Table 7. Statistics of the tested datasets.

DATASET	IMAGE SIZE	DATASET SIZE	TASK	ANNOTATION
NIH CHESTX-RAY 14	224×224	112,120	CLS	14 CLASSES
VINDR-CXR	512×640	18,000	CLS	28 CLASSES, BBOXES
COVIDX CXR-4	1024×1024	84,818	CLS	2 CLASSES
SIIM-ACR PTX	512×512	12,047	CLS, SEG	2 CLASSES, MASKS
RSNA PNEUMONIA	1024×1024	26,684	CLS, SEG	BBOXES
IU-XRAY	512×640	3,955	RRG	IMAGE-REPORT PAIRS
OBJECT CXR	2048×2624	10,000	DET	BBOXES, ELLIPSE, POLYGONS
TBX11K	512×512	11,200	CLS, SEG	3 CLASSES, BBOXES
MIMIC 5X200	512×512	1,000	RET	IMAGE-REPORT PAIRS

A. Appendix

A.1. Tested Datasets

We evaluate MedVLMs on 9 public datasets across 4 tasks including classification (CLS), report generation (RRG), segmentation (SEG), and image-text retrieval (RET). Table 7 provided the statistics for the tested datasets. The detailed dataset information is as follows:

- NIH ChestX-ray (Wang et al., 2017) consists of 112,120 frontal-view CXRs with 14 disease labels from 30,805 unique patients. To make our results comparable with those reported by existing works, we follow (Zhou et al., 2022; 2023) to use the same training, validation, and test split corresponds to 70%, 10%, and 20% of the entire dataset, respectively.
- VinDr-CXR (Nguyen et al., 2022) contains more 18,000 CXRs collected from two major hospitals in Vietnam, where each image is annotated with both class labels and bounding boxes for 28 findings or diseases. We use the official data split with the training set of 15,000 images and the test set of 3,000 images, respectively. We further randomly selected 3,000 images from the training set to construct a validation set for parameter selection. Therefore, the final training, validation, and test sets contain 12,000, 3,000, and 3,000 samples, respectively.
- COVIDx-CXR4 (Wang et al., 2020) consists of 84,818 images from 45,342 subjects for COVID-19 detection, which is a binary classification task. We employ the official data split corresponds to 80%, 10%, and 10% of the entire dataset, respectively.
- SIIM-ACR Pneumothorax Segmentation (SIIM) (Zawacki et al., 2019) is designed to support the development of segmentation models for identifying pneumothorax in CXRs. SIIM contains 12,047 frontal-view CXRs with mask annotations of pneumothorax. Following (Huang et al., 2021), we adopt the same training, validation, and test split, where each constitutes 70%, 15%, and 15% of the entire dataset, respectively.
- RSNA Pneumonia (Shih et al., 2019) contains 26,684 images with mask annotations of pneumonia. We build the data split corresponds to 70%, 15%, and 15% of the entire dataset, respectively.
- IU X-RAY (Demner-Fushman et al., 2016) consists of 7,470 chest X-ray images and 3,955 reports. We follow (Chen et al., 2020b) to exclude the samples without reports and use the same training, validation, and test split corresponds to 70%, 10%, and 20% of the entire dataset, respectively.
- Object CXR (Healthcare, 2020) contains 10,000 frontal-view CXRs with annotations of foreign objects, where 5000 CXRs have foreign objects and the other 5000 CXRs have no foreign object. We use the official data split with the training, validation, and test sets consisting of 8,000, 1,000, and 1,000 images, respectively.
- TBX11K (Liu et al., 2020) consists of 11,200 X-rays with bounding box annotations for tuberculosis (TB) areas, where there are 5,000 healthy cases, 5,000 sick but non-TB cases, and 1,200 cases with manifestations of TB. We use the official data split with the training, validation, and test sets consisting of 6,600, 1,800, and 2,800 samples, respectively.

- We follow (Huang et al., 2021; Cheng et al., 2023) to construct MIMIC 5x200 to detect 5 diseases including Atelectasis, Cardiomegaly, Edema, Pleural, Effsion by randomly sampling 200 exclusive samples for each class from the MIMIC-CXR dataset.

A.2. Implementation Details

- Overall Setup: We run each experiments three times with different random seeds and report the average results. We check the performance on the validation set at each epoch and select the best checkpoint for the final evaluation. To ensure fair comparison, we use the standard grid search to select the best hyper-parameters and model configurations for each method based on the performance on the validation set. Due to the high computational costs, we only perform one run for the segmentation experiments, while our preliminary results show that the segmentation results are insensitive to the choice of random seeds.

- Classification: We follow most existing methods to add a linear classifier on top of the pre-trained image encoder, and fine-tune both the image encoder and the classifier on each dataset. We use the binary cross entropy loss for multi-label classification and the cross entropy loss for multi-class classification. We set the maximum training epoch to 200. For grid search, we explore a large search space of hyper-parameters by selecting the learning rate form $\{3\times10^{-2}, 1\times10^{-2}, 3\times10^{-3}, 1\times10^{-3}, 1\times10^{-4}, 5\times10^{-5}, 1\times10^{-5}\}$, the batch size from $\{32, 64, 128\}$, the optimizer from $\{\text{SGD}, \text{Adam}\}$, and whether custom refinements including Layer Normalization (LN) and Discriminative Learning Rates (DLR) discussed in Section 3.3 are applied or not.

- Segmentation: We adapt the UperNet architecture (Xiao et al., 2018) based on the implementation of the open-source mmsegmentation package (Contributors, 2020) and fine-tune UperNet with a frozen backbone from the pre-trained MedVLP image encoder. To incorporate the segmentation head, we only make minimal modifications by matching the dimensions of the pre-trained image encoder and the UperNet network for each method. Following the recommended settings of mmsegmentation, we use the cross entropy loss for training and SGD as the optimizer with a momentum of 0.9 and a polynomial decay schedule. We set the maximum number of training iterations to 20,000, the batch size to 32, and select the best learning rate from $\{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\}$ for each dataset.

- Report Generation: We adapt R2Gen (Chen et al., 2020b) as the task-specific head for report generation, with the image encoder frozen from a specified MedVLP model. We follow the same setting of R2Gen to train the model with the cross entropy loss and the Adam optimizer, and set the maximum training epoch to 100 and the batch size to 16. We select the best learning rate from $\{1 \times 10^{-2}, 1 \times 10^{-3}, 1 \times 10^{-4}\}$.

- Image-Text Retrieval: We follow the original setting of each CLIP-based method to get the image and text embeddings from the corresponding pre-trained models, and then compute the cosine similarity between a query image and all candidate reports to find the target reports.

A.3. Selected Hyper-Parameters

In this section, we provide the selected hyper-parameters per method and dataset.

- Classification: Tables 8, 9, 10, 11, 12 show the selected hyper-parameters per method and dataset.

- Segmentation: When the pre-trained image encoder is frozen, we find the hyper-parameters are consistent in terms of the MedVLP methods. As a result, we select $lr = 1 \times 10^{-4}$ for Object CXR, $lr = 1 \times 10^{-4}$ for RSNA, $lr = 1 \times 10^{-3}$ for SIIM, and $lr = 1 \times 10^{-3}$ TBX11K.

- Report Generation: Similar to the segmentation experiments, the hyper-parameters are consistent in terms of the MedVLP methods. We find $lr = 1 \times 10^{-3}$ is the best learning rate for the IU X-ray dataset.

Table 8. Selected hyper-parameters per method on the NIH dataset.

Method	Learning Rate	Batch Size	Optimizer	LN	DLR
ConVIRT	1×10^{-4}	64	Adam	Yes	Yes
GLoRIA	1×10^{-4}	64	Adam	Yes	Yes
MedCLIP-R50	1×10^{-5}	64	Adam	No	No
MedCLIP-ViT	1×10^{-5}	32	Adam	No	No
MedKLIP	1×10^{-4}	128	Adam	No	Yes
M-FLAG	1×10^{-4}	32	Adam	Yes	No
MGCA-R50	1×10^{-5}	32	Adam	Yes	No
MGCA-ViT	1×10^{-2}	64	SGD	Yes	Yes
MRM	3×10^{-2}	64	SGD	Yes	Yes
REFERS	3×10^{-2}	32	SGD	Yes	No

Table 9. Selected hyper-parameters per method on the VinDr dataset.

Method	Learning Rate	Batch Size	Optimizer	LN	DLR
ConVIRT	5×10^{-5}	32	Adam	Yes	Yes
GLoRIA	1×10^{-4}	64	Adam	Yes	Yes
MedCLIP-R50	1×10^{-4}	128	Adam	No	No
MedCLIP-ViT	1×10^{-4}	128	Adam	No	No
MedKLIP	1×10^{-4}	64	Adam	No	Yes
M-FLAG	1×10^{-4}	64	Adam	Yes	No
MGCA-R50	5×10^{-5}	64	Adam	Yes	No
MGCA-ViT	3×10^{-2}	64	SGD	Yes	Yes
MRM	1×10^{-2}	64	SGD	Yes	Yes
REFERS	3×10^{-2}	128	SGD	Yes	No

Table 10. Selected hyper-parameters per method on the COVIDx dataset.

Method	Learning Rate	Batch Size	Optimizer	LN	DLR
ConVIRT	1×10^{-4}	32	Adam	Yes	Yes
GLoRIA	1×10^{-4}	32	Adam	Yes	Yes
MedCLIP-R50	1×10^{-4}	128	Adam	No	No
MedCLIP-ViT	5×10^{-5}	128	Adam	No	No
MedKLIP	1×10^{-4}	128	Adam	No	Yes
M-FLAG	1×10^{-4}	128	Adam	Yes	No
MGCA-R50	5×10^{-5}	64	Adam	Yes	No
MGCA-ViT	1×10^{-4}	64	Adam	Yes	Yes
MRM	1×10^{-4}	64	Adam	Yes	Yes
REFERS	5×10^{-5}	64	Adam	Yes	No

Table 11. Selected hyper-parameters per method on the SIIM dataset.

Method	Learning Rate	Batch Size	Optimizer	LN	DLR
ConVIRT	1×10^{-4}	128	Adam	Yes	Yes
GLoRIA	1×10^{-5}	128	Adam	Yes	Yes
MedCLIP-R50	1×10^{-5}	128	Adam	No	No
MedCLIP-ViT	1×10^{-5}	32	Adam	No	No
MedKLIP	1×10^{-4}	64	Adam	No	Yes
M-FLAG	1×10^{-4}	64	Adam	Yes	No
MGCA-R50	1×10^{-5}	128	Adam	Yes	No
MGCA-ViT	1×10^{-2}	128	SGD	Yes	Yes
MRM	1×10^{-2}	64	SGD	Yes	Yes
REFERS	3×10^{-2}	64	SGD	Yes	No

Table 12. Selected hyper-parameters per method on the RSNA dataset.

Method	Learning Rate	Batch Size	Optimizer	LN	DLR
ConVIRT	5×10^{-5}	64	Adam	Yes	Yes
GLoRIA	1×10^{-4}	32	Adam	Yes	Yes
MedCLIP-R50	1×10^{-5}	32	Adam	No	No
MedCLIP-ViT	1×10^{-5}	32	Adam	No	No
MedKLIP	1×10^{-4}	128	Adam	No	Yes
M-FLAG	1×10^{-4}	64	Adam	Yes	No
MGCA-R50	1×10^{-5}	32	Adam	Yes	No
MGCA-ViT	1×10^{-2}	32	SGD	Yes	Yes
MRM	1×10^{-2}	32	SGD	Yes	Yes
REFERS	1×10^{-2}	32	SGD	Yes	No