Multi-modal Pre-training for Medical Vision-language Understanding and Generation: An Empirical Study with A New Benchmark

Li Xu LI-CONTROL.XU@CONNECT.POLYU.HK
Bo Liu BOKELVIN.LIU@CONNECT.POLYU.HK
Ameer Hamza Khan AMEER-HAMZ.KHAN@POLYU.EDU.HK
Lu Fan CSLFAN@COMP.POLYU.EDU.HK
Xiao-Ming Wu✉ XIAO-MING.WU@POLYU.EDU.HK
Department of Computing, The Hong Kong Polytechnic University

Abstract

With the availability of large-scale, comprehensive, and general-purpose vision-language (VL) datasets such as MSCOCO, vision-language pre-training (VLP) has become an active area of research and proven to be effective for various VL tasks such as visual-question answering. However, studies on VLP in the medical domain have so far been scanty. To provide a comprehensive perspective on VLP for medical VL tasks, we conduct a thorough experimental analysis to study key factors that may affect the performance of VLP with a unified vision-language Transformer. To allow making sound and quick pre-training decisions, we propose RadioGraphy Captions (RGC), a high-quality, multi-modality radiographic dataset containing 18,434 image-caption pairs collected from an open-access online database MedPix. RGC can be used as a pre-training dataset or a new benchmark for medical report generation and medical image-text retrieval. By utilizing RGC and other available datasets for pre-training, we develop several key insights that can guide future medical VLP research and new strong baselines for various medical VL tasks.

Data and Code Availability In this study, we conduct experiments with 6 public datasets: ROCO (Pelka et al., 2018), MedICaT (Subramanian et al., 2020), MIMI-CXR (Johnson et al., 2019), SLAKE (Liu et al., 2021b), VQA-RAD (Lau et al., 2018) and IU X-Ray (Demner-Fushman et al., 2016). We also collect and filter image-caption pairs from MedPix[1] – an online open-access database and construct the proposed RGC dataset, which is hosted on NIH website[2] under the MedPix license. The source code and pre-trained models for reproducing the reported results are available at this link[3].

Institutional Review Board (IRB) This study has no human-subject research and only uses publicly available and de-identified data, which does not need an IRB approval.

1. Introduction

Background and Motivation. Medical vision-language (Med-VL) tasks include visual question answering (Med-VQA) (Liu et al., 2021a; Nguyen et al., 2019), medical report generation (Chen et al., 2020b), and medical image-text retrieval (Zhang et al., 2020), as illustrated in Figure 1. Due to their great potential in computer assisted diagnosis and healthcare automation, Med-VL tasks have recently attracted increasing attention from the academia. Solving VL tasks requires cross-modal understanding and generation, and vision-language pre-training (VLP) has shown great promise in various VL tasks. However, unlike the general domain, where the study of VLP has rapidly advanced (Zhou et al., 2020; Li et al., 2019b; Tan and Bansal, 2019; Li et al., 2020; Kim et al., 2021; Su et al., 2020; Lu et al., 2019; Li et al., 2021b,d; Gan et al., 2020; Li et al., 2021a; Dou et al., 2022; Chen et al., 2020a; Huang et al., 2021b), the research of medical VLP has so far been scanty.

Related Works. The main reason may be the lack of publicly available, large-scale, and high-

1. https://medpix.nlm.nih.gov/

2. https://openi.nlm.nih.gov/imgs/collections/RGC.zip

3. https://github.com/Control-xl/Medical-Vision-Langauge-Transformer

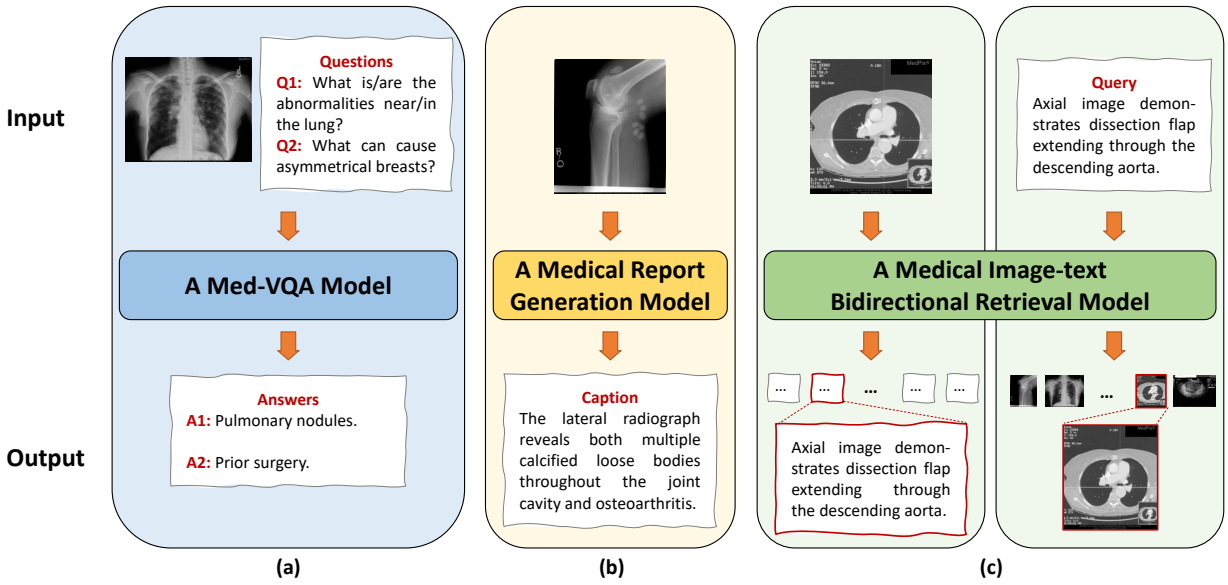


Figure 1: Medical vision-language tasks. (a) Medical visual question answering (Med-VQA). (b) Medical report generation. (c) Medical image-text retrieval.

quality medical VL datasets for pre-training. Due to privacy concerns and copyright issues, the acquisition of medical image-text data is difficult. Furthermore, annotating medical data requires significant domain knowledge and medical expertise, which is cost-prohibitive and time-consuming. So far, there are only a few studies investigating medical VLP, and they are not comprehensive enough. MMBERT (Khare et al., 2021) only targets for one downstream task – Med-VQA. MedViLL (Moon et al., 2021) and Clinical-BERT (Yan and Pei, 2022) utilize MIMIC-CXR (Johnson et al., 2019) – a large single-modality dataset with chest X-ray images for pre-training, and hence the pre-trained models cannot effectively deal with downstream VL tasks involving other imaging modalities or body regions. The most recent works ARL (Chen et al., 2022b) and M3AE (Chen et al., 2022a) utilize stronger backbone networks including pre-trained CLIP-ViT (Radford et al., 2021) and RoBERTa (Liu et al., 2019) along with external knowledge or masked autoencoder (He et al., 2022) to improve pre-training performance. However, they cannot deal with generation tasks due to the use of bidirectional Transformer like RoBERTa as text encoder.

Present Work. To provide a comprehensive perspective on medical VLP, we conduct a thorough empirical study based on a *unified* framework – vision-language Transformer (VLT) (Tan and Bansal, 2019; Zhou et al., 2020), which can deal with *both generation and understanding tasks*. To prepare datasets for pre-training, we examine existing large-scale radiographic VL datasets including ROCO (Pelka et al., 2018), MedICaT (Subramanian et al., 2020), and MIMIC-CXR (Johnson et al., 2019). Since these datasets are either noisy or of single imaging modality, to make sound and quick decisions on pre-training settings, we propose to construct a high-quality dataset of diverse radiographic imaging modalities. Specifically, we collected and filtered image-caption pairs from MedPix an online open-access database, and manually cleaned both image and text data to obtain 18,434 image-caption pairs, forming the RadioGraphy Captions (RGC) dataset, which can be used for pre-training and making pre-training decisions, or as a benchmark to evaluate the performance of the pre-trained models and existing ones. With RGC and the above mentioned datasets, we conduct a comprehensive empirical study on pre-training decisions including visual backbone, pre-training objective, and pre-training dataset. Further, we evaluate

the effectiveness of the pre-trained VLTs on downstream Med-VL tasks including Med-VQA, report generation, and image-text retrieval, compared with state-of-the-art methods. The key findings from our empirical study include:

- A small but high-quality in-domain dataset is useful for medical VLP and can be more effective than some existing radiographic datasets of much larger size. Data distribution (*e.g.*, diversity of imaging modality), data quantity, and data quality all significantly impact the performance of the pre-trained VLTs.
- The pre-trained VLTs demonstrate high effectiveness in downstream understanding tasks including Med-VQA and image-text retrieval. However, they are not effective for medical report generation, which may suggest the inadequacy of the pre-training data and pre-training method for generation tasks.

2. Unified Vision-language Pre-training

This section presents a unified vision-language pre-training framework for Med-VL applications including both understanding and generation tasks.

2.1. Model: Vision-language Transformer

We choose vision-language Transformer (VLT) as the model for pre-training, which has been the model of choice for multi-modal pre-training in the general domain. As shown in Figure 2 (left), it consists of a visual module, a text module, and a cross-modal Transformer encoder.

Visual Module. The visual module is a visual feature extractor that processes the input image and outputs its feature representation $V = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_N] \in \mathbb{R}^{N \times d}$, where N is the number of visual tokens, and d is the number of hidden states of the cross-modal Transformer. The visual feature extractor can be a pre-trained ResNet as used in some medical vision-language tasks (Gong et al., 2021; Chen et al., 2020b), or others such as linear patch (Kim et al., 2021). Here, we choose Vision Transformer (Dosovitskiy et al., 2021) and its variant (*e.g.*, Swin Transformer (Liu et al., 2021f)), which have shown promising performance in the general domain but never been used in medical vision-language tasks. Specifically, an image is split into non-overlapping patches, which are treated as visual tokens and fed to the Vision Transformer to extract visual features. Note that in the general domain, it is common to treat bottom-up attention (Anderson et al., 2018) obtained by an object detector (*e.g.*, Faster R-CNN) as visual tokens for VLT. However, it is not applicable in the medical domain due to the lack of annotated medical images with object labels to train the object detector.

Text Module. The text module is a tokenizer for processing the input text sequence. Specifically, we use WordPiece (Schuster and Nakajima, 2012) to tokenize the text sequence and obtain learnable embeddings $T = [\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_L] \in \mathbb{R}^{L \times d}$ by embedding lookup, where L is the length of the tokenized sequence. Three special tokens, [CLS], [SEP], and [END] with learnable embeddings t_{CLS}, t_{SEP}, and t_{END} respectively, are also introduced as in many other VL tasks, and so are learnable positional embeddings and segmentation embeddings. Note that similar to previous VLP works (Zhou et al., 2020; Yan and Pei, 2022; Li et al., 2020), we do not use an additional text encoder such as BERT (Devlin et al., 2019) as in Dou et al. (2022) to improve model capacity, because a bidirectional text encoder is not compatible with the downstream generation task.

Cross-modal Transformer Encoder. The cross-modal Transformer encoder consists of several multi-head self-attention layers. In each layer, an attention mask is used to control whether one token can attend to others. As illustrated in Figure 2 (right), a bidirectional attention mask (as used in BERT or RoBERTa) allows each token to see the tokens on its both sides, whereas a seq2seq attention mask only allows each token to attend to those on its left – so it can be used for generation tasks. Following Unified VLP (Zhou et al., 2020), we use both types of attention masks in the Transformer encoder so the VLP can deal with both understanding and generation tasks.

We describe how to fine-tune the pre-trained VLT on downstream Med-VL tasks including Med VQA, report generation, and image-text retrieval in Section 2.3.

2.2. Pre-training Objectives

To train the VLT, we adopt two widely-used self-supervised pre-training objectives: masked language modeling and image-text matching.

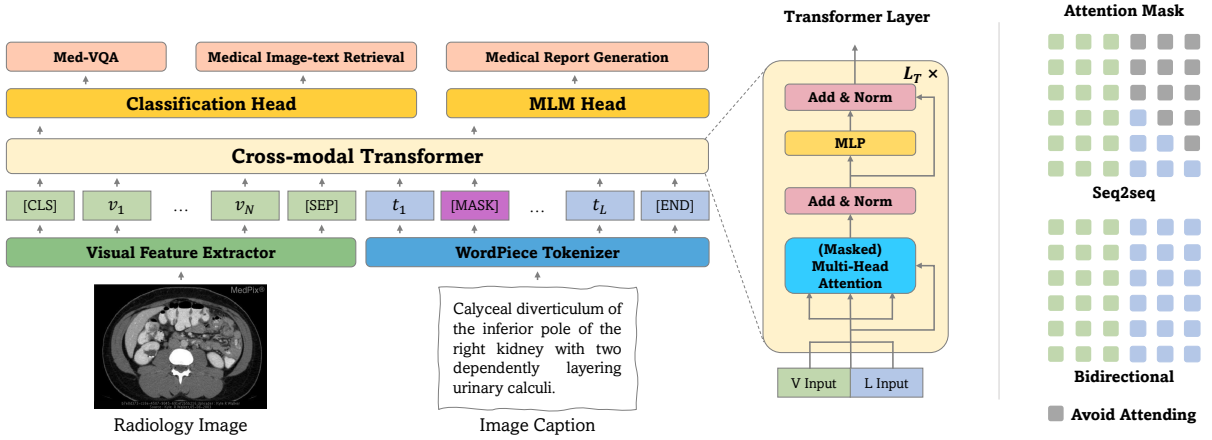


Figure 2: Overall architecture of the unified vision-language Transformer for both pre-training and fine-tuning. Both stages share the same network architecture. The main differences are in the input data and the training objectives. For pre-training, the input data is image-caption pairs from the pre-training corpus, with MLM or ITM as the training objective. For fine-tuning, the input data is from the training set of downstream tasks, and the training objective is task-dependent.

Masked Language Modeling (MLM). MLM (Devlin et al., 2019) replaces 15% of text tokens with a special token [MASK], a random token, or the original token with probabilities of 80%, 10%, and 10% respectively. The model is trained to predict the masked tokens given unmasked ones. Following Unified VLP (Zhou et al., 2020), we train with two sub-objectives, seq2seq MLM and bidirectional MLM, based on different attention masks as introduced in Figure 2, which alternates with probability α and $1-\alpha$ respectively.

Image-text Matching (ITM). ITM (Tan and Bansal, 2019; Li et al., 2019b) is similar to next sentence prediction (Devlin et al., 2019). For each image-text pair, the text will be replaced with another text with a probability of 50%. The model is trained to determine whether a given image-text pair is matched or not.

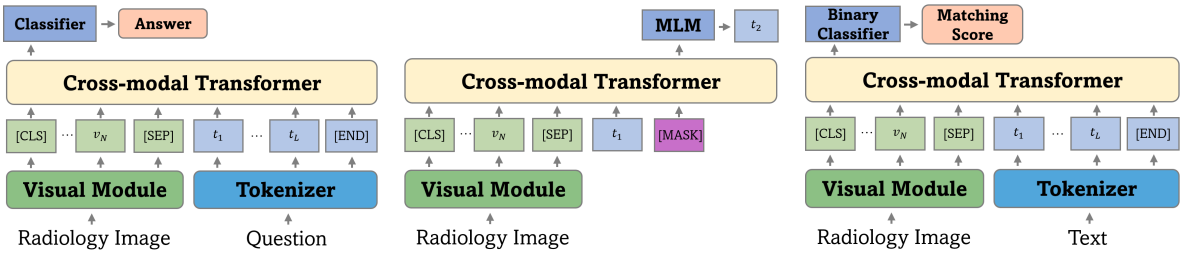
2.3. Fine-tuning the Pre-trained VLT for Medical Vision-language Tasks

In the following, we describe how to fine-tune the pre-trained VLT on downstream medical VL tasks.

Medical Visual Question Answering. Similar to VQA, a Med-VQA model aims to find a correct answer given a clinical question related to a medical image. In existing literature, it is commonly formulated as a classification task, *i.e.*, the model is trained to choose the correct answer from a list of candidate answers. To fine-tune the pre-tained VLT, a classifier (*e.g.*, MLP) is attached on top of the output of the [CLS] token by the cross-modal Transformer to predict the answer. The fine-tuned VLT and trained classifier can then be directly used for inference on test data, as shown in Figure 3(*a*).

Medical Report Generation. Given a radiography image, it aims to generate a clinical text that can accurately describe the image. Following Unified VLP (Zhou et al., 2020), we use seq2seq MLM as the objective to fine-tune the pre-trained VLT on the training data of this task. The fine-tuned VLT can be directly used for report generation, as illustrated in Figure 3(*b*), which shows a single step of the generation process with the seq2seq attention mask. The token t_1 is generated by the previous step, and a special token [MASK] is appended to t_1 to predict the next token, *i.e.*, t_2, which is generated by a token classifier (*e.g.*, a seq2seq MLM head).

Medical Image-text Retrieval. Image-text retrieval considers two sub-tasks, image retrieval given text queries and text retrieval given image queries. Following (Qi et al., 2020), we attach a binary classifier on top of the cross-modal Transformer and fine-



(a) Med-VQA. (b) Medical report generation. (c) Medical image-text retrieval.

Figure 3: Apply the pre-trained and fine-tuned VLT for inference on downstream medical VL tasks.

tune the pre-trained VLT on a training set of image-text pairs. Note that we augment the training data by randomly swapping images or texts to produce negative pairs. The model is trained to determine whether a given image-text pair is matched. For inference, the fine-tuned VLT and trained classifier is directly applied to predict the matching score of any image-text pair, as shown in Figure 3(*c*). The matching scores will be used for ranking relevant images or texts w.r.t. the query.

3. Radiographic Vision-language Datasets for Pre-training

In this section, we briefly introduce existing radiographic VL datasets and present our RGC dataset.

Existing Radiographic Vision-language Datasets. To date, there are five large public radiographic VL datasets that can be potentially used for VLP, including MIMIC-CXR (Johnson et al., 2019), IU X-Ray (Demner-Fushman et al., 2016), MedICaT (Subramanian et al., 2020), and ROCO (Pelka et al., 2018). However, as summarized in Table 2, they all have limitations. Both MIMIC-CXR and IU X-Ray only include *images of a single modality on a specific body region* – chest X-ray images. Many images in MedICaT have multiple sub-figures whose captions are concatenated to form a text, resulting in weak image-caption alignment. ROCO is built on archives/documents from PubMed Central[4] and automatically filters out non radiographic images and images with sub-figures. However, it may inevitably contain invalid captions or images with unwanted annotations. There are other VL datasets including CheXpert (Irvin et al., 2019), BIMCV COVID-19+ (Vayá et al., 2020), PadChest (Bustos et al., 2020), and FFA-IR (Li et al., 2021c). However, CheXpert is not a public dataset; BIMCV COVID-19+ is in Spanish; PadChest is similar to MIMIC-CXR but with a smaller size; FFA-IR contains fundus fluorescein angiography images and is of single modality like MIMIC-CXR. Hence, we do not use them for pre-training.

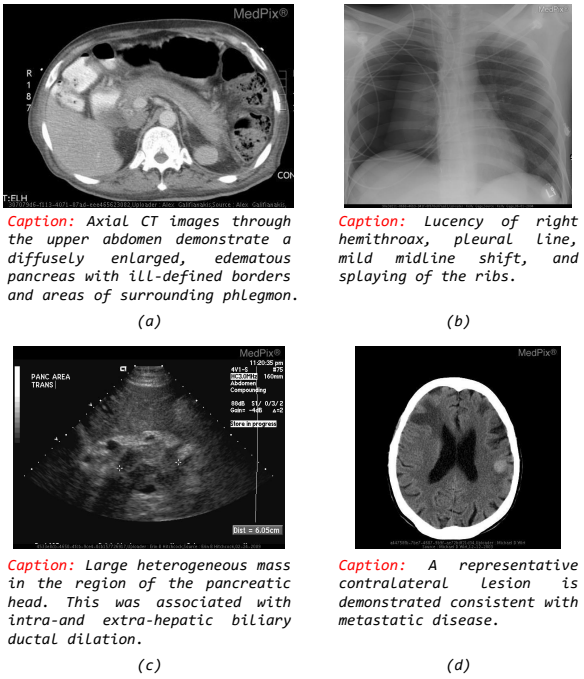


Figure 4: Samples from our RGC dataset with various radiographic imaging modalities. (a) Abdomen CT. (b) Chest X-ray. (c) Liver ultrasound. (d) Brain MRI.

We propose to construct a high-quality radiographic VL dataset with various imaging modalities,

4. https://www.ncbi.nlm.nih.gov/pmc/

Modality	MRI	CT	X-ray	Ultrasound	Not provided
# samples	6471	6922	3510	832	699

Table 1: The statistics of imaging modalities of RGC.

Datasets	ROCO	MedICaT	MIMIC-CXR	IU X-Ray	RGC
# image-caption pairs	87952	217060	473057	7470	18434
Multiple imaging modalities	✓	✓	✗	✗	✓
Various body regions	✓	✓	✗	✗	✓
Good image-caption alignment	✓	✗	✓	✓	✓
Manually cleaned	✗	✗	✓	✗	✓
Rich annotations	✗	✗	✗	✗	✓

Table 2: Comparison of different radiographic vision-language datasets.

which can be effectively used as a pre-training corpus and for making pre-training decisions, as well as a medical VL benchmark for evaluating the performance of the pre-trained models and existing methods.

3.1. Construction of RGC Dataset

The construction of our proposed Radiography Captions (RGC) dataset is built on MedPix, a free open-access online database with 37,997 image-caption pairs[5]. MedPix is intended to train medical students and physicians. Each case in MedPix contains multiple image-caption pairs of different imaging modalities (e.g., MRI, CT and Ultrasound), as well as detailed descriptions and rich annotations, including history, findings, diagnosis, treatment, and discussion, providing valuable resources for medical AI research. For each case, we form image-caption pairs using all images under the case and the summary caption associated with each image.

However, MedPix also includes some incomplete cases. Since captions are not compulsory, some of them are placeholders or invalid, and useful information may be provided in the explanation note but not in the caption. In addition, some images and captions contain noisy information. Hence, it is necessary to clean the raw data.

Filtering Based on Images and Captions. First, roughly 1,500 images are automatically removed because the caption is too short or meaningless. Second, about 10,000 images are manually inspected and removed. Next, about 8,000 images are manually removed because the captions are not useful for training a machine learning model. Specifically, the manual cleaning process includes

- Manually removing non-radiology images or images with too many (*e.g.*, > 16) sub-figures.
- Manually removing images with teaching annotations (e.g., arrows, words), because these marks may introduce bias during training.
- Manually removing images whose caption contains specific numerical information (e.g., "tumor with 2 cm area of central necrosis", "a 4 mm osteochondral defect", "a previous CT 4 days prior") or comparative descriptions (e.g., "compared to prior diastolic image") that involve multiple images, because these information cannot be learned by existing machine learning models, and including them would lower the quality of RGC as a benchmark.

After manual cleaning, we obtain 18,434 image-caption pairs. We then fix typos and clean noisy words from the captions, i.e., redundant words that do not have semantics (e.g., "Figure 1" and HTML tags). Finally, we divide the 18,434 clean image-caption pairs into a training set and a test set with a ratio of 9:1, which can be used as a new multi-modality benchmark for medical report generation and image-text retrieval. Table 1 shows the statistics of imaging modalities of RGC. Note that most of existing benchmarks such as MIMIC-CXR (Johnson

5. As of the submission of this manuscript, the number has grown to over 59,000, but the newly added cases are not downloadable yet.

Body region	Head	Chest	Abdomen	Neck	Pelvic cavity
VQA-RAD	104(CT/MRI)	107(X-Ray)	104(CT)	-	-
SLAKE	140(CT/MRI)	219(X-ray)	201(CT/MRI)	41(CT)	41(CT)

Table 3: Statistics of Med-VQA datasets.

et al., 2019) and IU X-Ray (Demner-Fushman et al., 2016) are of single modality.

4. Experiments

In this section, we conduct extensive experiments to study the effectiveness of unified VLP for various medical VL tasks. We first compare different pre-training objectives, visual modules, and pre-training datasets. Then, we report the results of VLT on three downstream tasks. We also report additional experimental results in Appendix A.

4.1. Implementation Details

We conduct experiments on PyTorch[6]. The structure of cross-modal Transformer is the same as BERT-base(Devlin et al., 2019) from HuggingFace[7] (Wolf et al., 2020), which has 12 Transformer layers with 12 attention heads and 768 hidden units. As for visual backbones, we conduct experiments on ResNet (He et al., 2016), Vision Transformer (Dosovitskiy et al., 2021), Swin Transformer (Liu et al., 2021f), and linear patch (Kim et al., 2021). We use the pre-trained weights of ResNet101 and ViT-B/16 from PyTorch. We use the official implementation of Swin-S[8]. All the images are resized into $(3, 224, 224)$. In ResNet and Swin Transformer, the output of the last adaptive pooling layer with size $(2048, 7, 7)$, is transposed into $(49, 768)$ by a fully connected layer and then fed into the cross-modal Transformer. During pre-training, the alternating probability α between 2 types of attention masks is set to 0.5 during pre-training. In the fine-tuning stage, α is set to 1 on Med-VQA and image-text retrieval tasks (*i.e.*, the attention mask is always bidirectional), and 0 on report generation tasks. All experiments reported are conducted on a single NVIDIA GeForce RTX 3090 24GB. The code for the experiments and materializing the dataset is submitted along with the paper in the OpenReview.

6. https://pytorch.org/
7. https://huggingface.co/transformers/index.html
8. https://github.com/microsoft/Swin-Transformer

Optimizer. We use AdamW optimizer (Loshchilov and Hutter, 2017) for all the experiments. The weight decay is set to 1e-4. During pre-training, the learning rate is 4e-5. On the downstream tasks, the learning rate is set to 1e-6 for image-text retrieval, and 2e-5 for Med-VQA and 1e-5 for report generation.

Batch Size. The batch size for pre-training and report generation is 32, and 64 for others. When ViT-B is the visual backbone, with a batch size of 64, it will lead to an out-of-memory error for a 24GB GPU and hence we reduce the batch size to 48 on downstream tasks.

4.2. Influence of Pre-training Settings

Following the common practice in (Dou et al., 2022), we mainly conduct empirical comparisons on Med-VQA tasks, since it is the fastest way to assess the comprehension ability of the pre-trained models. We experiment with two Med-VQA benchmarks: VQA-RAD (Lau et al., 2018) and SLAKE (Liu et al., 2021b). VQA-RAD contains 315 images also from MedPix with 3064 question-answer pairs for training and 451 for testing. In this paper, we use the English version of SLAKE, which contains 642 radiology images with 4919 question-answer pairs for training, 1053 for validation, and 1061 for testing. Table 3 shows the statistics of Med-VQA datasets w.r.t body regions. For SLAKE, there are 282 CT images, 181 MRI images, and 179 X-Ray images. For VQA-RAD, they do not provide detailed modality information. We report average accuracy (%) with standard deviation over 10 runs. 0.8

On Pre-training Objectives. We use the small version of Swin Transformer (Swin-S) as the visual backbone to compare different pre-training objectives. The results are provided in Table 4, where we can draw the main observation:

- MLM is much more effective than ITM as a pre-training objective, and the latter is not consistently useful.

Datasets	Pre-training Objectives	VQA-RAD	SLAKE
No Pre-training		65.28 ± 0.98	79.74 ± 0.97
RGC	MLM	**69.77 ± 0.66**	**81.50 ± 0.36**
	ITM	65.47 ± 1.01	80.31 ± 0.95
	MLM+ITM	67.54 ± 1.16	80.71 ± 0.45
RGC+ROCO	MLM	**70.43 ± 0.51**	**82.34 ± 0.44**
	ITM	64.96 ± 1.03	81.18 ± 0.70
	MLM+ITM	67.29 ± 1.23	82.27 ± 0.57

Table 4: Comparison of pre-training objectives with Swin-S as the visual backbone. Average accuracy (%) with standard deviation over 10 runs is reported.

Models	VQA-RAD	SLAKE
Linear/16 (w/o pt)	59.87 ± 1.46	73.71 ± 1.14
Linear/16	69.56 ± 0.96	78.23 ± 0.36
ViT-B/16 (w/o pt)	61.97 ± 1.35	75.45 ± 0.61
ViT-B/16	67.81 ± 1.07	79.25 ± 0.48
Resnet-101 (w/o pt)	64.34 ± 1.48	77.58 ± 0.67
Resnet-101	69.62 ± 1.03	80.03 ± 0.49
Swin-S (w/o pt)	65.28 ± 0.98	79.74 ± 0.97
Swin-S	**69.77 ± 0.66**	**81.50 ± 0.36**

Table 5: Comparison of vision modules. Average accuracy (%) with standard deviation over 10 runs is reported. The models are pre-trained on RGC. pt: pre-training; B: base; S: small.

Different from the observation in (Dou et al., 2022), ITM does not help to improve model performance in medical VLP, which is an echo of the finding in (Zhou et al., 2020). It is probably due to the differences in image modality and scale of pre-training data. The radiology images tend to be similar, and ITM may not be adequate for the model to learn meaningful representations, whereas in the general domain there is often a large difference between two images. In addition, the scale of the pre-training data for general VLP is around 10M, much larger than that used in our pre-training. Since the number of negative samples in our pre-training data is limited, ITM is less effective.

On Visual Backbones. We compare different visual backbones in Table 5. All the models with different visual backbone are first pre-trained on RGC with MLM and then fine-tuned on Med-VQA datasets. We can make the following observations:

- With each different visual backbone, pre-training on RGC helps to boost the performance of VLT on both Med-VQA datasets.
- Visual backbones with locality (*i.e.*, Resnet and Swin Transformer) outperform those without it, and Swin Transformer achieves the best overall performance on both datasets.

We can also observe that Linear/16 achieves comparable performance with Swin-S on VQA-RAD after pre-training, but fail to generalize well on SLAKE. It is probably because the number of images in VQA-RAD is quite limited.

On Pre-training Datasets. We compare the medical VL datasets introduced in Table 2. We also try using the unfiltered raw MedPix data (about 38k image-caption pairs) for pre-training. We use Swin-S as the visual backbone and MLM as the pre-training objective. The results are summarized in Table 6. Note that the statistics of image-caption pairs (# of samples) may slightly differ from the original datasets as we filter out some invalid samples. According to the results, we can make the following observations:

- The best corpus for pre-training is the combination of RGC, ROCO and MedICaT. Further including MIMIC-CXR for pre-training decreases model performance. It shows both data quantity and data distribution are important, and the large mass of single-modality data in MIMIC-CXR may introduce bias during pre-training.
- The VLT pre-trained with RGC can achieve comparable results with those pre-trained with much larger datasets (*e.g.*, MedICaT, MIMIC-CXR), demonstrating its effectiveness.
- The single-modality dataset MIMIC-CXR is the least effective for pre-training, despite being the largest dataset, which shows the importance of pre-training with multi-modality images.
- The VLT pre-trained with ROCO outperforms the one pre-trained with RGC, showing the influence of dataset scale on pre-training. Note that while the differences in performance are small, the results are statistically significant (the p-values for differences between RGC and ROCO are 0.076 on VQA-RAD and 0.00056 on SLAKE).
- The VLT pre-trained with RGC performs slightly better than the one pre-trained with MedPix.

Pre-training datasets					# samples	Accuracy (pre-trained model)	
RGC	ROCO	MedICaT	MIMIC-CXR	MedPix		VQA-RAD	SLAKE
✗	✗	✗	✗	✗	0	65.28 ± 0.98	79.74 ± 0.97
✓	✗	✗	✗	✗	∼17k	69.77 ± 0.66	81.50 ± 0.36
✗	✓	✗	✗	✗	∼65k	70.31 ± 0.62	82.16 ± 0.43
✗	✗	✓	✗	✗	∼217k	68.92 ± 1.26	81.81 ± 0.65
✗	✗	✗	✓	✗	∼271k	65.34 ± 1.01	81.53 ± 0.48
✗	✗	✗	✗	✓	∼38k	69.53 ± 0.59	81.54 ± 0.51
✓	✓	✗	✗	✗	∼82k	70.43 ± 0.51	82.34 ± 0.44
✓	✗	✓	✗	✗	∼234k	70.07 ± 0.84	82.18 ± 0.47
✗	✓	✓	✗	✗	∼282k	**71.25 ± 1.02**	83.08 ± 0.55
✓	✓	✓	✗	✗	∼299k	**71.20 ± 0.83**	**83.40 ± 0.57**
✓	✓	✓	✓	✗	∼570k	70.09 ± 0.76	83.13 ± 0.54

Table 6: Comparison of different pre-training datasets and their combinations. Average accuracy (%) with standard deviation over 10 runs is reported. ✓ means used in pre-training, and ✗ means not.

Pre-training Datasets	VQA-RAD
RGC w/ overlapping images	69.77 ± 0.66
RGC w/o overlapping images	69.80 ± 0.84

Table 7: Evaluating the impact of pre-training with overlapping images from RGC and VQA-RAD on model performance.

Note that the size of RGC is less than half of MedPix, which demonstrates the effectiveness of our data cleaning strategy.

Since both RGC and VQA-RAD are collected from MedPix, they have some overlapping data. However, it does not influence the conclusions drawn from our experiments. This is because the train-test split of VQA-RAD is based on questions instead of images (each image is associated with several questions), and the training set contains all the images in VQA-RAD. So, when a model is trained/fine-tuned on VQA-RAD, it will see all the images anyway. To support this claim, we exclude the 188 images overlapped with VQA-RAD from our RGC dataset and use the rest for pre-training. The results in Table 7 show that prediction accuracy on the test set of VQA-RAD is 69.80 ± 0.84 (averaged over 30 runs), which is very close to the result of 69.77 ± 0.66 obtained by using all images in RGC for pre-training. This supports our statement that the overlapping images used in pre-training will not affect the conclusions drawn from the reported results.

4.3. Effect of Pre-training on Med-VL Tasks

In this section, we evaluate the effectiveness of the pre-trained VLT on three downstream medical VL tasks. Unless otherwise stated, the VLT is pre-trained with MLM on a combination of three datasets including RGC, ROCO, and MedICaT.

Medical Visual Question Answering. We conduct experiments on VQA-RAD and SLAKE and compare VLT with general VQA models (*i.e.*, SAN (Yang et al., 2016), MFH (Yu et al., 2018), MCB (Fukui et al., 2016), MUTAN (Ben-Younes et al., 2017), BAN (Kim et al., 2018)) as well as Med-VQA models (*i.e.*, MEVF (Nguyen et al., 2019), CPRD (Liu et al., 2021a), CMSA-MTPT (Gong et al., 2021), MMBERT (Khare et al., 2021), MedViLL (Moon et al., 2021), M3AE (Chen et al., 2022a)). Specifically, MEVF is trained in a semi-supervised manner to overcome the lack of medical training data. Both CPRD and CMSA-MTPT propose to pre-train three visual feature extractors (*i.e.*, Resnet) for three image modalities (*i.e.*, abdomen CT, chest X-ray and brain MRI) respectively, on current datasets (Lau et al., 2018; Liu et al., 2021b), which can be integrated with cross-modal networks such as BAN and BERT (Devlin et al., 2019). CR utilizes question types to improve model performance. MMBERT and MedViLL pre-train a VLT on ROCO and MIMIC-CXR, respectively. To have a fair comparison, we do not compare with previous works that use additional label information as in Zhan et al. (2020); Liu et al. (2023) or external knowledge as in Chen et al. (2022b). Following most previous works,

Models	VQA-RAD			SLAKE		
	Overall	Open-ended	Closed-ended	Overall	Open-ended	Closed-ended
SAN (Lau et al., 2018; Yang et al., 2016)	54.3	31.3	69.5	76.0	74.0	79.1
MFH (Yu et al., 2018)	57.9	35.2	72.8	75.9	73.6	79.3
MCB (Lau et al., 2018; Fukui et al., 2016)	58.1	38.0	71.3	76.1	73.2	80.5
MUTAN (Ben-Younes et al., 2017)	58.1	34.1	73.9	76.8	73.6	81.7
BAN (Nguyen et al., 2019; Kim et al., 2018)	58.3	37.4	72.1	76.3	74.6	79.1
MEVF+BAN (Nguyen et al., 2019)	66.1	49.2	77.2	78.6	77.8	79.8
CP+BAN (Liu et al., 2021a)	68.1	53.1	77.9	80.9	79.1	83.7
CMSA-MTPT (Gong et al., 2021)	71.4	60.9	78.3	80.5	78.2	84.0
MMBERT* (Khare et al., 2021)	72.0	**63.1**	77.9	-	-	-
MedViLL* (Moon et al., 2021)	70.9	59.7	78.2	-	-	-
Swin-S+VLT (w/o pt)	66.5	53.6	75.0	80.5	79.2	82.5
Swin-S+VLT (w/ pt)	**72.1**	60.9	**79.4**	**84.0**	**81.9**	**87.3**

Table 8: Test accuracy (%) of our VLTs and baselines on VQA-RAD and SLAKE. * indicates the results are not reproduced but copied from the original paper due to unavailability of open source codes or key information not provided in the released code.

we report test accuracy on VQA-RAD and SLAKE. We can observe from the results in Table 8 that:

- Pre-training can significantly improve model performance on Med-VQA tasks.
- Our pre-trained VLT outperforms state-of-the-art methods on two benchmarks.

Medical Report Generation. We conduct experiments on MIMI-CXR, IU X-Ray (Demner-Fushman et al., 2016), and RGC. Current works on report generation (Jing et al., 2018; Li et al., 2018, 2019a; Jing et al., 2019; Chen et al., 2020b; Liu et al., 2021c; Chen et al., 2020b; Liu et al., 2021d; Yan and Pei, 2022; Liu et al., 2021e; Jain et al., 2021) mainly focus on chest X-ray images and experiment with MIMIC-CXR and IU X-Ray. R2GEN (Chen et al., 2020b) proposes a relational memory and memory-driven layer normalization in a standard Transformer for generation. CMCL (Liu et al., 2021c) proposes to use the most suitable samples to train the model based on current model competence. PPKED (Liu et al., 2021d) proposes a knowledge explorer and distiller to generate reports. KGAE (Liu et al., 2021e) leverages an external knowledge graph. We follow R2GEN to preprocess MIMI-CXR and IU X-Ray and compare with existing works using the same pre-processing pipeline. Beam search with the beam size 3 is used. CMCL, PPKED, KGAE, and Clinical-BERT are tailored for generating reports for chest X-ray images, and the source codes are not released. Hence, we only compare with R2GEN on RGC. The results in Table 9 suggest:

- Pre-training does not help to improve the performance on generation tasks and can even decrease the model performance on IU X-Ray. It is probably because generation tasks are different and more complicated than classification tasks (*e.g.*, Med-VQA and image-text retrieval), which suggests the inadequacy of current pre-training data and strategy for generation tasks.
- Resnet and Swin Transformer achieve comparable performance on MIMIC-CXR and IU X-Ray, but there is a large gap between their performance on RGC. It shows that RGC as a multi-modality dataset is a better benchmark for report generation than single-modality datasets, as it can reflect the difference in model capacity for learning visual features.

Medical Image-text Retrieval. We conduct experiments on RGC and IU X-Ray. Previous works on this task (Hsu et al., 2018; Moon et al., 2021; Wang et al., 2021; Huang et al., 2021a; Zhang et al., 2020) are mainly designed for chest X-ray datasets, *e.g.*, MIMIC-CXR and CheXpert (Irvin et al., 2019). However, CheXpert is not released, and the source codes for (Zhang et al., 2020; Wang et al., 2021; Hsu et al., 2018) are unavailable. Hence, we only compare with VSE++ (Faghri et al., 2018), a strong baseline for image-text retrieval. Following previous works, we report Recall@K ($K = 1, 5, 10$). The results presented in Table 10 suggest that:

- Pre-training can greatly improve model performance on image-text retrieval tasks.

Dataset	Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER-D
MIMIC-CXR	R2GEN (Chen et al., 2020b)	0.353	0.218	0.145	0.103	0.142	0.277	-
	CMCL* (Liu et al., 2021c)	0.344	0.217	0.140	0.097	0.133	0.281	-
	PPKED* (Liu et al., 2021d)	0.360	0.224	0.149	0.106	0.149	0.284	-
	KGAE* (Liu et al., 2021e)	0.369	**0.231**	**0.156**	**0.118**	0.153	0.295	-
	Clinical-BERT* (Yan and Pei, 2022)	**0.383**	0.230	0.151	0.106	0.144	0.275	-
	Resnet-101+VLT(w/ pt)	0.339	0.197	0.124	0.093	0.154	0.298	-
	Swin-S+VLT (w/o pt)	0.340	0.198	0.127	0.090	0.146	0.289	-
	Swin-S+VLT (w/ pt)	0.340	0.209	0.139	0.099	**0.166**	**0.306**	-
IU X-Ray	HRGR (Li et al., 2018)	0.438	0.298	0.208	0.151	-	0.322	-
	CMAS-RL (Jing et al., 2019)	0.464	0.301	0.210	0.154	-	0.362	-
	R2GEN (Chen et al., 2020b)	0.470	0.304	0.219	0.165	0.187	0.371	-
	CMCL* (Liu et al., 2021c)	0.473	0.305	0.217	0.164	0.186	0.378	-
	PPKED* (Liu et al., 2021d)	0.483	0.315	0.224	0.168	0.190	0.376	0.351
	KGAE* (Liu et al., 2021e)	**0.512**	0.327	**0.240**	**0.179**	**0.195**	0.383	-
	Clinical-BERT* (Yan and Pei, 2022)	0.495	**0.330**	0.231	0.170	-	0.376	0.432
	Resnet-101+VLT (w/ pt)	0.423	0.266	0.186	0.136	0.180	0.393	0.388
	Swin-S+VLT (w/o pt)	0.461	0.297	0.214	0.154	0.190	**0.404**	**0.462**
	Swin-S+VLT (w/ pt)	0.429	0.265	0.185	0.137	0.184	0.387	0.439
RGC	R2GEN (Chen et al., 2020b)	0.404	0.335	0.316	0.298	0.193	0.360	2.381
	Resnet-101+VLT (w/ pt)	0.403	0.352	0.330	0.318	0.214	0.357	2.554
	Swin-S+VLT (w/o pt)	0.490	**0.453**	0.435	0.419	0.280	**0.459**	**3.645**
	Swin-S+VLT (w/ pt)	**0.491**	0.452	**0.436**	**0.420**	**0.282**	0.455	3.535

Table 9: Results of generation tasks on MIMIC-CXR, IU X-Ray, and RGC. BLEU-N (N=1,2,3,4) (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and CIDER-D (Vedantam et al., 2015) scores are reported. * indicates the results are not reproduced but copied from the original paper due to unavailability of open source codes.

	RGC						IU X-Ray					
	Text Retrieval			Image Retrieval			Text Retrieval			Image Retrieval		
Models	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
VSE++	26.27	43.27	50.89	30.02	47.70	55.71	0.21	3.43	5.38	0.57	3.24	5.83
Resnet-101+VLT (w/ pt)	24.18	50.46	62.36	30.77	59.00	70.09	1.02	3.39	5.59	1.19	4.24	7.12
Swin-S+VLT (w/o pt)	17.02	38.52	51.70	18.31	43.71	57.16	0.68	2.37	4.24	0.34	1.86	5.59
Swin-S+VLT (w/ pt)	**29.42**	**54.59**	**67.55**	**34.72**	**63.76**	**75.66**	**1.53**	**6.78**	**10.51**	**2.03**	**5.93**	**10.00**

Table 10: Image-text retrieval results on RGC and IU X-Ray. Recall@k (%) is used as the evaluation metric.

- Swin Transformer significantly outperforms Resnet on retrieval tasks, just as in Med-VQA tasks and report generation tasks on RGC.

5. Conclusions and Limitations

This paper makes two main contributions. 1) We present a comprehensive empirical study on medical VLP, providing analysis on pre-training decisions and evaluating the effectiveness of the pre-trained VLTs on both generation and understanding Med-VL tasks. We distill the experimental results into several key observations which can be used as a guide to future VLP research. The pre-trained VLTs can also serve as strong baselines for future research. 2) We propose RGC, a high-quality radiographic VL dataset of multiple imaging modalities, which can be used as a pre-training dataset or a new benchmark for medical report generation and medical image-text retrieval, to supplement the very small pool of existing Med-VL benchmarks. One limitation of RGC is that its size is relatively small. However, since the MedPix database keeps growing, it provides opportunities to expand and improve RGC in the future.

Acknowledgment

We sincerely thank the MedPix team including Dr. Dina Demner-Fushman, Mr. Soumya Gayen, and Dr. James G. Smirniotopoulos for their kind help in hosting the RGC dataset on MedPix website. We would also like to thank the anonymous reviewers for their helpful comments. This research was partially supported by the grant of project P0038194 (1-ZVVX) funded by PolyU (UGC).

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, page 6077–6086, June 2018.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, June 2005.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *CVPR*, pages 2612–2620, 2017.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, Dec 2020. ISSN 1361-8415.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

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 *ECCV*. Springer, 2020a.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. Generating radiology reports via memory-driven transformer. In *EMNLP*, 2020b.

Zhihong Chen, Yuhao Du, Jinpeng Hu, Yang Liu, Guanbin Li, Xiang Wan, and Tsung-Hui Chang. Multi-modal masked autoencoders for medical vision-and-language pre-training. In *MICCAI*. Springer, 2022a.

Zhihong Chen, Guanbin Li, and Xiang Wan. Align, reason and learn: Enhancing medical vision-and-language pre-training with knowledge. In *ACMMM*, page 5152–5161, 2022b.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *JAMIA*, 2016.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019.

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 *ICLR*, 2021.

Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, Zicheng Liu, and Michael Zeng. An empirical study of training end-to-end vision-and-language transformers. In *CVPR*, 2022.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12, 2018.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *EMNLP*, 2016.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*, 2020.

Haifan Gong, Guanqi Chen, Sishuo Liu, Yizhou Yu, and Guanbin Li. Cross-modal self-attention with

multi-task pre-training for medical visual question answering. In *ICMR*, pages 456–460, 2021. doi: 10.1145/3460426.3463584.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, June 2022.

Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

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 *ICCV*, pages 3942–3951, 2021a.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. Seeing out of the box: End-to-end pre-training for vision-language representation learning. In *CVPR*, 2021b.

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 *AAAI*, volume 33, pages 590–597, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In *NeurIPS Datasets and Benchmarks Track (Round 1)*, 2021.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. In *ACL*, pages 2577–2586, 2018. doi: 10.18653/v1/P18-1240.

Baoyu Jing, Zeya Wang, and Eric P. Xing. Show, describe and conclude: On exploiting the structure information of chest x-ray reports. In *ACL*, pages 6570–6580, 2019.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 2019.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *ISBI*. IEEE, 2021.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NIPS*, 2018.

Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, pages 5583–5594, 2021.

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*, 2018.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Knowledge-driven encode, retrieve, paraphrase for medical image report generation. In *AAAI*, 2019a.

Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021a.

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 arXiv:1908.03557*, 2019b.

Liunian Harold Li, Haoxuan You, Zhecan Wang, Alireza Zareian, Shih-Fu Chang, and Kai-Wei Chang. Unsupervised vision-and-language pre-training without parallel images and captions. In *NAACL*, pages 5339–5350, 2021b.

Mingjie Li, Wenjia Cai, Rui Liu, Yuetian Weng, Xiaoyun Zhao, Cong Wang, Xin Chen, Zhong Liu, Caineng Pan, Mengke Li, yingfeng zheng, Yizhi Liu, Flora D. Salim, Karin Verspoor, Xiaodan Liang, and Xiaojun Chang. FFA-IR: Towards an explainable and reliable medical report generation benchmark . In *NeurIPS Datasets and Benchmarks Track (Round 2)*, 2021c.

Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. In *ACL/IJCNLP*, pages 2592–2607, 2021d.

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 *ECCV*. Springer, 2020.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. In *NIPS*, volume 31, 2018.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL workshop on Text Summarization Branches Out*, pages 74–81, July 2004.

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 *MICCAI*. 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 *ISBI*. IEEE, 2021b.

Bo Liu, Li-Ming Zhan, Li Xu, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning and contrastive learning. *IEEE Transactions on Medical Imaging*, 42(5):1532–1545, 2023.

Fenglin Liu, Shen Ge, and Xian Wu. Competence-based multimodal curriculum learning for medical report generation. In *ACL*, pages 3001–3012, 2021c.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. Exploring and distilling posterior and prior knowledge for radiology report generation. In *CVPR*, pages 13753–13762, 2021d.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Sheng Wang, and Xu Sun. Auto-encoding knowledge graph for unsupervised medical report generation. In *NeurIPS*, pages 16266–16279, 2021e.

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, 2019.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002. IEEE, 2021f.

Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101, 2017.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23, 2019.

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 arXiv:2105.11333*, 2021.

Binh D Nguyen, Thanh-Toan Do, Binh X Nguyen, Tuong Do, Erman Tjiputra, and Quang D Tran. Overcoming data limitation in medical visual question answering. In *MICCAI*. Springer, 2019.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, July 2002.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. Radiology objects in context (roco): a multimodal image dataset. In *CVII-STENT and LABELS*, pages 180–189. Springer, 2018.

Di Qi, Lin Su, Jia Song, Edward Cui, Taroon Bharti, and Arun Sacheti. Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data. *arXiv preprint arXiv:2001.07966*, 2020.

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 *ICML*, pages 8748–8763. PMLR, 2021.

Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *ICASSP*, pages 5149–5152, 2012.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

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 *ICLR*, 2020.

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 *EMNLP Findings*, pages 2112–2120, 2020.

Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*, pages 5100–5111, 2019.

Maria De La Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco García-García, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

Xiaosong Wang, Ziyue Xu, Leo Tam, Dong Yang, and Daguang Xu. Self-supervised image-text pre-training with mixed data in chest x-rays. *arXiv preprint arXiv:2103.16022*, 2021.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. Transformers: State-of-the-art natural language processing. In *EMNLP*, 2020.

Bin Yan and Mingtao Pei. Clinical-bert: Vision-language pre-training for radiograph diagnosis and reports generation. In *AAAI*, 2022.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. Stacked attention networks for image question answering. In *CVPR*, 2016.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. on neural networks and learning systems*, 2018.

Li-Ming Zhan, Bo Liu, Lu Fan, Jiaxin Chen, and Xiao-Ming Wu. Medical visual question answering via conditional reasoning. In *ACM MM*, 2020.

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 arXiv:2010.00747*, 2020.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, volume 34, pages 13041–13049, 2020.

Appendix A. Additional Experiments

A.1. More Ablation Studies on Pre-training Settings

We provide additional experimental results of VLTs pretrained with different visual backbones and different datasets on downstream tasks including Med-VQA, report generation, and medical image-text retrieval.

Med-VQA. The results are shown in Table 12, and we can make the following observations:

- For each different visual backbone, the VLT pre-trained with larger dataset (*i.e.*, RGC + ROCO + MedICaT) consistently achieves better performance on both Med-VQA benchmarks.
- When pre-training with RGC + ROCO + MedICaT, ResNet and ViT achieve similar performance as Swin Transformer on VQA-RAD , but the latter performs much better on SLAKE.

Report Generation. The results are shown in Table 13. We can observe:

- Pre-training with larger dataset (*i.e.*, RGC + ROCO + MedICaT) does not help to improve model performance on downstream generation tasks.
- The VLTs with Swin-S as visual backbone achieve the best performance, though pre-training does not seem to make a difference.

Medical Image-text Retrieval. The results are summarized in Table 14, and we can make the following observations:

- Pre-training is highly effective for downstream medical image-text retrieval tasks. For each visual backbone, more pre-training data leads to overall better performance.
- The VLTs with Swin-S as visual backbone outperform others.

A.2. Pre-training with Out-of-domain Datasets

To demonstrate the importance of using in-domain data for pre-training, we use MSOCO Captions (Chen et al., 2015) and Conceptual Captions (CC) (Sharma et al., 2018) to pre-train VLTs with Swin-S as visual backbone and report the results on Med-VQA tasks in Table 11. We can observe that:

Datasets	Samples	VQA-RAD	SLAKE
None	N/A	65.28 ± 0.98	79.74 ± 0.97
RGC	∼17k	69.77 ± 0.66	81.50 ± 0.36
MSCOCO	∼414k	62.86 ± 1.67	80.57 ± 0.32
CC	∼2M	59.31 ± 1.32	79.23 ± 0.48

Table 11: Comparisons of VLTs pre-trained with in-domain and out-of-domain datasets. The visual backbone is Swin-S.

- Pre-training with a small in-domain dataset (*e.g.*, RGC) is much more effective than with large-scale out-of-domain datasets.
- Pre-training with out-of-domain datasets may have an adverse effect and significantly decrease model performance.

Visual Backbone	RGC	ROCO	MedICaT	VQA-RAD	SLAKE
Linear/16	✓	✗	✗	69.56 ± 0.96	78.23 ± 0.36
Linear/16	✓	✓	✓	70.41 ± 0.41	79.76 ± 0.57
ViT-B/16	✓	✗	✗	67.81 ± 1.07	79.25 ± 0.48
ViT-B/16	✓	✓	✓	**71.21 ± 0.89**	80.96 ± 0.63
Resnet-101	✓	✗	✗	69.62 ± 1.03	80.03 ± 0.49
Resnet-101	✓	✓	✓	**71.27 ± 0.46**	81.39 ± 0.64
Swin-S	✓	✗	✗	69.77 ± 0.66	81.50 ± 0.36
Swin-S	✓	✓	✓	**71.20 ± 0.83**	**83.40 ± 0.57**

Table 12: Comparison of VLTs pre-trained with different visual modules and different datasets for Med-VQA.

Visual Backbone	RGC	ROCO	MedICaT	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDER-D
Linear/16	✓	✗	✗	0.332	0.285	0.264	0.252	0.195	0.342	2.291
Linear/16	✓	✓	✓	0.346	0.302	0.281	0.270	0.213	0.361	2.500
ViT-B/16	✓	✗	✗	0.417	0.373	0.351	0.339	0.244	0.408	3.066
ViT-B/16	✓	✓	✓	0.404	0.365	0.338	0.326	0.241	0.397	2.919
Resnet-101	✓	✗	✗	0.404	0.359	0.332	0.320	0.218	0.361	2.560
Resnet-101	✓	✓	✓	0.403	0.352	0.330	0.318	0.214	0.357	2.554
Swin-S	✗	✗	✗	0.490	0.453	0.435	0.419	0.280	**0.459**	3.645
Swin-S	✓	✗	✗	0.491	**0.455**	0.435	0.420	0.281	0.455	**3.676**
Swin-S	✓	✓	✓	**0.491**	0.452	**0.436**	**0.420**	**0.282**	0.455	3.535

Table 13: Comparison of VLTs pre-trained with different visual modules and different datasets for report generation on RGC.

Visual Backbone	Pre-training datasets			Text Retrieval			Image Retrieval		
	RGC	ROCO	MedICaT	R@1	R@5	R@10	R@1	R@5	R@10
Linear/16	✗	✗	✗	15.31	26.39	34.72	15.52	31.96	40.08
Linear/16	✓	✗	✗	18.12	28.07	32.72	24.50	34.45	39.81
Linear/16	✓	✓	✓	21.85	40.78	51.27	27.37	48.40	57.54
ViT-B/16	✗	✗	✗	3.19	9.30	14.44	4.65	12.93	19.69
ViT-B/16	✓	✗	✗	19.04	37.21	45.43	26.07	47.97	54.89
ViT-B/16	✓	✓	✓	17.39	39.16	52.24	22.96	47.81	61.17
Resnet-101	✗	✗	✗	5.30	12.49	18.71	7.41	18.17	26.01
Resnet-101	✓	✗	✗	22.71	39.05	46.46	31.10	46.78	54.30
Resnet-101	✓	✓	✓	24.18	50.46	62.36	30.77	59.00	70.09
Swin-S	✗	✗	✗	17.02	38.52	51.70	18.31	43.71	57.16
Swin-S	✓	✗	✗	**30.86**	46.22	54.68	**35.32**	56.30	63.17
Swin-S	✗	✓	✓	22.88	47.65	61.71	29.53	57.17	71.61
Swin-S	✓	✓	✓	29.42	**54.59**	**67.55**	34.72	**63.76**	**75.66**

Table 14: Comparison of VLTs pre-trained with different visual modules and different datasets for medical image-text retrieval on RGC.