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

Supplementary Material

7. Architecture of latent adapter

Here, we provide detailed network architecture for latent adapters. We devise a novel latent adapter to transform different levels' multi-modal features into latent representations with its internal structure presented in Figure 4.

8. Implementation details

Pre-training. For the image-text encoder, we adopt the pre-trained ViT-g from [18] as preliminary visual embeddings extractor (E_I) and its hidden dimension (d_v) is set to 1408. We adopt a base-size Transformer encoder with 12 layers initialized from [4] and its hidden dimension (d_q, d_t) is set to 768. The length of learnable embeddings was set to 32 (L_q). For tokenizing medical reports specialized medicine vocabulary dictionary from [4] was employed with vocabulary size of 30522. All text inputs were confined to a maximum length of 95 (L_t). All images size was resized into $3 \times 224 \times 224$ ($C \times H \times W$) and normalized into range $(0, 1)$. Raw images were split into non-overlap patches by stride of 14×14 ($h \times w$) so that patch embeddings length was $16*16 = 256$ (L_v). The feature dimension of ITC linear heads was set to 256. For the text generator, we set the dimension of the prediction head to 768. For the image encoder, the size of latent feature maps were set to be $8 \times 8, 16 \times 16$ for top level and bottom level, respectively. For optimization, we adopt the AdamW optimizer [43] with $\beta_1 = 0.9, \beta_2 = 0.95$, and a weight decay of 0.05. We used a cosine learning rate decay with a peak learning rate of 1e-4. We used the warm-up strategy during the first 5% of the total number of steps and an initial learning rate of 1e-5. The pre-training process was running on four 80G NVIDIA A100 GPUs. We used the mix precision wuth Accelerate open-source library [21] to speed up training and save computational costs.

Details of downstream tasks. For uni-modal downstream tasks, we used the AdamW optimizer [43] with the learning rate set to 3e-6 and 3e-4 for the pre-trained model and task-specific layers, respectively. We conducted binary classification and multi-class classification for different datasets. For image-text retrieval tasks, we computed pairwise similarity to rank paired data by relevance. For zero-shot categorization we used images as queries and generated expert text prompts as in [26]. The text prompt with the highest score would be considered a predictive positive sample. We used images as the prompt for the medical report generation task to guide text generation. We trained two auto-regressive models in image generation, e.g. Pixelsnail [8], to model multi-modal priors. Then we sampled latent encodings and fed them into a hierarchical decoder to generate new images.

9. Datasets

MIMIC-CXR[30] This is the largest radiology dataset currently available, comprising chest X-ray images and corresponding reports from Beth Israel Deaconess Medical Center. It contains over 370,000 images from more than 65,000 patients, making it one of the most extensive collections of de-identified chest X-rays available for research. Each image is accompanied by detailed textual reports, which provide diagnostic information and contextual clinical notes. For the use of this dataset, We exclude samples that lack a "findings" or "impression" section within their clinical reports and retain only images in the frontal view. For dataset splitting, we utilize the official train-test splits provided by MIMIC-CXR. For downstream tasks, following the approach in [11], we sample the **MIMIC 5x200** subset and remove it from the training set to ensure robust evaluation.

CheXpert 5x200 The original CheXpert dataset's chest radiographs [28] are multi-labeled to accommodate numerous medical observations occurring at the same time. Because our zero-shot classification relies on identifying the most comparable target, having numerous alternative labels for a target can lead to results that are inconsistent across categories. As a result, following setting in [26, 69], we employ CheXpert's partial data to construct the CheXpert 5x200 dataset, which has 200 solely positive images for each of the CheXpert competition tasks: atelectasis, cardiomegaly, pneumonia, edema, and pleural effusion. In this dataset, each image has positive labels for only one condition.

RSNA Pneumonia[53] The RSNA Pneumonia Detection Challenge dataset, developed by the Radiological Society of North America (RSNA), is a large, annotated collection of chest X-ray images specifically labelled for pneumonia detection. We use the stage 2 version. This dataset contains 30k frontal view chest radiographs labeled either as "normal" or "Pneumonia". We sample raw 500 positives and 500 negatives for zero-shot classification. For fine-tuning,

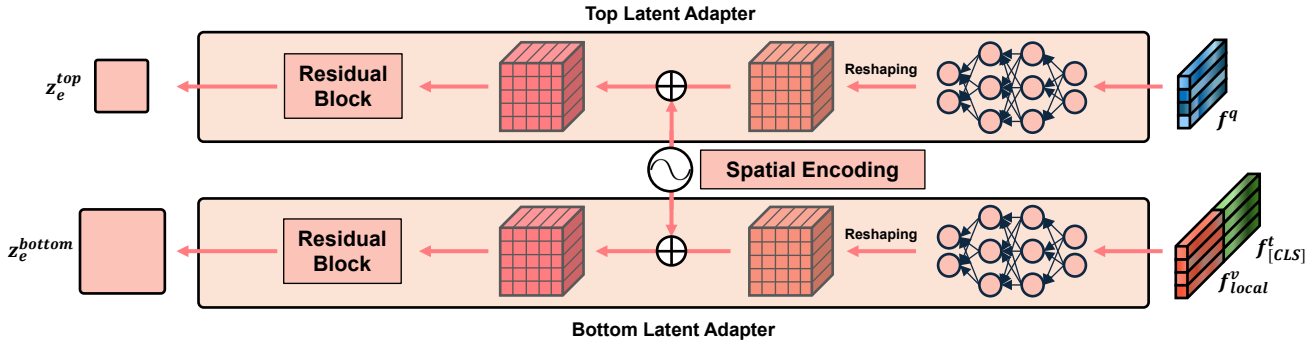

Figure 4. The detailed model structure of latent adapters. The learned embeddings are fed into top adapter as input while text representation concatenated with local preliminary visual embeddings are put into bottom adapters.

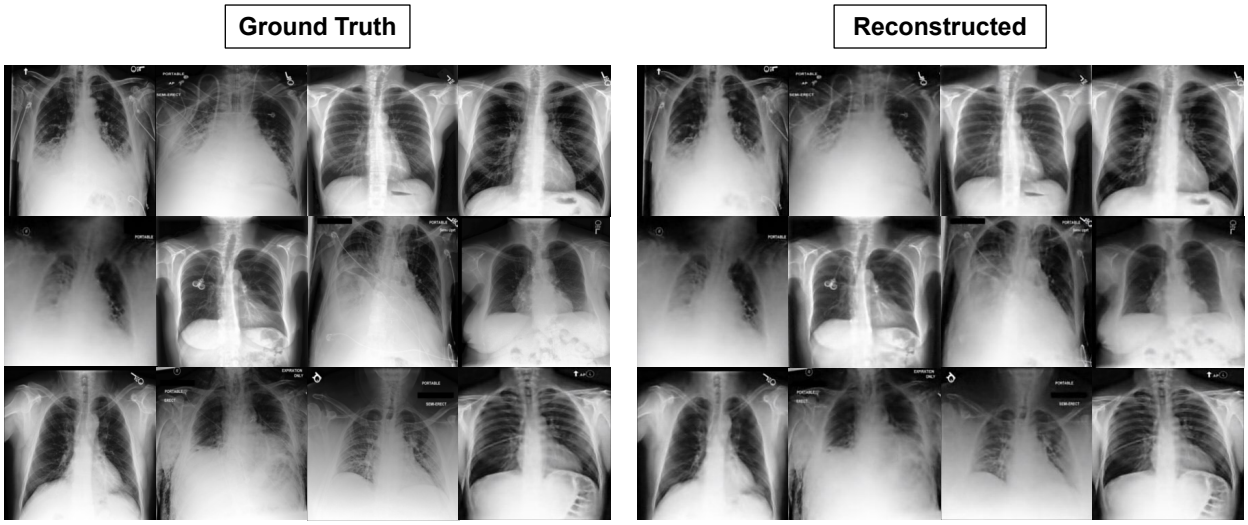

Figure 5. The additional comparison between real radiographs and reconstructed ones.

the train/valid/test split each constitutes 70%/15%/15% of the dataset, following [11].

COVIDx [59] It includes more than 30,000 CXR pictures from a global group of more than 16,600 patients. 16, 490 positive COVID-19 pictures from more than 2,800 patients are included in this collection. We make use of version 6 of this dataset. The task is to categorize each radiograph into three groups: normal, non-COVID pneumonia, and COVID-19. The data split follows [58].

SIIM-ACR Pneumothorax[66] The SIIM-ACR Pneumothorax Segmentation Challenge is a collaborative machine-learning competition organized by the Society for Imaging Informatics in Medicine (SIIM) and the American College of Radiology (ACR). SIIM-ACR Pneumothorax contains 12954 X-ray chest images, together with image-level pneumothorax annotation and pixel-level segmentation mask if pneumothorax exists. We use them for downstream supervised classification as in [11].

10. Additional visual reconstruction

We randomly select several reconstructed visual contents and compare them with real images (see Figure 5), which shows that our framework could capture visual details.

11. Detailed ablation study

We further conducted comprehensive ablation studies to evaluate the performance on in-/out-of-distribution datasets and various downstream tasks in Table 7. Results highlight the effectiveness of the TIG module for enhancing cross-modal alignment. Additionally, the various loss weights were determined through empirical testing, as illustrated in Table 7 ID 5, 6, 7. We examined the impact of varying the

ID	Learning objectives				Zero-shot cls	Fine-tuned cls		
	ITC	ITM	ITG	TIG	MIMIC 5x200 (ACC)	RSNA (AUC)	SIIM (AUC)	COVIDx (ACC)
1	✓				41.4	87.0	89.4	90.8
2	✓	✓			44.3	87.1	89.8	91.5
3	✓	✓	✓		44.8	88.1	92.3	92.8
4	✓	✓		✓	46.2	89.1	92.6	91.3
5 $	✓	✓	✓	✓	**50.4**	**91.7**	**94.8**	**93.5**
6 &	✓	✓	✓	✓	49.7	87.5	92.0	92.5
7 #	✓	✓	✓	✓	49.1	90.8	93.6	92.0

Table 7. Ablation studies of proposed components. $, & and # represent the weight of TIG loss set to 1.0, 0.8, 1.2, respectively.

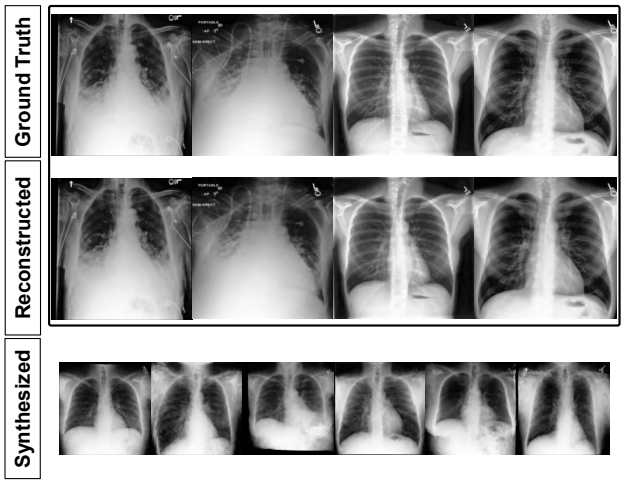

Figure 6. Visualization of generated images. Top: real radiographs. Middle: reconstructed images corresponding to the real samples. Bottom: generated radiographs through trained Pixelsnail models and VAE decoder.

TIG loss weight while maintaining the other losses' weights constant.