Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing

Benedikt Boecking*, Naoto Usuyama*, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, Ozan Oktay†

Microsoft Health Futures

ECCV TEL AVIV 2022

Microsoft

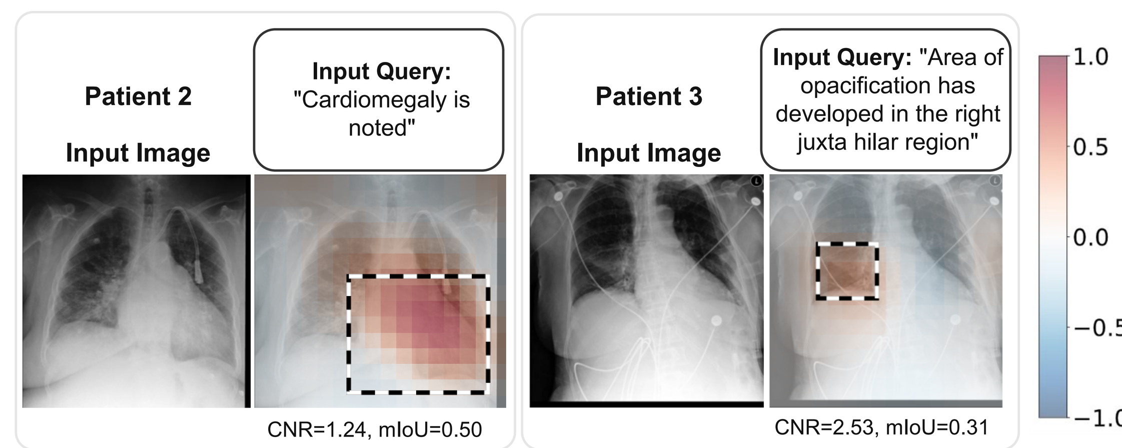
Highlights

- A new chest X-ray (CXR) domain-specific **language model**, CXR-BERT (Fig. 1), available on HuggingFace: `https://aka.ms/biovil-models`
- A **self-supervised Vision-Language Processing (VLP)** approach for paired biomedical data (BioViL, Fig.2). `https://aka.ms/biovil-code`
- `MS-CXR`: a **phrase grounding dataset** for chest X-ray data, released on PhysioNet: `https://aka.ms/ms-cxr`

Motivation

- **Clinical motivation**: Growing backlogs of medical image reporting puts pressure on radiologists and leads to errors and omissions.
- **Scalability** ML models require a vast number of manual annotations (experts' time is precious). Existing models are often limited to a fixed set of abnormalities or body-part.
- **Domain-specific challenges**: Lack of foundation models suitable for health data, smaller scale datasets, domain specific-language.

MS-CXR Phrase Grounding Dataset



`MS-CXR` allows fine-grained evaluation of joint text-image understanding in a biomedical domain.

- 1162 image bounding-box & sentence pairs,
- covering 8 different clinical findings,
- manually annotated and curated by radiologists.

Approach

- **CXR-BERT** is specialised to chest X-ray reports via masked language modelling (MLM), domain-specific vocabulary, contrastive learning and augmentations (sentence shuffle) **(Fig. 1)**.
- **BioViL** is a self-supervised VLP approach that uses the domain specific CXR-BERT as a text encoder, maintains an MLM loss, and utilises a global/local contrastive loss to match image-report pairs **(Fig. 2)**.

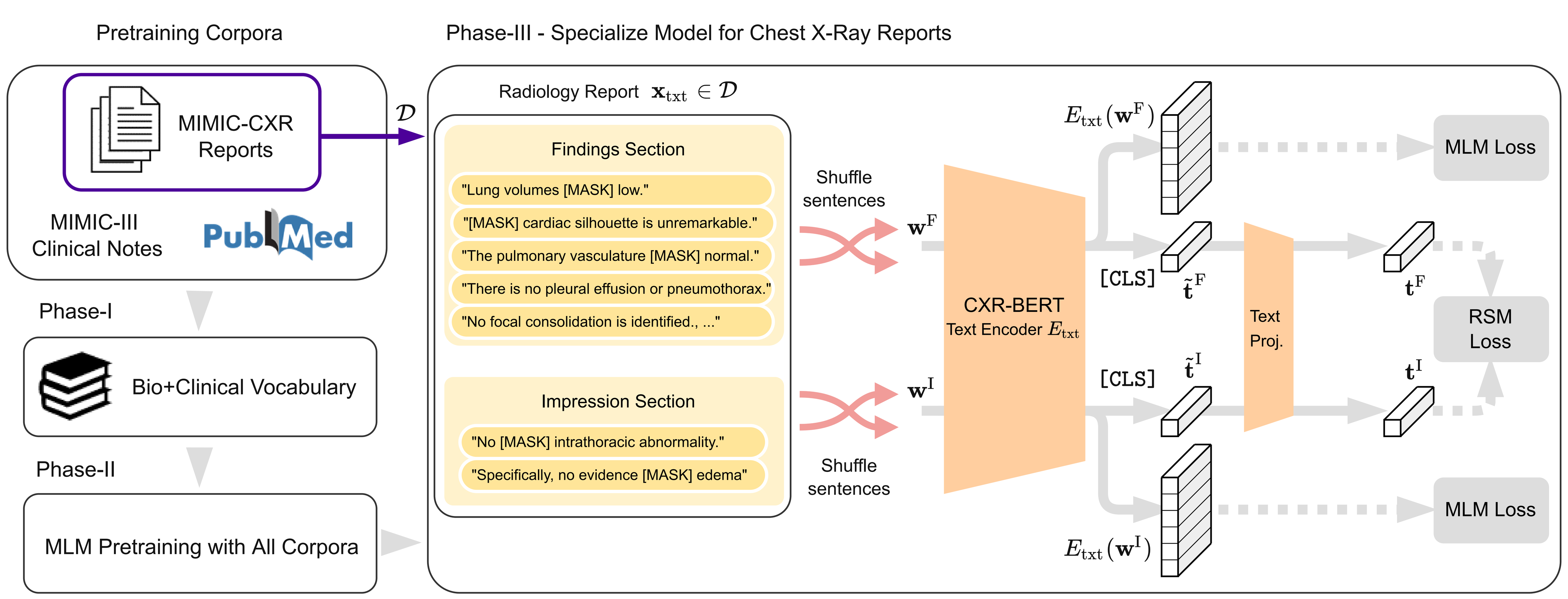


Figure 1: The proposed CXR-BERT text encoder has three phases of pretraining and uses a domain-specific vocabulary, masked language modelling (MLM) and radiology section matching (RSM) losses, regularisation, and text augmentations.

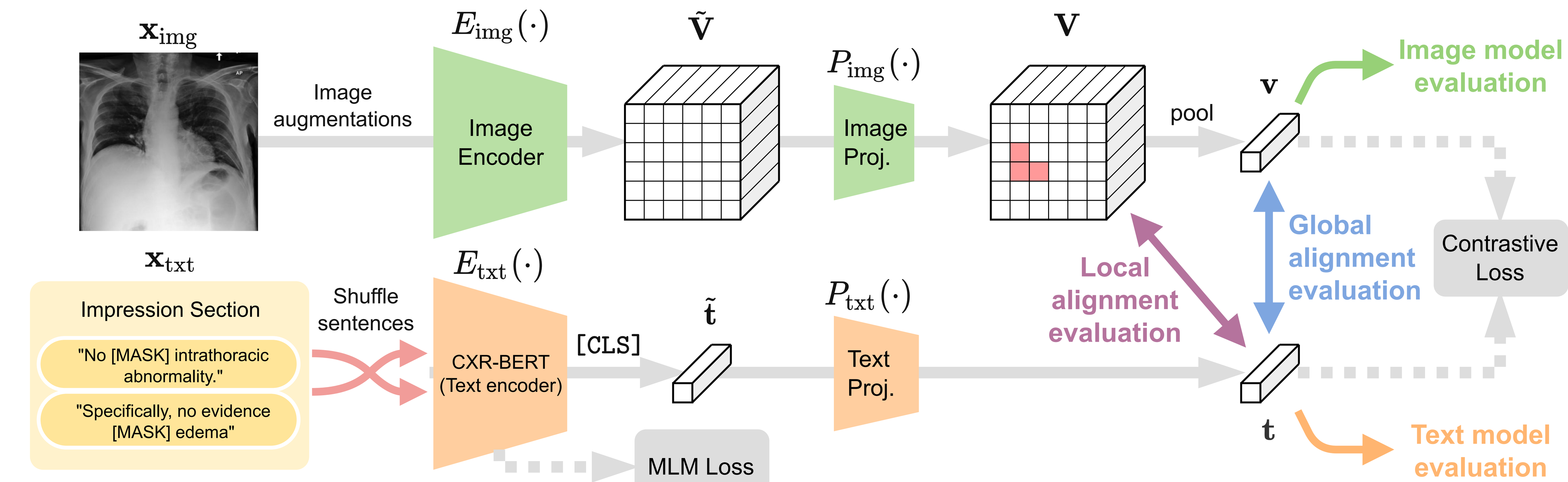


Figure 2: BioViL leverages our radiology-specific text encoder (CXR-BERT), text augmentation, regularisation, and maintains language model quality via a masked language modelling (MLM) loss.

Experiments Preview

We conduct a broad evaluation including zero-shot classification, phrase grounding, and natural language inference (NLI). **Data:** MIMIC-CXR v2 [2] chest radiograph dataset. After processing we have 146.7k training and 22.2k validation samples. Downstream evaluation samples are kept in a held-out test set.

Table 1: Text encoder evaluation: radiology domain **natural language inference**, fine-tuned and averaged over 5 runs.

	RadNLI accuracy
RadNLI baseline	53.30
ClinicalBERT	47.67
PubMedBERT	57.71
CXR-BERT (after Phase-III)	60.46
CXR-BERT (Phase-III + Joint Training)	65.21

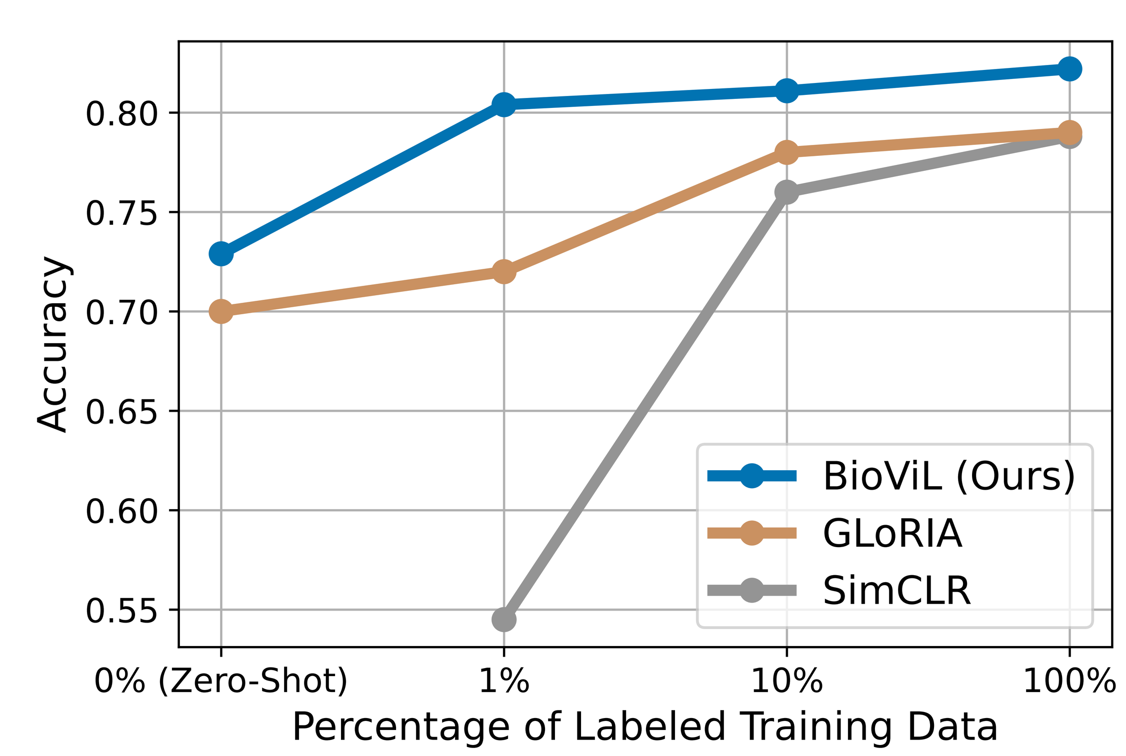


Figure 3: Pneumonia **classification**, zero-shot and fine-tuned.

Table 2: **Zero-shot phrase grounding** results on our **MS-CXR Benchmark**. Contrast-to-Noise Ratio (CNR) and Intersection over Union (mIoU) averaged over all findings.

Method	Contrastive Obj.	CNR	mIoU
Baseline (w/ ClinicalBERT)	global	0.76	.224
Baseline (w/ PubMedBERT)	global	0.77	.225
GLoRIA [1]	global & local	0.93	.246
BioViL	global	1.02	.266
BioViL-L	global & local	1.14	.284

References

[1] S.-C. Huang, L. Shen, M. P. Lungren, and S. 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*, pages 3942–3951, 2021.

[2] A. Johnson, T. Pollard, S. Berkowitz, R. Mark, and S. Horng. MIMIC-CXR database (v2). PhysioNet, 2019.

*The authors contributed equally.
Correspondence: † ozan.oktay@microsoft.com