Institute of High Performance Computing
IHPC

BenchX: A Unified Benchmark Framework for Medical Vision-Language Pre-Training on Chest X-Rays

Yang Zhou, Tan Li Hui Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu*, Rick Siow Mong Goh*
Senior Scientist
IHPC, A*STAR, Singapore

CREATING GROWTH, ENHANCING LIVES

Background

What is Medical Vision-Language Pre-Training (MedVLP)?

- MedVLP learns generalizable visual representations from both medical **images** and **reports**

Why MedVLP?

- **Rich** and **cross-modal** knowledge captured from medical images and text
- **Strong transferability** for a wide range of medical tasks
- **Core** of multimodal medical foundation models

Question: Which MedVLP?

CREATING GROWTH, ENHANCING LIVES

Challenges

Compared MedVLP Methods

Method	Pre-Train Data	Image Encoder	Text Encoder	Training Loss
ConVIRT	MIMIC-CXR	R50	ClinicalBERT	ITC
GLORIA	CheXpert	R50	ClinicalBERT	ITC
MedCLIP	CheXpert, MIMIC-CXR	R50/Swin-tiny	ClinicalBERT	SML
MedKLIP	MIMIC-CXR	4-Stage R50	ClinicalBERT	ITC, CE
M-FLAG	MIMIC-CXR	R50	CXR-BERT	RegL2
MGCA	MIMIC-CXR	R50/ViT-base	ClinicalBERT	ITC, CPA
MRM	MIMIC-CXR	ViT-base	Custom BERT	MIM, MLM
PTUnifier	ROCO, MediCaT, MIMIC-CXR	ViT-base	BioMed ROBERTa	ITC, MLM, ITM
REFERS	MIMIC-CXR	ViT-base	BERT	ITC, CLM

Challenges in Benchmarking MedVLP Methods

- *Inconsistent* Pre-Training Setup: Datasets, Train-Test Splits, ...
- *Incompatible* Fine-Tuning Protocol: Pre-processing, Training Strategies, Head, ...
- *Incomprehensive* Comparison: Limited Baselines and Tasks

Main Contributions

We *proposed* **BenchX**, a unified MedVLP benchmark framework on CXRs

- **Standardized Pre-Training Setup**
- **Unified Fine-Tuning Protocol**
- **Comprehensive Test Datasets and Tasks**

We *retrained* and *established* baselines for **9** MedVLP methods across **4** tasks

Goal: Address *Discrepancies* in Datasets, Pre-Training, and Fine-Tuning

Enable **Head-to-Head** Comparison and **Systematic** Analysis

CREATING GROWTH, ENHANCING LIVES

BenchX Design: Training and Test Data

Pre-Training Data

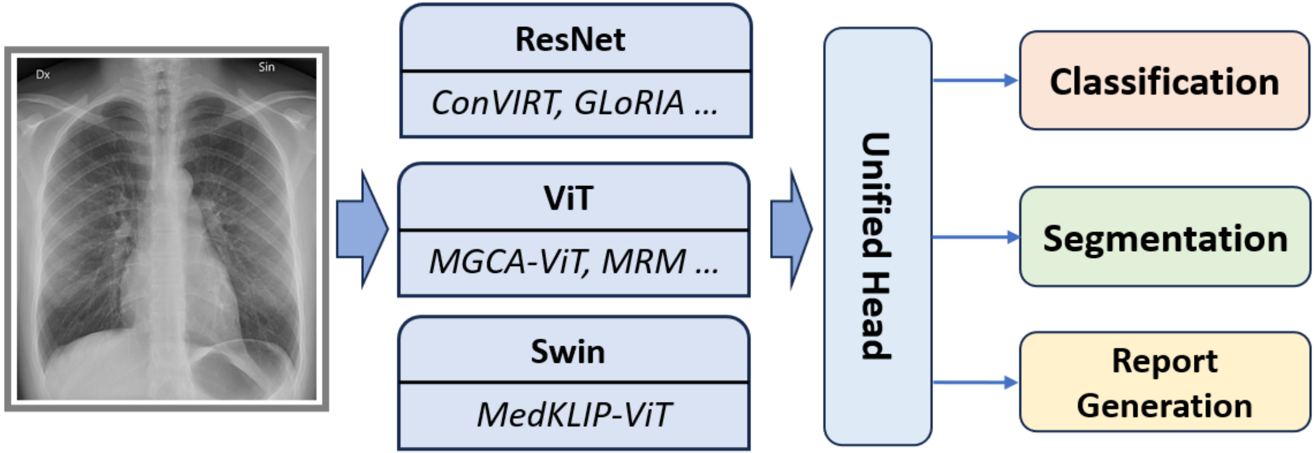
- MIMIC-CXR: ~ 220,000 frontal images with reports in the official training set
- Transform: Resize 256x256 → random crop 224x224

Fine-Tuning Data

- **4** Tasks: Classification, Segmentation, Report Generation, Image-Text Retrieval
- **9** Datasets from Diverse Resources
- **Consistent Preprocessing**: All scripts are provided

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

BenchX Design: Fine-Tuning Pipeline



Flexible Architectures

- ResNet, ViT, Swin, and more

Compatible Task-Specific Heads

- Classification: Linear Classifier
- Segmentation: UperNet
- Report Generation: R2Gen

✓ **Training or testing in *one line***

```
# Training
python bin/train.py train.yml

# Testing
python bin/test.py test.yml
```

Summary of Experimental Results and Key Findings

Table 1. Overall performance (%) across different tasks (**Best**, Second Best)

Method	M-CLS (AUC)↑	B-CLS (F1)↑	SEG (mDice)↑	RRG (BLEU4)↑	Avg. Rank↓
ConVIRT	85.37	65.56	78.89	14.8	6.38
GLoRIA	84.68	64.06	77.05	**17.0**	5.88
MedCLIP-R50	83.02	67.17	79.80	16.3	5.25
MedCLIP-ViT	84.00	**68.33**	78.76	15.1	5.75
MedKLIP	82.77	65.56	79.42	16.7	6.13
M-FLAG	77.73	62.96	72.77	14.7	10.00
MGCA-R50	83.47	64.69	79.85	15.9	6.50
MGCA-ViT	86.10	67.03	80.32	**17.0**	2.38
MRM	**86.18**	67.72	**80.66**	16.5	**2.00**
REFERS	84.65	66.06	79.93	16.1	4.75

Key Findings

- *Performance Leadership*: **MRM** and **MGCA-ViT** consistently outperform others
- *Progress Assessment*: Some recent methods show **less improvement** than initially reported
- *Unexpected Strength of ConVIRT*: Properly trained **earlier** MedVLP methods could perform **comparably or better** than more recent approaches

Conclusion

BenchX Framework

- **Broad Coverage**
  - **Nine** Datasets & **Four** Medical Tasks
- **Fair and Transparent Comparison**
  - **Standardized** Benchmark Suites
  - **Unified** Finetuning Protocols
- **Good Extensibility**
  - Supports **Diverse** Model Architectures
  - Easily **Adaptable** to New Models
  - Facilitates **New Dataset** Integration

It is time to reassess prior advancements in MedVLP

CREATING GROWTH, ENHANCING LIVES

Institute of High Performance Computing
IHPC

Thank You

Code and Models are Available