

# PTUNIFIER: PSEUDO TOKENS AS PARADIGM UNIFIERS IN MEDICAL VISION-AND-LANGUAGE PRE-TRAINING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Medical vision-and-language pre-training (Med-VLP) has shown promising improvements on many downstream medical tasks owing to its applicability to extracting generic representations from medical images and texts. Practically, there exist two typical paradigms, i.e., the **fusion-encoder paradigm** and the **dual-encoder paradigm**, depending on whether a heavy fusion module is used. The former outperforms on multi-modal tasks owing to the sufficient interaction between modalities; the latter outperforms on uni-modal and cross-modal tasks due to the single-modality encoding ability. To take advantage of these two paradigms, we propose an effective yet straightforward scheme named PTUnifier to unify the two paradigms thanks to the identical input format by introducing visual and textual pseudo tokens, which serve as a feature bank that stores the most representative images/texts. By doing so, a single model could process various tasks adopting different input formats (i.e., image-only, text-only, and image-text-pair). Furthermore, we construct a pool of pseudo tokens (instead of static ones) to improve diversity and scalability. Experimental results show that our approach achieves state-of-the-art results on a broad range of tasks, spanning uni-modal tasks (i.e., image/text classification and text summarization), cross-modal tasks (i.e., image-to-text generation and image-text/text-image retrieval), and multi-modal tasks (i.e., visual question answering), demonstrating the effectiveness of our approach. Note that the adoption of pseudo tokens is orthogonal to most existing Med-VLP approaches, and we believe that our approach could be a beneficial and complementary extension to these approaches.<sup>1</sup>

## 1 INTRODUCTION

Medical data is multi-modal in general, among which vision and language are two critical modalities. It includes visual data (e.g., radiography, magnetic resonance imaging, and computed tomography) and textual data (e.g., radiology reports and medical texts). More importantly, such images and texts are pair-collected in routine clinical practice (e.g., X-ray images and their corresponding radiology reports). Medical vision-and-language pre-training (Med-VLP) aims to learn generic representation from large-scale medical image-text pairs and then transfer it to various medical tasks. Med-VLP is beneficial in addressing the data scarcity problem in the medical field.

Recently, substantial progress has been made toward research on Med-VLP (Zhang et al., 2020; Li et al., 2020b; Huang et al., 2021; Khare et al., 2021; Moon et al., 2021). In general, most existing Med-VLP models can be classified into two paradigms: the dual-encoder paradigm and the fusion-encoder paradigm, where the former encodes images and texts separately to learn *cross-modal* representations following a shallow interaction layer (i.e., an image-text contrastive layer), and the latter performs an early fusion of the two modalities through the self-attention/co-attention mechanisms to learn *multi-modal* representations.<sup>2</sup>

For dual-encoders, the purpose of existing studies (Zhang et al., 2020; Huang et al., 2021; Müller et al., 2021) is to develop label-efficient algorithms to learn effective *uni-modal/cross-modal* representations

<sup>1</sup>Our code will be released in the final version of this paper.

<sup>2</sup>Although the terminologies “cross-modal” and “multi-modal” have been used interchangeably in the literature, we treat them as terms with different meanings in this paper.

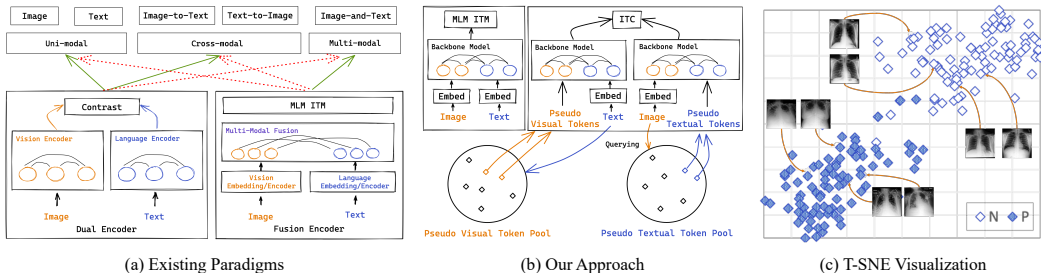


Figure 1: (a) Illustrations of two Med-VLP paradigms and their advantages (pointed by green arrows) and disadvantages (pointed by red arrows) in downstream tasks; (b) The overall architecture of our proposed approach, where the backbone models share the same parameters, and we duplicate them for illustration; (c) T-SNE visualization of the [CLS] representations of pseudo texts, where P and N denote the cases with and without abnormalities, respectively.

and the learned representations can improve the effectiveness of uni-modal (*i.e.*, vision-only or language-only) tasks<sup>3</sup> and the efficiency of cross-modal (*i.e.*, image-to-text or text-to-image) retrieval tasks significantly. For fusion-encoders, existing studies (Li et al., 2020b; Khare et al., 2021; Moon et al., 2021) aim to jointly process these two modalities with an early interaction to learn *multi-modal* representations to solve those tasks requiring multi-modal reasoning (*e.g.*, medical visual question answering and medical image-text classification). However, it seems that “*you can’t have your cake and eat it, too.*”: the fusion-encoders can not perform uni-modal tasks effectively and cross-modal tasks efficiently due to the lack of single modal encoding, while the dual-encoders underperform on multi-modal tasks owing to the insufficient interaction between modalities as shown in Figure 1(a).

In this paper, we aim to learn a unified medical vision-and-language pre-trained model. Although there exist some solutions (*e.g.*, mixture-of-modality experts (Wang et al., 2021) and a modified VLP architecture (Singh et al., 2022)) to achieve a similar goal in the general domain, we propose an architecture- and task-agnostic approach named PTUnifier (as shown in Figure 1(b)), which is much simpler and lighter-weight. Technically, we develop the approach from the following perspectives: (i) *Compatibility*: we introduce visual and textual pseudo tokens to make the Med-VLP model compatible with different kinds of inputs (*i.e.*, image-only inputs, text-only inputs, and image-text pairs); (ii) *Scalability*: we improve the diversity of the pseudo tokens by constructing pseudo token pools for different modalities from which different inputs are able to select their corresponding pseudo tokens, which enhances the capacity and makes it scalable to larger-scale Med-VLP. In intuition, the introduced pseudo tokens serve as a feature bank that stores the most representative images/texts, which is qualitatively analyzed by our experiments shown in Figure 1(c), where the [CLS] representations of pseudo texts are grouped into two clusters according to whether the images contain abnormalities or not.

As a result, the proposed approach can be employed in unifying Med-VLP with many existing VLP model architectures (*e.g.*, classic ones (Li et al., 2019; Dou et al., 2021) or even a single vanilla Transformer model) and does not require extra modality-dependent architectures, resulting in better applicability. We perform the pre-training on three large-scale medical image-text datasets, *i.e.*, ROCO (Pelka et al., 2018), MedICaT (Subramanian et al., 2020), and MIMIC-CXR (Johnson et al., 2019). To verify the effectiveness of our approach and facilitate further research, we construct a medical vision-language benchmark including *uni-modal* tasks (*i.e.*, image classification (IC) for vision and text classification (TC) and text summarization (TS) for language), *cross-modal* tasks (*i.e.*, image-to-text retrieval (ITR), text-to-image retrieval (TIR), and image-to-text generation<sup>4</sup> (ITG)), and *multi-modal* tasks (*i.e.*, visual question answering (VQA)). The proposed PTUnifier achieves state-of-the-art performance on all datasets, demonstrating its effectiveness.

<sup>3</sup>It is worth noting that most existing studies only conduct the evaluation on the vision-only tasks and disregard the language-only tasks although the text representations are simultaneously learned in the image-text contrastive procedure.

<sup>4</sup>Medical image-to-text generation refers to medical/radiology report generation in previous studies.

## 2 UNIFYING MED-VLP

§ 2.1 details the problem to be addressed. This work proposes to unify inputs using pseudo tokens (in § 2.2), and thus one could jointly train various tasks even with different input formats (in § 2.3).

### 2.1 PROBLEM DEFINITION

When dealing with vision and language modalities, suppose that we have embedded a medical image  $I$  as  $\mathbf{X}^v \in \mathbb{R}^{D_v \times N_v}$  or a medical text  $T$  as  $\mathbf{X}^l \in \mathbb{R}^{D_l \times N_l}$ . The input could be one of the following cases: (i) image-only input, (ii) text-only input, and (iii) image-text pair.

$$\mathbf{X} = \begin{cases} (\mathbf{X}^v) & \text{if } \textit{image-only} \\ (\mathbf{X}^l) & \text{if } \textit{text-only} \\ (\mathbf{X}^v, \mathbf{X}^l) & \text{if } \textit{image-text} \end{cases} \quad (1)$$

We aim to learn a single backbone model (denoted as  $\mathcal{M}_\theta$ , which is parameterized by  $\theta$ ) to process these three inputs indiscriminately. The challenge is to make **the backbone model  $\mathcal{M}_\theta$  deal with such variable-size and heterogeneous input**. To overcome this, one could unify various vision-language tasks. Assuming there are  $S$  pretext tasks with different input formats (i.e., image-only, text-only input, or image-text pair), the learning process can be formulated as

$$\theta^*, \theta_1^*, \dots, \theta_S^* = \arg \min_{\theta, \theta_1, \dots, \theta_S} \sum_{s=1}^S \mathcal{L}_s(Y_s, \mathcal{H}_{\theta_s}(\mathcal{M}_\theta(\mathbf{X}))), \quad (2)$$

where  $\mathcal{L}_s$  are the loss functions of pretext tasks,  $Y_s$  are the corresponding ground-truth labels, and  $\mathcal{H}_{\theta_s}$  are the prediction heads with their parameters  $\theta_s$ .

### 2.2 UNIFYING INPUTS USING PSEUDO TOKENS

To unify inputs, we design a basic solution for *compatibility* and an advanced solution for *scalability*. In this work, we use the advanced solution in default if not specified.

#### 2.2.1 COMPATIBILITY USING PSEUDO TOKENS

To make the backbone model compatible with variable-size and heterogeneous input, this work proposes a simple yet effective approach, namely using Pseudo Tokens (PT) as placeholders for missing modality.  $\mathcal{M}_\theta$  naturally accepts two inputs (visual and textual embeddings  $\mathbf{X}^v, \mathbf{X}^l$ ), which is by definition compatible to inputs with image-text pairs. For image-only/text-only inputs, we propose to introduce visual/textual pseudo tokens to enable the backbone model to perceive the missing input in a specific modality:

$$\mathbf{X} = \begin{cases} (\mathbf{X}^v, \mathbf{PT}^l) & \text{if } \textit{image-only} \\ (\mathbf{PT}^v, \mathbf{X}^l) & \text{if } \textit{text-only} \\ (\mathbf{X}^v, \mathbf{X}^l) & \text{if } \textit{image-text} \end{cases}, \quad (3)$$

where  $\mathbf{PT}^v \in \mathbb{R}^{D_v \times k}$  and  $\mathbf{PT}^l \in \mathbb{R}^{D_l \times k}$ .

#### 2.2.2 SCALABILITY OF PSEUDO TOKENS

The above solution adopts a static fashion to introduce pseudo tokens, which might have limited diversity and therefore harm its capacity. Hence, we construct a pool of visual/textual pseudo tokens *instead of static pseudo tokens*. Importantly, the selection of pseudo tokens is *conditioned on the input embeddings*.

**Pseudo Token Pool** Formally, we define a visual pseudo token pool  $\mathbf{V} \in \mathbb{R}^{D_v \times N_v}$  and a textual pseudo token pool  $\mathbf{T} \in \mathbb{R}^{D_l \times N_l}$ .  $N_v$  and  $N_l$  are the size of the visual/textual pseudo token pool, respectively.

**Pseudo Token Selection** Given the image-only input with its visual embedding sequence  $\mathbf{X}^v$  or language-only input with its textual embedding sequence  $\mathbf{X}^l$ , we conduct a pooling operation (e.g., average/max pooling) to obtain a *query vector* for existing modality (denoted as  $\mathbf{q}^v$  or  $\mathbf{q}^l$ ), namely,  $\mathbf{q}^v = \text{pooling}(\mathbf{X}^v)$  and  $\mathbf{q}^l = \text{pooling}(\mathbf{X}^l)$ , respectively. To get the pseudo tokens of the missing modality, the selection of pseudo tokens is based on the similarity scores between the query vector and all pseudo tokens in the pool from the missing modality.

$$\begin{aligned} PT^l &= \text{top-}k \left[ \mathbf{w}^T \mathbf{q}^v \right]_{\mathbf{w} \in V}, \\ PT^v &= \text{top-}k \left[ \mathbf{w}^T \mathbf{q}^l \right]_{\mathbf{w} \in T}, \end{aligned} \quad (4)$$

where  $\mathbf{w}$  is an embedding vector in the pseudo token pool, and we select  $k$  closest pseudo tokens as the input embeddings of the missing modality.

Without loss of any generality, we take a text-only scenario as an example, but it also holds for the image-only scenario. To select the best visual pseudo tokens for the text-only input, the proposed method chooses the most similar ones compared to the given textual query vector. As an **intuitive explanation**, one could treat the visual pseudo token pool as a feature bank that stores the most representative images of a given dataset. Eq. 4 aims to choose the visual pseudo tokens that might convey a similar semantic meaning as the given text by conducting dot products. In other words, *it might, at least to some extent, automatically fill (originally unprovided) semantically-similar images conditioned on purely the given text.*

**Linking to Prompts** We find that the PTUnifier (especially the static one in § 2.2.1) is quite similar to the prompt tuning (PT) (Li & Liang, 2021; Liu et al., 2021c). They both introduce special tokens or vectors as a certain signal for training or inference. One notable difference is that in a special version of PTUnifier using pseudo token pools (see § 2.2.2), the selection of additional tokens/vectors is conditioned on the input, while prompts are generally static and constant to input.

### 2.3 UNIFYING MULTIPLE PRE-TRAINING OBJECTIVES

Owing to the unified image and/or text input formulation, we can adopt pretext tasks of both fusion-encoders and dual-encoders (see Eq. 2). Following previous studies (Li et al., 2019; Tan & Bansal, 2019; Zhang et al., 2020; Radford et al., 2021), we develop two commonly used pre-text tasks (i.e., masked language modeling (MLM) and image-text matching (ITM)) for fusion-encoders and the image-text contrast (ITC) pre-text task for dual-encoders. To produce the prediction for the aforementioned MLM and ITM tasks, we use two independent prediction heads  $\mathcal{H}_{\text{MLM}}$  and  $\mathcal{H}_{\text{ITM}}$  (i.e., two two-layer multilayer perceptrons (MLPs)).

**Masked Language Modeling (MLM)** Following BERT (Devlin et al., 2019), we randomly mask 15% of the words (denoted as  $Y_{\text{MLM}}$ ) of the input text  $T$  and recover them according to the remaining text ( $T_M$ ) and the input  $I$ . The MLM objective is given by:

$$\mathcal{L}_{\text{MLM}} = - \sum_{(I, T)} \log p_{\text{MLM}}(Y_{\text{MLM}} | I, T_M), \quad (5)$$

where  $p_{\text{MLM}}$  is obtained by applying  $\mathcal{H}_{\text{MLM}}$  followed by a softmax operation on the corresponding representations of [MASK] in  $\mathcal{Z}^l$ .

**Image-Text Matching (ITM)** This task aims to distinguish whether an image-text pair is a match. In detail, a positive image-text pair and a randomly sampled negative pair are fed into  $\mathcal{M}_\theta$  and the concatenation of  $\mathbf{z}_{[\text{CLS}]}^v$  and  $\mathbf{z}_{[\text{CLS}]}^l$  is processed by  $\mathcal{H}_{\text{MLM}}$  followed by a softmax layer to output a binary probability  $p_{\text{ITM}}$ . Therefore, the ITM objective is given by

$$\mathcal{L}_{\text{ITM}} = - \sum_{(I, T)} \log p_{\text{ITM}}(Y_{\text{ITM}} | I, T). \quad (6)$$

**Image-Text Contrast (ITC)** This task aims to learn better uni-modal/cross-modal representation from the instance-level contrast. In this work, given an image-text pair, we use two different forward procedures on the image-only input  $I$  and the text-only input  $T$ , respectively, to obtain the image-only

representation (denoted as  $z^v$ ) and text-only representation (denoted as  $z^l$ ). Afterward, we adopt the similarity function  $s(I, T) = z^v \top z^l$  to compute the image-to-text similarity and text-to-image similarity between  $z^v$  and  $z^l$ . Subsequently, the similarities are normalized as follows:

$$p_n^{\text{i2t}} = \frac{\exp(s(I, T_n) / \tau)}{\sum_{n=1}^N \exp(s(I, T_n) / \tau)}, \quad (7)$$

$$p_n^{\text{t2i}} = \frac{\exp(s(I_n, T) / \tau)}{\sum_{n=1}^N \exp(s(I_n, T) / \tau)}, \quad (8)$$

where  $N$  is the size of the mini-batch. The ground-truth labels  $Y^{\text{i2t}}$  and  $Y^{\text{t2i}}$  are two  $N \times N$  one-hot matrices, where negative pairs have a probability of 0 and the positive pair has a probability of 1. Therefore, the ITC objective is given by

$$\mathcal{L}_{ITC} = -\frac{1}{2} \sum_{(I, T)} \log p^{\text{i2t}}(Y^{\text{i2t}} | I, T) - \frac{1}{2} \sum_{(I, T)} \log p^{\text{t2i}}(Y^{\text{t2i}} | I, T). \quad (9)$$

### 3 OVERALL ARCHITECTURE

The previous section documents the unification at the input and task levels. This section will introduce the overall architecture of our work. As a pipeline, we first map visual and textual tokens into embeddings space ( $\mathbf{X}^v$  and  $\mathbf{X}^l$  as specified in §3.1). Such token embeddings with or without pseudo tokens will be jointly processed by an identical backbone model  $\mathcal{M}_\theta$  (§3.2). An overview of the proposed approach is shown in Figure 1(b). The training objectives are introduced in (§2.3).

#### 3.1 VISUAL AND TEXTUAL EMBEDDINGS

**Visual embedding** For an input image  $I$ , it is first segmented into patches following Dosovitskiy et al. (2021). Then the patches are linearly projected into patch embeddings  $\mathbf{X}^v = (\mathbf{x}_1^v, \mathbf{x}_2^v, \dots, \mathbf{x}_{N_v}^v)$ ,  $\mathbf{x}_i^v \in \mathbb{R}^{D_v}$  through a linear transformation and a special learnable token embedding  $\mathbf{x}_{[\text{CLS}]}^v$  is prepended for the aggregation of visual information. Therefore, the image embedding sequence is obtained by summing up the patch embeddings and learnable 1D position embeddings  $\mathbf{E}_{pos}^v \in \mathbb{R}^{D_v \times (N_v+1)}$ :

$$\mathbf{X}^v = [\mathbf{x}_{[\text{CLS}]}^v; \mathbf{x}_1^v; \mathbf{x}_2^v; \dots; \mathbf{x}_{N_v}^v] + \mathbf{E}_{pos}^v, \quad (10)$$

where  $[\cdot; \cdot]$  represents the column concatenation.<sup>5</sup>

**Textual embedding** Similarly, for an input text  $T$ , we follow BERT (Devlin et al., 2019) to tokenize the input text to subword tokens by WordPiece (Wu et al., 2016). Afterwards, the tokens are linearly projected into embeddings  $\mathbf{X}^l = (\mathbf{x}_1^l, \mathbf{x}_2^l, \dots, \mathbf{x}_{N_l}^l)$ ,  $\mathbf{x}_i^l \in \mathbb{R}^D$  through a linear transformation with a start-of-sequence token embedding  $\mathbf{x}_{[\text{CLS}]}^l$ , and a special boundary token embedding  $\mathbf{x}_{[\text{SEP}]}^l$  added. Therefore, the text embedding sequence is obtained by summing up the sub-word token embeddings and text position embeddings  $\mathbf{E}_{pos}^l \in \mathbb{R}^{D \times (N_l+2)}$ :

$$\mathbf{X}^l = [\mathbf{x}_{[\text{CLS}]}^l; \mathbf{x}_1^l; \dots; \mathbf{x}_{N_l}^l; \mathbf{x}_{[\text{SEP}]}^l] + \mathbf{E}_{pos}^l. \quad (11)$$

#### 3.2 BACKBONE MODEL ARCHITECTURE

Since the input image and/or text are represented as a unified image-text sequence, the backbone model can be any model for sequential modeling. In this work, we adopt an attention-based Med-VLP model with the multi-modal interaction, which can be an effective model (including uni-modal encoders and a multi-modal fusion module) or an efficient one (i.e., a single Transformer model). Formally, for a given input (defined in Eq. 3), the whole representation process can be formulated as

$$\mathbf{Z}^v, \mathbf{Z}^l = \mathcal{M}_\theta(\mathbf{X}), \quad (12)$$

where  $\mathbf{Z}^v = (z_{[\text{CLS}]}^v, z_1^v, z_2^v, \dots, z_{N_v}^v)$  and  $\mathbf{Z}^l = (z_{[\text{CLS}]}^l, z_1^l, \dots, z_{N_l}^l, z_{[\text{SEP}]}^l)$  are the image and text representations from the backbone model.

<sup>5</sup>We abuse the notation  $\mathbf{X}^v$  for simplicity (same for  $\mathbf{X}^l$ ).

## 4 EXPERIMENTAL SETTINGS

### 4.1 PRE-TRAINING DATASETS

In our experiments, we perform the pre-training on three datasets, which are described as follows:

- **ROCO** (Pelka et al., 2018): a dataset of radiology figure-caption pairs from PubMed Central, an open-access biomedical literature database.
- **MedICaT** (Subramanian et al., 2020): a dataset of medical figure-caption pairs also extracted from PubMed Central. Different from ROCO, 75% of its figures are compound figures, including several sub-figures.
- **MIMIC-CXR** (Johnson et al., 2019): the largest radiology dataset to date from the Beth Israel Deaconess Medical Center.

For all the datasets, we exclude those samples with the length of their texts less than 3. For ROCO and MedICaT, we filter non-radiology samples, and for MIMIC-CXR, we only keep images in the frontal view. As for the dataset split, we adopt the official splits of ROCO and MIMIC-CXR. For MedICaT, we randomly sample 1,000 image-text pairs for validation and 1,000 for testing, and the remaining image-text pairs are used for training.<sup>6</sup> Different from the texts in general-domain VLP, the medical texts are long narratives consisting of multiple sentences. To deal with this case, we randomly sample a sentence from the input text in each iteration.

### 4.2 MEDICAL VISION-LANGUAGE BENCHMARK

To evaluate the performance, we construct a medical vision-language evaluation benchmark including three types of tasks, i.e., uni-modal, cross-modal, and multi-modal evaluations.<sup>7</sup>

**Uni-modal Evaluation** requires the model to process a single modality with vision-only or language-only inputs. For vision-only tasks, we conduct the image classification (IC) experiments on CheXpert (Irvin et al., 2019) and RSNA Pneumonia (Shih et al., 2019). For language-only tasks, we perform both the understanding task (i.e., text classification (TC)) and the generation task (i.e., text summarization (TS)) on the RadNLI (Romanov & Shivade, 2018; Miura et al., 2021) and MIMIC-CXR datasets, respectively.

**Cross-modal Evaluation** requires the model to align the vision and language modalities. We conduct experiments on three kinds of tasks (i.e., image-to-text retrieval (ITR), text-to-image retrieval (TIR), and image-to-text generation (ITG)). For ITR and TIR, we adopt the ROCO dataset. For ITG, we conduct experiments on the MIMIC-CXR dataset to evaluate its ability for radiology report generation.

**Multi-modal Evaluation** requires the model to reason over both the image and text inputs through the multi-modal interaction. We conduct the experiments on the medical visual question answering (VQA) task, which requires the model to answer natural language questions about a medical image. We adopt three publicly available Med-VQA datasets (i.e., VQA-RAD (Lau et al., 2018), SLACK (Liu et al., 2021b) and MedVQA-2019 (Abacha et al., 2019)).

The fine-tuning strategies can be divided into three categories according to the type of tasks. Specifically, for the classification tasks (i.e., IC, TC, and VQA), we feed the concatenation of the image/visual pseudo token and text/textual pseudo token representations to a randomly initialized two-layer MLP to predict the labels. For the retrieval tasks (i.e., ITR and TIR), we adopt the prediction head for the image-text contrast pre-text task and test its zero-shot and fine-tuned performance. For the generation tasks (i.e., TS and ITG), we feed the concatenation of the sequence of image/visual pseudo token and text/textual pseudo token representations to a Transformer decoder with its parameters (except for the parameters of cross-attention layers) initialized from the pre-trained language encoder. For the evaluation metrics, we follow the previous studies to adopt AUROC for IC, accuracy for TC and VQA, Recall@K (K=1, 5, 10) for ITR and TIR, and natural language generation (NLG) metrics (i.e.,

<sup>6</sup>More details of the pre-training datasets are reported in Appendix A.

<sup>7</sup>More details of the downstream evaluations are reported in Appendix B.

Methods	Uni-Modal				Cross-Modal				Multi-Modal		
	Image		Text		Image-to-Text		Text-to-Image		VQA-RAD Acc	SLACK Acc	MedVQA-2019 Acc
	CheXpert AUROC	RSNA AUROC	RadNLI Acc	MIMIC RG-L	MIMIC BL-4	ROCO R@1	ROCO R@1	R@1			
SOTA <sub>1</sub>	87.3	81.3	72.6	43.8	8.0	11.9	9.8	72.7	82.1	-	
SOTA <sub>2</sub>	88.1	88.6	77.8	45.1	8.2	14.5	11.3	72.0	-	77.9	
<b>PTUnifier (ours)</b>	<b>90.1</b>	<b>90.6</b>	<b>80.0</b>	<b>46.2</b>	<b>10.7</b>	<b>21.0</b>	<b>20.8</b>	<b>78.3</b>	<b>85.2</b>	<b>79.3</b>	

Table 1: Comparisons of our proposed method with previous studies on three types of evaluations (*i.e.*, uni-modal, cross-modal, and multi-modal evaluations). SOTA<sub>1</sub> and SOTA<sub>2</sub> denote two state-of-the-art approaches of each type of tasks, respectively. BL-4 denotes BLEU score using 4-grams and RG-L denotes ROUGE-L. Dark and light grey colors highlight the top and second best results on each metric. Note that the results of text summarization and image-to-text generation are replicated using our pre-processed data (See Appendix B and F).

BLEU (Papineni et al., 2002), METEOR (Denkowski & Lavie, 2011), CIDEr (Vedantam et al., 2015) and ROUGE (Lin, 2004)) for TS and ITG.<sup>8</sup>

To demonstrate the effectiveness of the proposed approach, we compare it with previous studies, including ConVIRT (Zhang et al., 2020), GLoRIA (Huang et al., 2021), ClinicalBERT (Alsentzer et al., 2019), IFCC (Miura et al., 2021), TransABS (Liu & Lapata, 2019), WGSum (Hu et al., 2021), R2Gen (Chen et al., 2020c), R2GenCMN (Chen et al., 2021), ViLT (Kim et al., 2021), METER (Dou et al., 2021), CPRD (Liu et al., 2021a), and MMBERT (Khare et al., 2021).

## 5 RESULTS AND ANALYSES

### 5.1 MAIN RESULTS

As observed in Table 1, our approach achieves the best performance on all tasks.<sup>9</sup> It outperforms previous studies on uni-modal image classification (+2.0% AUROC), text classification (+2.2% Accuracy), text summarization (+1.1% Rouge-L), image-to-text generation (+2.4% BLEU-4), image-to-text retrieval (+7.5% Recall@1), text-to-image retrieval (+9.5% Recall@1), and multi-modal VQA (+3.4% Accuracy), which confirms the validity of the proposed approach. Furthermore, the proposed approach outperforms those complicated methods designed for specific tasks (*e.g.*, WGSum using extra word graphs and CPRD adopting representation distillation). Note that the existing studies are only designed for a single task, while our approach generally targets all vision- and/or language-related tasks, namely, without any tailored adaptations to a specific task.

### 5.2 EFFECTS OF DIFFERENT OBJECTIVES

To further illustrate the effectiveness of our proposed approach, we perform an ablation study on the pre-training objectives, including the ones from fusion-encoders (*i.e.*, MLM and ITM) and the one from dual-encoders (*i.e.*, ITC).

There are several observations drawn from different aspects. First, the fusion-encoders’ objectives (*i.e.*, MLM and ITM) guide the models (*i.e.*, ID 3 and 5) to learn the transferrable multi-modal representations, which achieve the promising performance on the downstream VQA task. Second, the dual-encoders’ objective (*i.e.*, ITC) assists the models (*i.e.*, ID 4 and 5) in learning the uni-modal image representations and the cross-modal representations, and the models pre-trained with the ITC objective outperform those pre-trained without the ITC objective. Third, interestingly, **the ITC objective does not promote the performance of the uni-modal text classification task**. The reason behind this might be that images and texts are abstracted at different levels, where pixels of images have a lower semantic level than tokens of texts. Therefore, in the ITC process, the texts can be treated as a kind of “supervision signals” for the learning of image encoding, yet, it is harder for the images to play such a role in contrast. This can be observed from previous studies, where

<sup>8</sup>The hyperparameter settings are reported in Appendix C.

<sup>9</sup>We report the validation results in Appendix D, the statistics of results in Appendix E, and the detailed NLG metrics in Appendix F.

ID	Objectives			Uni-Modal Image			Text	Cross-Modal Image-to-Text		Multi-Modal		
	MLM	ITM	ITC	1%	CheXpert	100%	RadNLI	MIMIC		VQA-RAD		Overall Acc
				AUROC	AUROC	AUROC	Acc	BL-4	CDr	Open Acc	Closed Acc	
1	✓			66.1	79.1	81.1	77.2	9.1	15.0	57.5	79.5	70.8
2		✓		56.9	83.0	85.8	77.5	10.0	18.2	23.5	82.8	59.3
3	✓	✓		74.5	87.2	88.4	78.3	9.9	17.1	67.0	84.6	77.7
4			✓	88.0	88.9	89.3	76.5	10.3	19.0	64.8	81.0	74.6
5	✓	✓	✓	88.7	89.0	90.1	80.0	10.7	21.0	68.7	84.6	78.3

Table 2: Ablation studies on the different types of objectives, including the fusion-encoders ones (*i.e.*, masked language modeling (MLM) and image-text matching (ITM)) and the dual-encoders one (*i.e.*, image-text contrastive (ITC)). 1%, 10%, and 100% represent the different portion of training data of CheXpert.

Pool Size	Pool Para.	MLM	ITM	ITC	Total
0	0	1.055	0.232	1.901	3.188
512	393K	1.053	0.215	1.787	3.055
1024	784K	1.049	0.211	1.778	3.038
2048	1,573K	1.057	0.229	1.861	3.147

Table 3: Pre-training losses (including MLM, ITM, and ITC) of our approach against different pool size, where the parameters of the pool (Pool Para.) are also shown.

the dual-encoders were only evaluated on the uni-modal vision tasks or cross-modal tasks. Fourth, the models pre-trained with the ITC objective (*i.e.*, ID 4 and 5) demonstrate their great transfer ability where the pre-trained models can achieve high performance with very little data (e.g., 1% and 10%). Fifth, performing both types of objectives promotes the model (*i.e.*, ID 5) to achieve the best performance across all the tasks, which confirms the feasibility of the research direction on unifying the fusion-encoders and dual-encoders.

### 5.3 EFFECTS OF PSEUDO TOKEN POOLS

To analyze the impacts of pseudo token pools, we perform the pre-training with different pool sizes (ranging from 0 to 2048) with the results shown in Table 3.<sup>10</sup> We have several observations: (i) Although enlarging pool size leads to increasing parameter numbers, it is demonstrated that there are not too many parameters (less than 0.5%) introduced compared with the total parameters (350M); (ii) All models with pseudo token pools have a better convergence (with a lower convergence loss) than the one without pseudo token pools (*i.e.*, pool size equal to 0), which demonstrates the effectiveness of introducing the pseudo token pools; (iii) It is found that setting a proper pool size is important, where the model achieves the best convergence when the pool size is set to 1024. This might owe to the fact that the pool size controls how much the modal information is stored during the pre-training procedure, and a large pool size with a large capacity might “absorb” too much noise in the pre-training corpus.<sup>11</sup>

## 6 RELATED WORK

**Vision-and-Language Pre-training (VLP)** Motivated by the success of the self-supervised pre-training recipe in natural language processing (NLP) (e.g., BERT (Devlin et al., 2019)) and computer vision (CV) (e.g., SimCLR (Chen et al., 2020a) and MoCo (He et al., 2020)), there has been an increasing interest in developing VLP methods to address a wide range of vision-and-language-related tasks. In general, VLP methods can be classified into two categories according to the vision-and-language interaction, *i.e.*, dual-encoders and fusion-encoders. Existing dual-encoder methods can

<sup>10</sup>We show the pre-training losses since they directly reflect how well the models perform the pre-text tasks.

<sup>11</sup>We show cases to illustrate the potential functions of the learned pseudo tokens in Appendix G.



be summarized according to the following aspects: (i) using medium-scale curated image-text data (Radford et al., 2021), (ii) using large-scale noisy image-text data (Jia et al., 2021), (iii) designing more fine-grained image-text contrast (Yao et al., 2021), (iv) adopting extra single modal contrastive learning (Mu et al., 2021). For fusion-encoders, existing studies can be further categorized with respect to these three perspectives: (i) Uni-modal encoders: different methods adopt different image features (e.g., region features (Li et al., 2019; Lu et al., 2019), patch embeddings (Kim et al., 2021), and grid features (Huang et al., 2020)) and distinct text features (e.g., statistic embeddings (Kim et al., 2021) and dynamic embeddings (Dou et al., 2021)); (ii) Multi-modal fusion modules: existing studies adopted the single-stream fusion scheme (Su et al., 2020; Li et al., 2020a) or dual-stream fusion scheme (Tan & Bansal, 2019; Yu et al., 2021); (iii) Pretext tasks: existing studies explore a variety of pre-training tasks, including masked language modeling (Li et al., 2019), masked image modeling (Lu et al., 2019; Chen et al., 2020b), image-text matching (Zhang et al., 2021). This paper adopts the model architecture of fusion-encoders and the pre-text tasks from both dual and fusion encoders.

**Medical Vision-and-Language Pre-Training (Med-VLP)** Being one of the applications and extensions of VLP to the medical domain, Med-VLP aims to understand the content of medical images and texts, which can be traced back to Zhang et al. (2020) for dual-encoders and Li et al. (2020b) for fusion-encoders. For dual-encoders, the follow-up studies (Huang et al., 2021; Müller et al., 2021) explored the global-local image-text contrastive learning to capture more fine-grained information among medical images and texts and have achieved state-of-the-art results in the medical image classification task. For fusion-encoders, Khare et al. (2021); Moon et al. (2021) performed pre-training to improve the multi-modal reasoning ability of the vision-and-language models for the downstream task (i.e., Medical VQA). Compared with these studies, we design a more comprehensive scheme for Med-VLP from four aspects (i.e., pre-training datasets, model designs, pre-training tasks, and evaluation benchmarks).

**Unified Vision-and-Language Pre-training** To unify the dual and fusion encoders, existing studies mainly adopted/designed specific model architectures to accommodate different pre-text tasks. The most common scheme is to add an extra multi-modal fusion module to the dual-encoders and perform the cross-modal pre-text task (i.e., image-text contrast) before the fusion and multi-modal pre-text tasks (e.g., MLM and ITM) after the fusion (Li et al., 2021; Singh et al., 2022). Besides, Wang et al. (2021) proposed a mixture-of-modality experts (MoME) Transformer to unify vision-and-language models by employing a set of modality experts to replace the feed-forward networks (FFN) in the standard Transformer. However, the aforementioned studies are architecture-dependent, and they perform the unifying through training different parts of the models when applying different types of VLP objectives. Therefore, it is expected to unify the existing Med-VLP paradigms in an *architecture- and task-agnostic* fashion to improve the generalization and extensionality ability of Med-VLP methods, as done in this paper.

## 7 CONCLUSION

In this paper, we proposed a simple yet effective Med-VLP scheme to take the advantages of both fusion-encoders and dual-encoders, where visual and textual pseudo token pools are used to make our model compatible with different kinds of inputs (i.e., image-only, text-only, and image-text-pair), and thus different types of objectives (e.g., MLM and ITM for fusion-encoders and ITC for dual-encoders) can be adopted for pre-training. It is worth noting that our proposed approach is complementary to most of the existing Med-VLP models. To perform a comprehensive evaluation, we construct a medical vision-language evaluation benchmark including three types of tasks. Experimental results confirm the validity of our approach, which achieves state-of-the-art performance on all the downstream tasks. Furthermore, the analyses investigate the effects of different types of objectives and different pool sizes, and such empirical studies might provide a valuable reference for future research, even for VLP in the general domain.

**Limitation** The proposed vision-and-language approach is orthogonal to the domains (e.g., the general domain, and the medical domain). However, limited by GPU resources, we do not perform the pre-training in the general domain. Instead, we simulate a similar experimental environment in the medical domain, which allows us pre-training the models with an academic budget. Nonetheless, we admit that it would be better to evaluate domain-agnostic approaches in the general domain to verify their generalization.

## REPRODUCIBILITY STATEMENT

For the data, all the pre-training and downstream datasets are publicly available. We report the details of the pre-training datasets in Appendix A and the details of the downstream evaluations in Appendix B. For the approach, we describe the model architecture and pre-training objectives in Section 2 and 3. For the hyperparameters, the details are reported in Appendix C. For experimental results, we report the validation results in Appendix D, the statistics of results in Appendix E, and the detailed NLG metrics in Appendix F. Besides, our code will be released in the final version of this paper.

## REFERENCES

- Asma Ben Abacha, Sadid A. Hasan, Vivek Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF*, 2019.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 72–78, Minneapolis, Minnesota, USA, 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909. URL <https://aclanthology.org/W19-1909>.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a. URL <http://proceedings.mlr.press/v119/chen20j.html>.
- 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*, pp. 104–120. Springer, 2020b.
- Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1439–1449, Online, 2020c. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112. URL <https://aclanthology.org/2020.emnlp-main.112>.
- Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5904–5914, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.459. URL <https://aclanthology.org/2021.acl-long.459>.
- 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*, pp. 85–91, Edinburgh, Scotland, 2011. Association for Computational Linguistics. URL <https://aclanthology.org/W11-2107>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL <https://aclanthology.org/N19-1423>.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL <https://openreview.net/forum?id=YicbFdNTTy>.

- Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Zicheng Liu, Michael Zeng, et al. An empirical study of training end-to-end vision-and-language transformers. *ArXiv preprint*, abs/2111.02387, 2021. URL <https://arxiv.org/abs/2111.02387>.
- Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9726–9735. IEEE, 2020. doi: 10.1109/CVPR42600.2020.00975. URL <https://doi.org/10.1109/CVPR42600.2020.00975>.
- Jinpeng Hu, Jianling Li, Zhihong Chen, Yaling Shen, Yan Song, Xiang Wan, and Tsung-Hui Chang. Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4980–4990, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.441. URL <https://aclanthology.org/2021.findings-acl.441>.
- 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*, pp. 3942–3951, 2021.
- Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *ArXiv preprint*, abs/2004.00849, 2020. URL <https://arxiv.org/abs/2004.00849>.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 590–597. AAAI Press, 2019. doi: 10.1609/aaai.v33i01.3301590. URL <https://doi.org/10.1609/aaai.v33i01.3301590>.
- Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4904–4916. PMLR, 2021. URL <http://proceedings.mlr.press/v139/jia21b.html>.
- 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*, abs/1901.07042, 2019. URL <https://arxiv.org/abs/1901.07042>.
- Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1033–1036. IEEE, 2021.
- Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5583–5594. PMLR, 2021. URL <http://proceedings.mlr.press/v139/kim21k.html>.
- 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.
- 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, 2021.

- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *ArXiv preprint*, abs/1908.03557, 2019. URL <https://arxiv.org/abs/1908.03557>.
- Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.
- Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pp. 121–137. Springer, 2020a.
- Yikuan Li, Hanyin Wang, and Yuan Luo. A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1999–2004. IEEE, 2020b.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. URL <https://aclanthology.org/W04-1013>.
- Bo Liu, Li-Ming Zhan, and Xiao-Ming Wu. Contrastive pre-training and representation distillation for medical visual question answering based on radiology images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 210–220. Springer, 2021a.
- 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)*, pp. 1650–1654. IEEE, 2021b.
- Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021c.
- Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3730–3740, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1387. URL <https://aclanthology.org/D19-1387>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692, 2019. URL <https://arxiv.org/abs/1907.11692>.
- Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL <https://openreview.net/forum?id=Bkg6RiCqY7>.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic vi-  
siolinguistic representations for vision-and-language tasks. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d’Alché-Buc, Emily B. Fox, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pp. 13–23, 2019. URL <https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html>.
- Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory F. Diamos, Erich Elsen, David García, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. Mixed precision training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL <https://openreview.net/forum?id=r1gs9JgRZ>.
- Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304, Online, 2021. Association for Computational

- Linguistics. doi: 10.18653/v1/2021.naacl-main.416. URL <https://aclanthology.org/2021.naacl-main.416>.
- Jong Hak Moon, Hyungyung Lee, Woncheol Shin, and Edward Choi. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *ArXiv preprint*, abs/2105.11333, 2021. URL <https://arxiv.org/abs/2105.11333>.
- Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *ArXiv preprint*, abs/2112.12750, 2021. URL <https://arxiv.org/abs/2112.12750>.
- Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rückert. Joint learning of localized representations from medical images and reports. *ArXiv preprint*, abs/2112.02889, 2021. URL <https://arxiv.org/abs/2112.02889>.
- 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*, pp. 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL <https://aclanthology.org/P02-1040>.
- 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*, pp. 180–189. Springer, 2018.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL <http://proceedings.mlr.press/v139/radford21a.html>.
- Alexey Romanov and Chaitanya Shivade. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1586–1596, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1187. URL <https://aclanthology.org/D18-1187>.
- 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), 2019.
- Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL <https://openreview.net/forum?id=SygXPaEYvH>.
- Sanjay Subramanian, Lucy Lu Wang, Ben Bogin, Sachin Mehta, Madeleine van Zuylen, Sravanthi Parasa, Sameer Singh, Matt Gardner, and Hannaneh Hajishirzi. MedICaT: A dataset of medical images, captions, and textual references. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 2112–2120, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.191. URL <https://aclanthology.org/2020.findings-emnlp.191>.
- Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language*

- Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5100–5111, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1514. URL <https://aclanthology.org/D19-1514>.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4566–4575. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7299087. URL <https://doi.org/10.1109/CVPR.2015.7299087>.
- Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *ArXiv preprint*, abs/2111.02358, 2021. URL <https://arxiv.org/abs/2111.02358>.
- Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google’s neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144, 2016. URL <https://arxiv.org/abs/1609.08144>.
- Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ArXiv preprint*, abs/2111.07783, 2021. URL <https://arxiv.org/abs/2111.07783>.
- Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 3208–3216, 2021.
- Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5579–5588, 2021.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *ArXiv preprint*, abs/2010.00747, 2020. URL <https://arxiv.org/abs/2010.00747>.

## A MORE DETAILS OF PRE-TRAINING DATASETS

Table 4 shows the statistics of the pre-training datasets.

Datasets	Image #	Text #	Avg. Len.	Avg. Sent. #
ROCO	81k	81k	20.42	1.46
MedICaT	124k	321k	40.88	2.82
MIMIC-CXR	232k	367k	36.49	5.07

Table 4: The statistics of the three pre-training datasets including the numbers of images, texts, the average word-based length (Avg. Len.) of texts, and the average number of sentences (Avg. Sent. #) of texts.

## B MORE DETAILS OF DOWNSTREAM EVALUATION

In this appendix, we detail the descriptions for each downstream evaluation dataset.

**CheXpert** This dataset contains 224,316 chest radiographs labeled for 14 medical observations. Following the previous studies (GLORIA), we only keep those front-view radiographs, hold out the expert-labeled validation set as the test set, and randomly sample 5,000 images from the training data for validation.

**RNAS Pneumonia** This dataset consists of 30,000 front-view chest radiographs labeled by “pneumothorax negative” or “pneumothorax positive”. Following the previous studies (GLORIA), the train/validation/test split constitutes 70%/15%/15% of the dataset, respectively.

**RadNLI** This dataset contains 19k sentence pairs labeled by “Entailment”, “Neutral”, or “Contradiction”. We follow IFCC to produce and pre-process the dataset, which contains the training data from an extra NLI dataset (i.e., MedNLI).

**ROCO** This dataset contains 81k image-text pairs. For the training and validation set, we adopt the official ones. For the test procedure, we sample 2,000 pairs from the test set and evaluate the models on the 2,000 pairs to obtain the Recall@K scores.

**MIMIC-CXR** This dataset contains 377,110 chest x-rays. Different from the pre-training, for downstream evaluations (i.e., text summarization and image-to-text generation), we only keep those front-view x-rays with both the findings and impression section.

**VQA-RAD** This dataset consists of 315 images and 3,515 questions. We adopt the commonly used version pre-processed by MEVF.

**SLACK** This dataset contains 642 images and 14,028 questions. We follow the original SLACK paper to prepare and pre-process the dataset and adopt the official dataset split.

**MedVQA-2019** This dataset contains 4,200 images and 15,292 questions. We follow previous studies to prepare and pre-process the dataset by keeping the main three categories of questions: Modality, Plane, and Organ system.

## C MORE DETAILS OF HYPER-PARAMETER SETTINGS

Table 5 reports the hyper-parameters adopted for pre-training and fine-tuning.

**Pre-training** We adopt the classical VLP model as the backbone model, including a vision encoder, a language encoder, and a multi-modal fusion module. For the vision and language encoders, we adopt base-size Transformer encoders with 12 layers initialized from CLIP-ViT-B (Radford et al., 2021) RoBERTa-base (Liu et al., 2019) and their hidden dimension is set to 768. For the multi-modal fusion module, we set the number of Transformer layers to 6, the dimension of the hidden states to 768, and the number of heads to 12. For the visual/textual pseudo token pools, the dimension and the pool size are set to 768 and 1,024, respectively, by default. For optimization, the pre-training takes

100,000 steps with AdamW optimizer (Loshchilov & Hutter, 2019) with a weight decay of 0.01. The learning rates for the vision and language encoders and the remaining parameters are set to  $1e-5$  and  $5e-5$ , respectively. We use the warm-up strategy during the first 10% of the total number of steps, and the learning rate is linearly decayed to 0 after warm-up. For data augmentation, we use center-crop to resize each image to the size of  $288 \times 288$ .

**Fine-tuning** For all downstream tasks, we use the AdamW optimizer with the learning rate set to  $5e-6$  and  $2.5e-4$  for the backbone model and task-specific layers, respectively.

All pre-training and fine-tuning experiments are conducted on 80GB NVIDIA A100 GPUs with mixed-precision (Mickevičius et al., 2018) to accelerate training and save memory.

Hyper-Parameters	Pre-training	Fine-tuning
Optimizer	AdamW	AdamW
Learning rate (backbone model)	$1e-5$	$5e-5$
Learning rate (prediction heads)	$5e-6$	$2.5e-4$
Weight decay	0.01	0.01
Optimizer momentum	(0.9, 0.98)	(0.9, 0.98)
Batch size	256	{16, 32}
Learning rate schedule	cosine	cosine
warmup ratio	0.01	0.01
Training steps	100,000	-
Training epochs	-	{15, 20, 30, 50}

Table 5: The hyper-parameters used for pre-training and fine-tuning.

## D RESULTS ON THE VALIDATION SET

Table 6 reports the scores on the validation set of different evaluation datasets.

	Uni-Modal		Text		Cross-Modal	Multi-Modal		
	Image				Image-to-Text			
	CheXpert AUROC	PNAS AUROC	RadNLI Acc	MIMIC RG-L	MIMIC BL-4	VQA-RAD Acc	SLACK Acc	MedVQA-2019 Acc
Validation	81.8	90.0	74.2	58.3	12.1	78.3	86.8	83.5

Table 6: Supplementary results of the proposed approach, where the scores on the validation set are shown.

## E MEAN AND STANDARD DEVIATION

We run each experiment three times with different random seeds and report the summarized statistics with the mean and standard deviation in Table 7.

	Uni-Modal		Text		Cross-Modal	Multi-Modal		
	Image				Image-to-Text			
	CheXpert AUROC	PNAS AUROC	RadNLI Acc	MIMIC RG-L	MIMIC BL-4	VQA-RAD Acc	SLACK Acc	MedVQA-2019 Acc
Mean (Test)	89.8	90.6	79.9	46.0	10.6	78.1	85.0	78.9
Std (Test)	0.35	0.02	0.24	0.19	0.13	0.22	0.25	0.41

Table 7: Supplementary results of the proposed approach, where the score statistics on the test set are shown.



## F DETAILED NLG METRICS FOR THE GENERATION TASKS

We report the detailed NLG Scores for the text summarization and image-text generation tasks in Table 8 and 9, respectively.

Methods	RG-1	RG-2	RG-L
TransABS	46.2	29.1	43.9
WGSUM	47.0	30.6	45.1
Ours	50.1	34.7	46.2

Table 8: Detailed NLG scores of the existing studies and the proposed approach on the text summarization task.

Methods	BL-1	BL-2	BL-3	BL-4	MTR	RG-L	CDr
R2Gen	28.6	17.1	11.4	8.0	12.8	23.7	13.6
R2GenCMN	29.9	17.8	11.7	8.2	13.0	23.1	14.3
Ours	35.8	22.1	14.9	10.7	15.3	25.8	21.0

Table 9: Detailed NLG scores of the existing studies and the proposed approach on the image-text generation task.

## G MORE ILLUSTRATIONS OF PSEUDO TOKENS

To illustrate the inherent mechanism of the proposed approach more clearly, we show those images that share the same pseudo token in Figure 2. It can be observed that after the pre-training, the pseudo tokens represent some topics of the images/texts implicitly. For example, those images sharing the 1947th pseudo textual token are in the potential topics “*Atelectasis*” and “*Cardiomegaly*”; those images sharing the 816th pseudo textual token are in the potential topic “*Edema*”; those images sharing the 601st pseudo textual token are in the potential topic “*Support Devices*”. This demonstrates that the semantically-similar information is preserved in the pseudo tokens during the pre-training procedure.

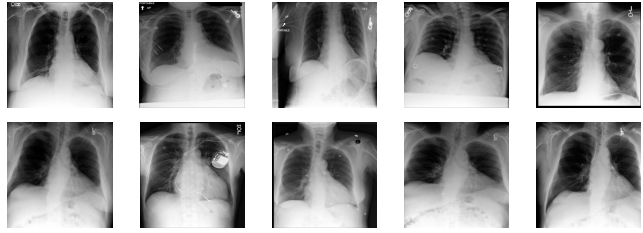
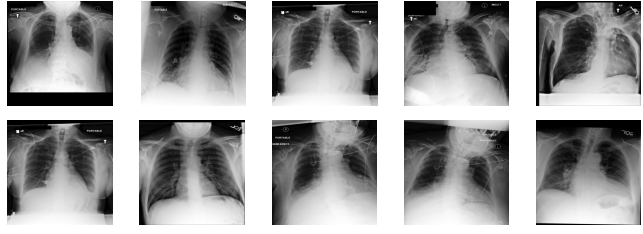
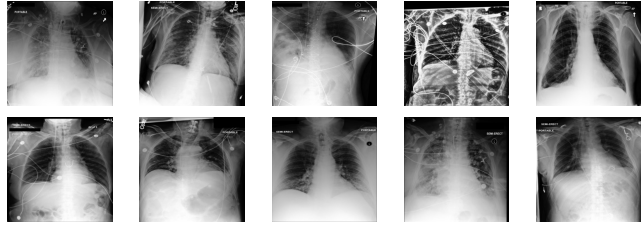
Pseudo Token ids	Potential Topics	Selected Images that Share the Same Pseudo Token
1947	<i>"Atelectasis", "Cardiomegaly"</i>	
816	<i>"Edema"</i>	
601	<i>"Support Devices"</i>	

Figure 2: Illustrations of pseudo tokens, where images shared the same pseudo token are shown with their potential topics.