 6, 2

[67] Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11502–11512, 2024. 3

[68] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842, 2021. 3

[69] 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. 1, 2, 6, 7

[70] Yi Zhong, Mengqiu Xu, Kongming Liang, Kaixin Chen, and Ming Wu. Ariadne's thread: Using text prompts to improve segmentation of infected areas from chest x-ray images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 724–733. Springer, 2023. 1

[71] Hong-Yu Zhou, Xiaoyu Chen, Yinghao Zhang, Ruibang Luo, Liansheng Wang, and Yizhou Yu. Generalized radio-

graph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022. 1, 2, 6, 7

[72] Hong-Yu Zhou, Chenyu Lian, Liansheng Wang, and Yizhou Yu. Advancing radiograph representation learning with masked record modeling. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 5