Multimodal visual-language pre-training in radiology

Bartosz Kochański
04.11.2024

Agenda

- **Introduction** (inspiration, context, rationale)
- **Prior work** (Microsoft papers '22, '23, MIMIC-CXR database)
- **ICML 2024 conference paper** (Yang et al. 2024)

Focus on broad picture

Inspiration

Unlocking the Power of Spatial and Temporal Information in Medical Multimodal Pre-training

Jinxia Yang [1] Bing Su [1,2] Wayne Xin Zhao [1,2] Ji-Rong Wen [1,2,3]

1 Gaoling School of Artificial Intelligence, Renmin University of China
2 Beijing Key Laboratory of Big Data Management and Analysis Methods
3 School of Information, Renmin University of China.

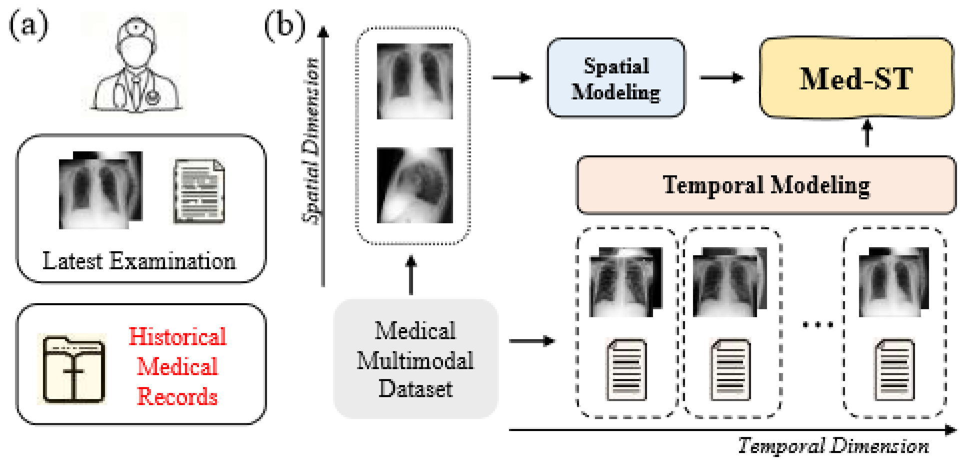


Figure 1: The motivation and our framework: (a) The practice of physicians in a clinical setting. (b) We propose the Med-ST framework, which explicitly designs spatial and temporal modeling by mining spatio-temporal supervision signals from the dataset.

Inspiration

Unlocking the Power of Spatial and Temporal Information in Medical Multimodal Pre-training

Jinxia Yang [1] Bing Su [1,2] Wayne Xin Zhao [1,2] Ji-Rong Wen [1,2,3]

1 Gaoling School of Artificial Intelligence, Renmin University of China
2 Beijing Key Laboratory of Big Data Management and Analysis Methods
3 School of Information, Renmin University of China.

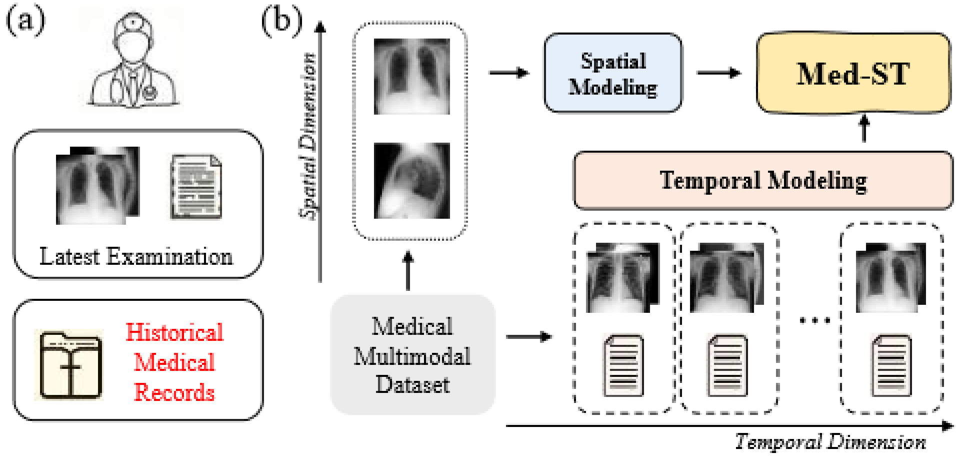


Figure 1: The motivation and our framework: (a) The practice of physicians in a clinical setting. (b) We propose the Med-ST framework, which explicitly designs spatial and temporal modeling by mining spatio-temporal supervision signals from the dataset.

Authors

Unlocking the Power of Spatial and Temporal Information in Medical Multimodal Pre-training

---

Jinxia Yang [1] **Bing Su** [1,2] **Wayne Xin Zhao** [1,2] **Ji-Rong Wen** [1,2,3]

1 Gaoling School of Artificial Intelligence, Renmin University of China
2 Beijing Key Laboratory of Big Data Management and Analysis Methods
3 School of Information, Renmin University of China.

MSc Student

Assoc. Prof.
Advisor - CV
H: 18

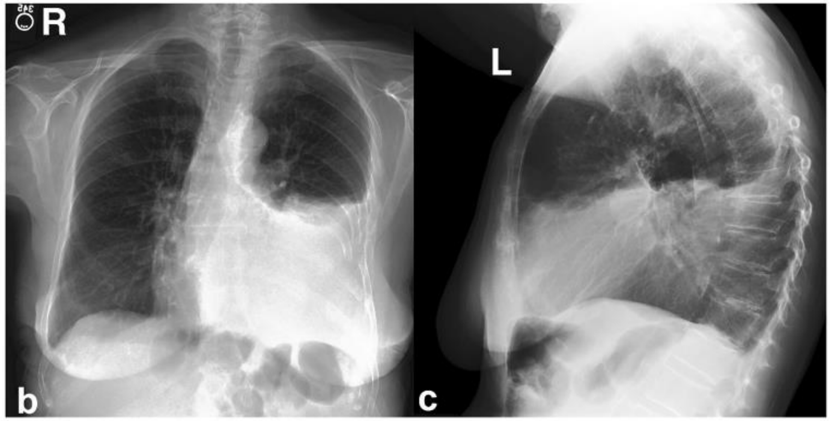
Assoc. Prof.
Advisor - NLP
H: 66

Full Professor

H: 91

Rationale: common image annotations in **ML**

The image



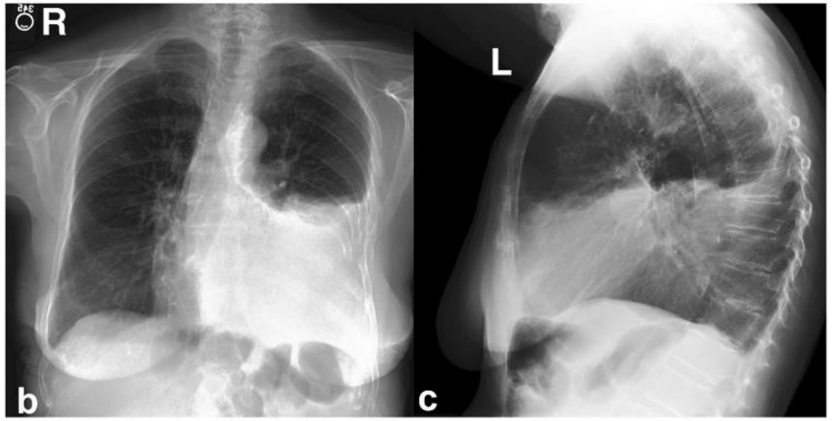
The annotation

id	edema	pl_eff	pneumo	pn_thorx
Sub_001	0	0	1	0

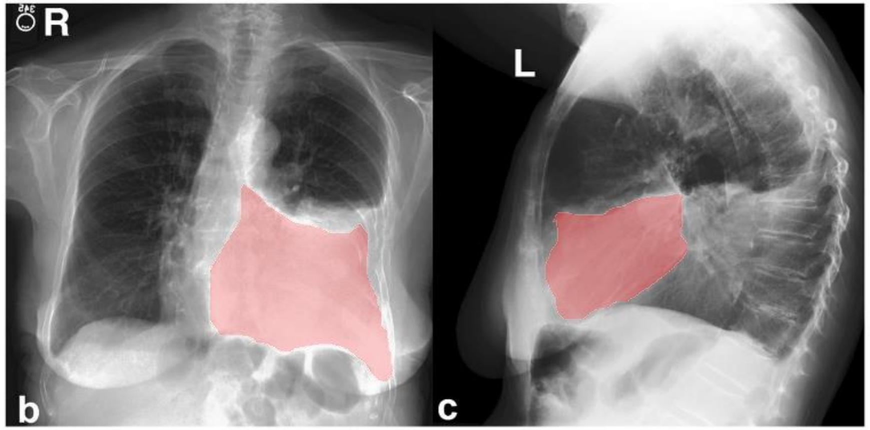
Johnson et al. 2019: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports

Rationale: common image annotations in **ML**

The image



The annotation



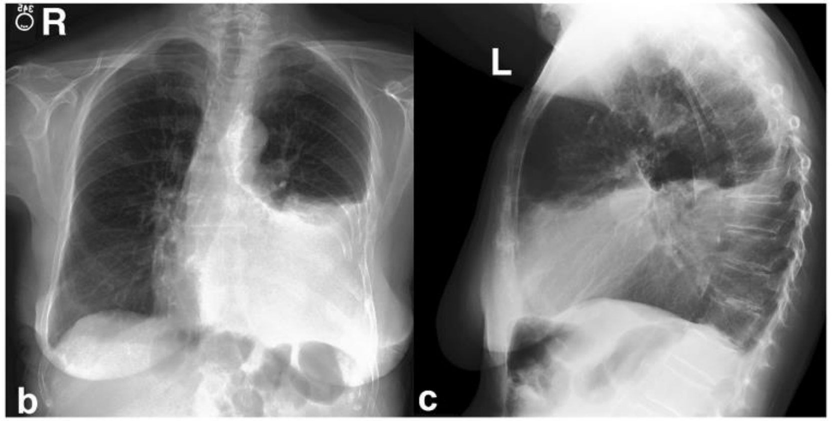
Johnson et al. 2019: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports

Rationale: common image annotations in **ML**

- **Task-specific**
- **Require a lot of radiology expert work-hours**

Rationale: common image annotations in **radiology**

The image



The annotation

```
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE:  Chest PA and lateral

COMPARISON:  ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

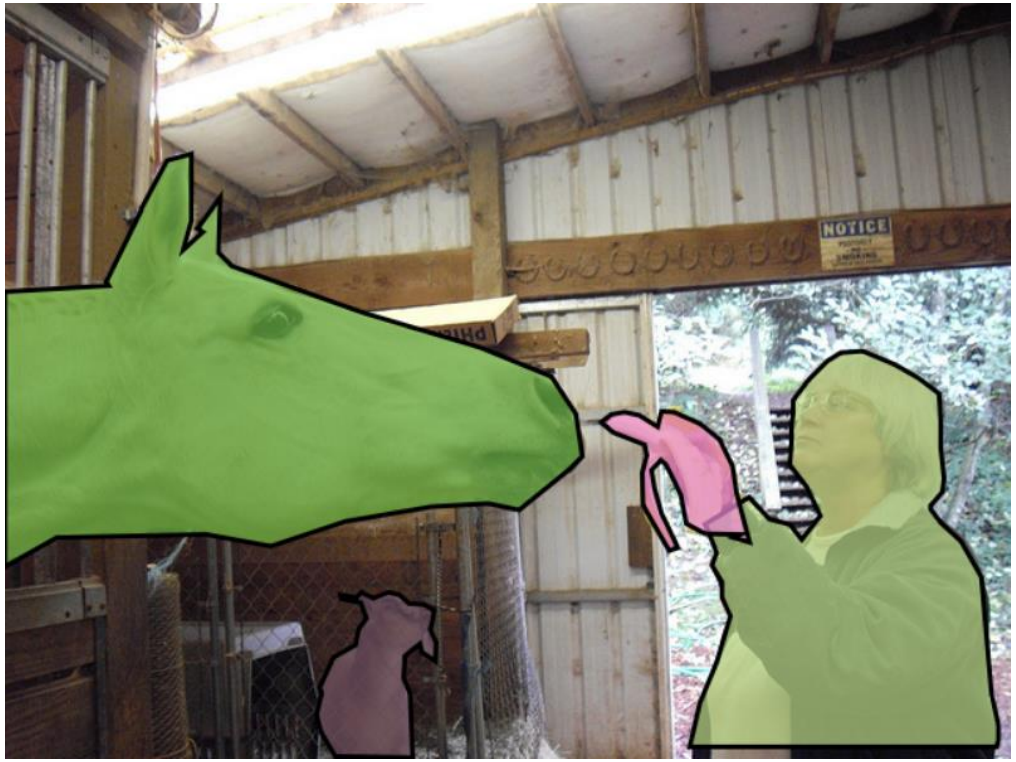
Johnson et al. 2019: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports

Rationale: common image annotations in **radiology**

- **Utilization of already-made annotation (anonymization needed)**
- **Possibility to train more general models**

Free-text annotations: what are we accustomed to

Microsoft COCO: Common Objects in Context
Lin et al. 2014



a horse eating a banana out of an older woman's hand.
a woman is feeding a banana to a horse
a woman feeds a banana to a horse.
an old woman feeding a horse a banana.
a woman is feeding a banana to a horse.

https://cocodataset.org/#explore

Radiology reports: domain-specific challenges

a horse eating a banana out of an older woman's hand.

vs.

```
FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

Negations:
"There is no evidence of pneumonia in the left lung"
vs. "There is no dog in this picture"

Long-range dependency:
"There are no new areas of consolidation to suggest the presence of pneumonia"

Text unrelated to scans:
"Preliminary findings were discussed with Dr. _ at the time of patient's admission"

Boecking et al. 2022 Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing

The data: MIMIC dataset and derivatives

Beth Israel Deaconess Medical Center

Massachusetts Institute of Technology

Prof. Roger Greenwod Mark

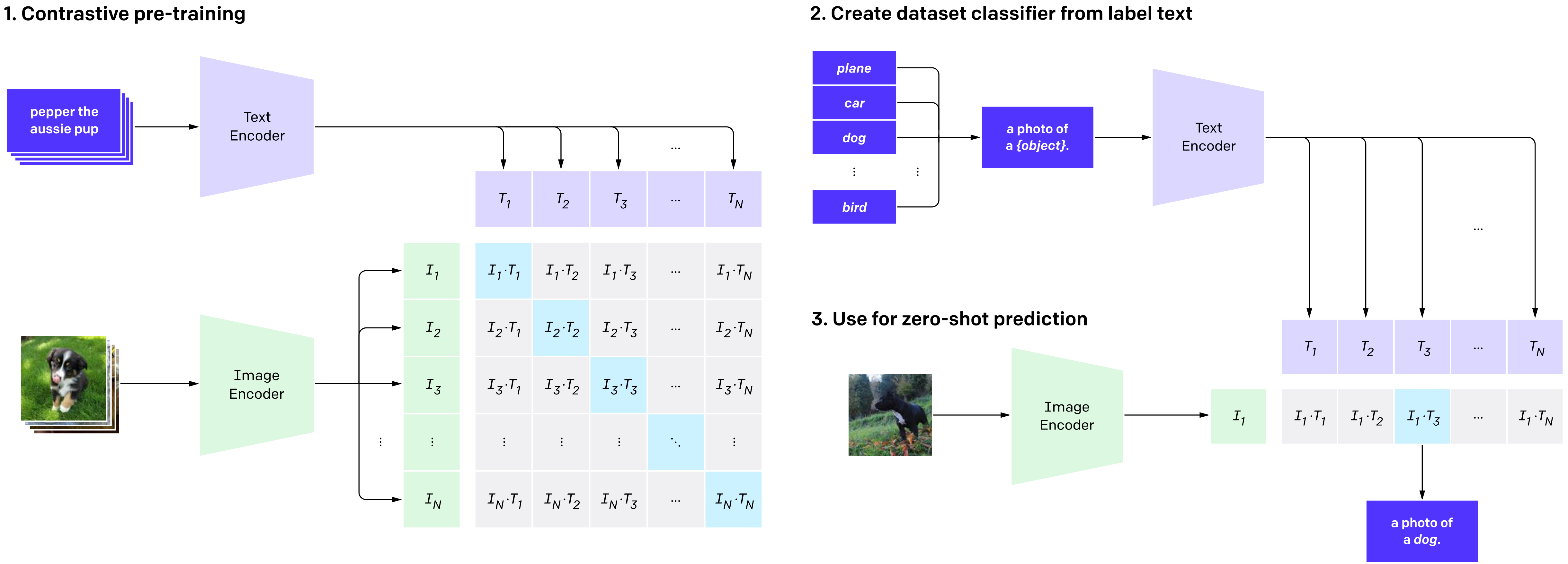
Microsoft

MIMIC 2001-now	MIMIC-CXR 2019	MS-CXR 2022	MS-CXR-T 2023
n = 223 452	n = 227 835	n = 1 162	n = 1 326

MIMIC = Multi-parameter Intelligent Monitoring in Intensive Care

The prior work: OpenAI CLIP (2021)

CLIP (*Contrastive Language–Image Pre-training*) builds on: zero-shot transfer, natural language supervision and multimodal learning.



CLIP pre-trains an image encoder and a text encoder to predict which images were paired with which texts in our dataset. We then use this behavior to turn CLIP into a zero-shot classifier. We convert all of a dataset's classes into captions such as "a photo of a dog" and predict the class of the caption CLIP estimates best pairs with a given image.

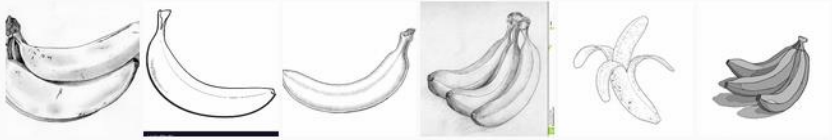
OpenAI CLIP

CLIP (*Contrastive Language–Image Pre-training*) builds on:

- zero-shot transfer,
- natural language supervision
- multimodal learning.

https://openai.com/index/clip/

Radford et al. 2021, Learning Transferable Visual Models From Natural Language Supervision

Dataset		ImageNet ResNet101	CLIP ViT-L
ImageNet		76.2%	76.2%
ImageNet V2		64.3%	70.1%
ImageNet Rendition		37.7%	88.9%
ObjectNet		32.6%	72.3%
ImageNet Sketch		25.2%	60.2%
ImageNet Adversarial		2.7%	77.1%

The prior work: MS-CXR (2022)

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, and Ozan Oktay‡

Microsoft Health Futures

TEL AVIV 2022

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`

https://www.microsoft.com/en-us/research/publication/making-the-most-of-text-semantics-to-improve-biomedical-vision-language-processing/

The prior work: MS-CXR (2022)

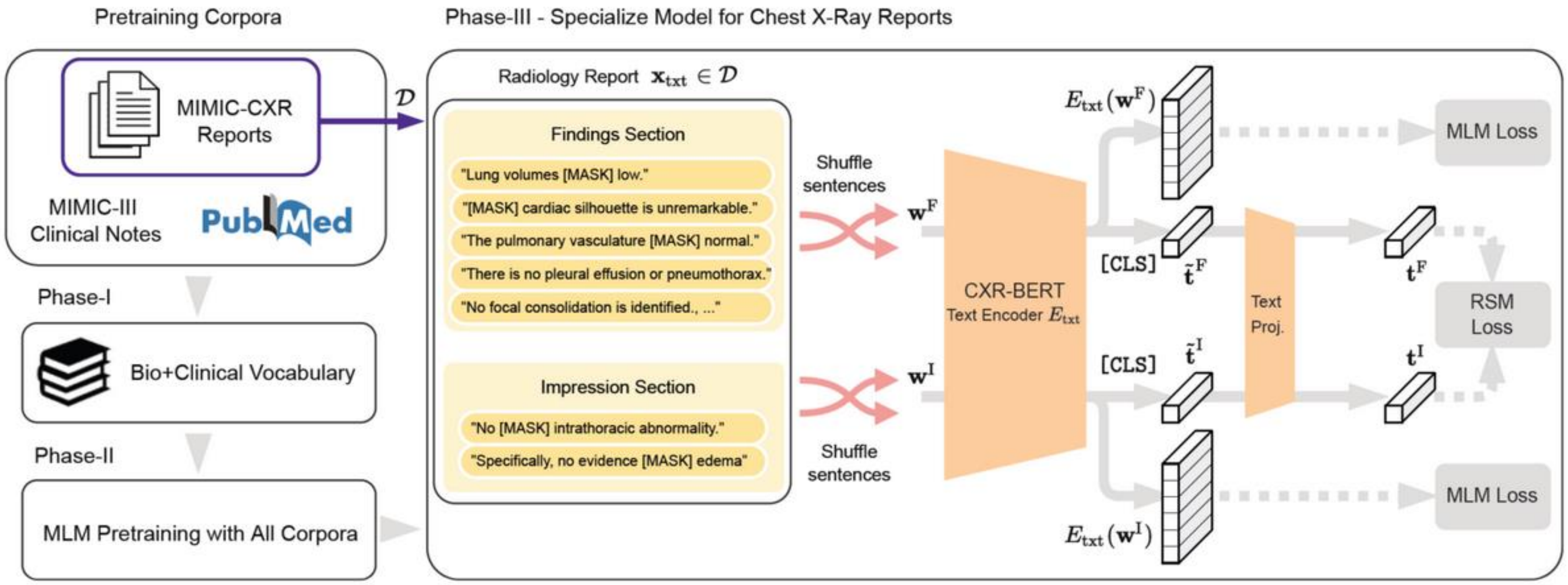


Figure 1: The proposed CXR-BERT text encoder has three phases of pretraining: (I) Pre-training with biomedical corpora (e.g., PubMed Abstracts, MIMIC-III clinical notes), (II) building a biomedical/clinical vocabulary, and (III) further specialising to chest radiology domain by performing contrastive learning between radiology reports and leveraging text augmentations.

The prior work: MS-CXR (2022)

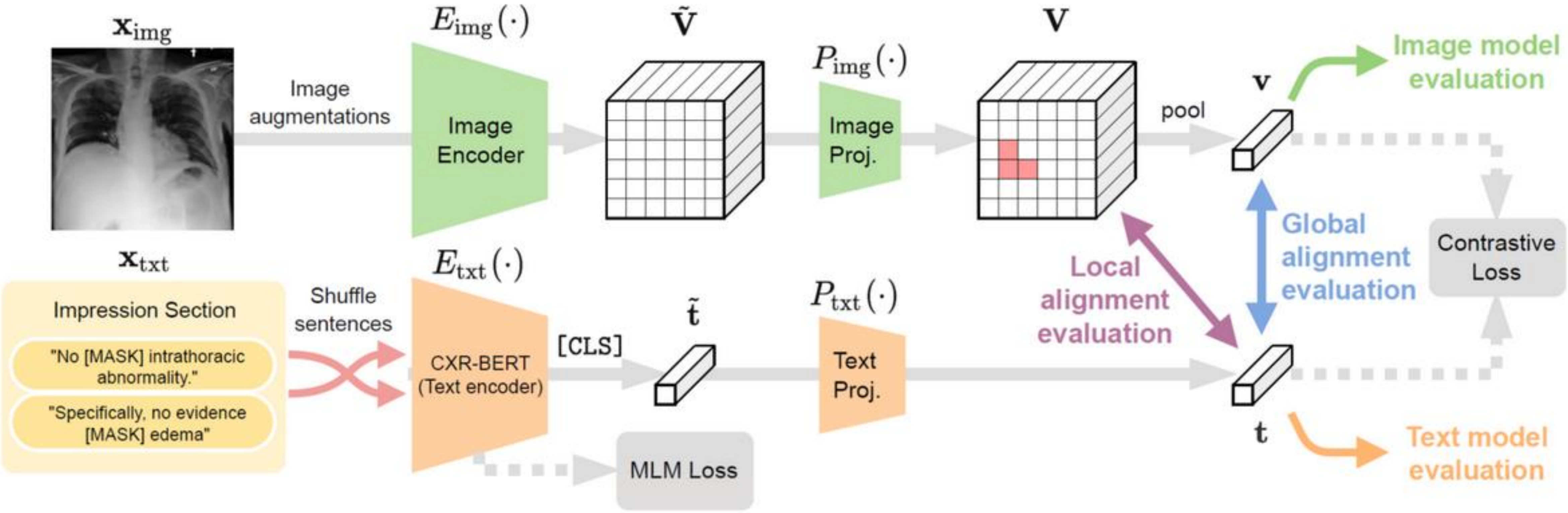


Figure 2: BioViL leverages our radiology-specific text encoder (CXR-BERT) in a multi-modal contrastive learning framework to train image and text encoders that can be aligned in the joint latent space. The proposed learning framework can be coupled with local-contrastive objectives as well.

The prior work: MS-CXR (2022)

Table 2: Comparing evaluations conducted in recent CXR image-text alignment studies.

Downstream task	Used in ref.*	Image encoder	Text encoder	Phrase reasoning	Findings localisation	Latent alignment	Annotation availability
Natural language inference	[B]	-	✓	✓	-	-	Scarce
Phrase grounding	[B]	✓	✓	✓	✓	✓	Scarce
Image classification	[B,C,G,L,M]	✓	-	-	-	-	High
Zero-shot image classif.	[B,G]	✓	✓	-	-	✓	Moderate
Dense image prediction (e.g. segmentation)	[B,G,L]	✓	-	-	✓	-	High
Global image–text retrieval	[C,G]	✓	✓	-	-	✓	High

*B, BioViL (Proposed); C, ConVIRT [85]; G, GLoRIA [31]; L, LoVT [56]; M, Local MI [45].

The prior work: MS-CXR (2022)

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

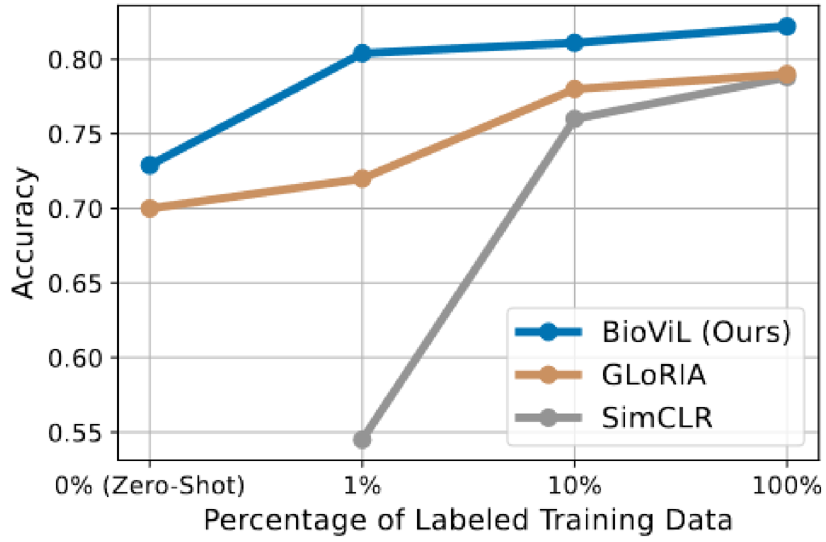


Figure 3: Pneumonia **classification**, zero-shot and fine-tuned.

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

The prior works: MS-CXR-T (2023)

Learning to Exploit Temporal Structure for Biomedical Vision–Language Processing

Shruthi Bannur*, Stephanie Hyland*, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori Javier Alvarez-Valle, and Ozan Oktay†

Microsoft Health Futures

JUNE 18-22, 2023
CVPR
VANCOUVER, CANADA

https://www.microsoft.com/en-us/research/publication/learning-to-exploit-temporal-structure-for-biomedical-vision-language-processing/

The prior works: MS-CXR-T (2023)

~40% of reports in MIMIC-CXR explicitly reference a prior image

EXAMINATION: CHEST (PA AND LAT)

INDICATION: ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE: Chest PA and lateral

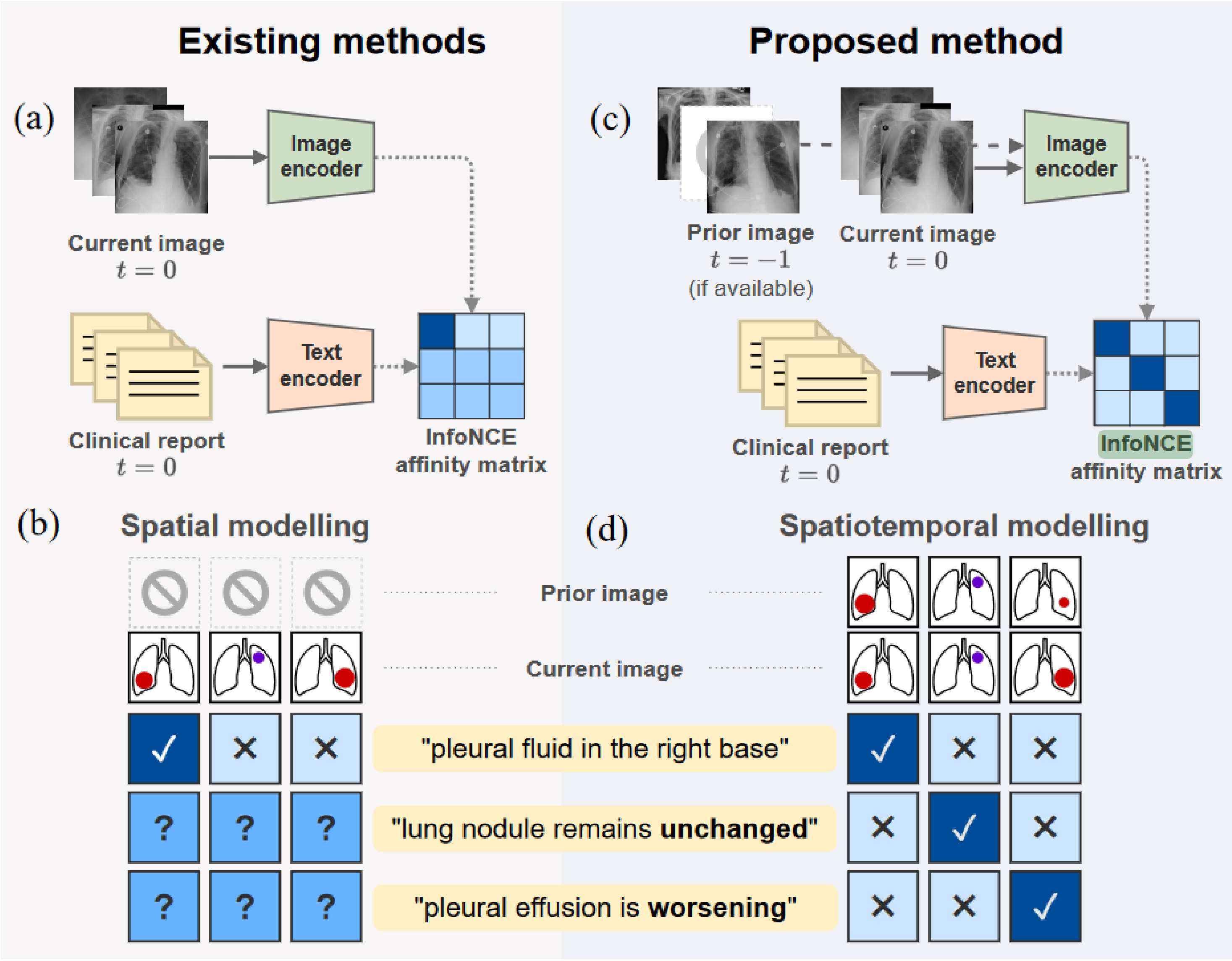
COMPARISON: ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new. Small right effusion is new. The upper lungs are clear. Right lower lobe opacities are better seen in prior CT. There is no pneumothorax. There are mild degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion



The prior works: MS-CXR-T (2023)

Multi-image encoder

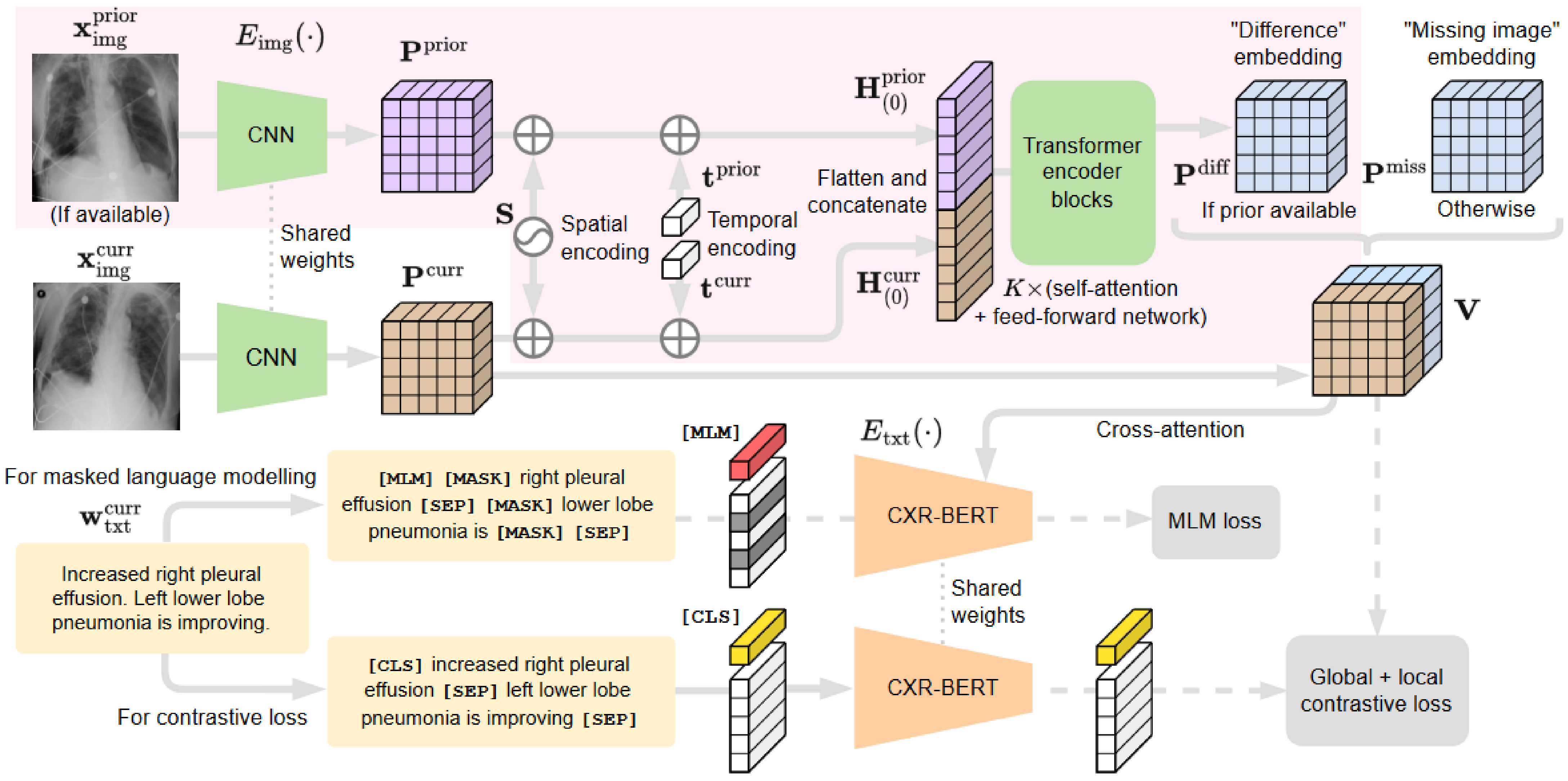
Requirements:

- A **flexible** encoder which can handle prior images **if available**, otherwise single images
- Avoid need for **registration:** it is not well defined between chest X-rays!

Approach: Hybrid **CNN-ViT** (vision transformer)

- CNN extracts patch-level features
- ViT integrates prior image information

The prior works: MS-CXR-T (2023)



The prior works: MS-CXR-T (2023)

Table 4. Image classification results on RSNA Pneumonia Detection Benchmark [60] for train and test splits of 70% – 30% respectively.

Method	% of Labels	Supervision	Acc.	F1	AUROC
GLoRIA [32]	✗	Zero-shot	0.70	0.58	-
BioViL [9]	✗	Zero-shot	0.732	0.665	0.831
BioViL-T	✗	Zero-shot	**0.805**	**0.706**	**0.871**
BioViL [9]	1%	Few-shot	0.805	0.723	0.881
BioViL-T	1%	Few-shot	**0.814**	**0.730**	**0.890**

Table 5. Results on *MS-CXR* benchmark [10] (5-runs with different seeds), "Multi-image" column indicates the input images used at test time.

Method	Multi-Image	Avg. CNR	Avg. mIoU
BioViL [9]	✗	1.07 ± 0.04	0.229 ± 0.005
+ Local loss [9, 32]	✗	1.21 ± 0.05	0.202 ± 0.010
BioViL-T	✗	**1.33 ± 0.04**	**0.243 ± 0.005**
BioViL-T	✓	**1.32 ± 0.04**	**0.240 ± 0.005**

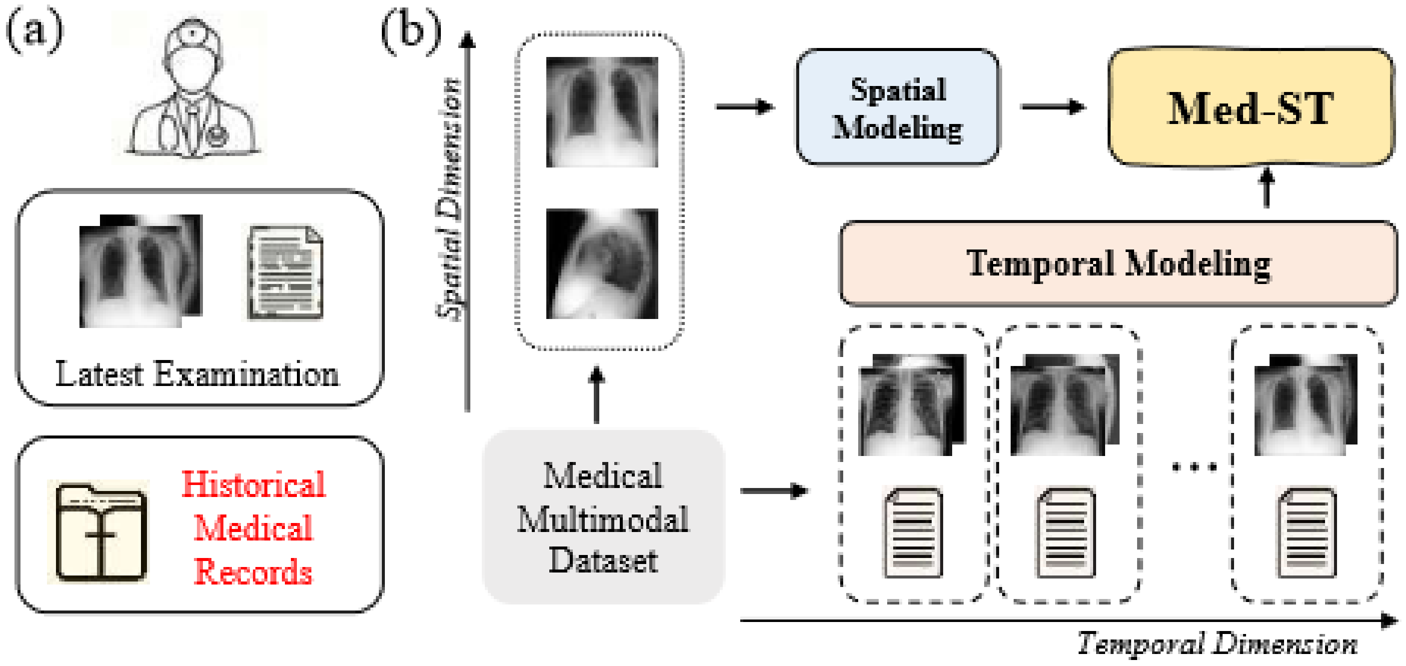
Table 6. Results on *MS-CXR-T* sentence similarity benchmark.

	MS-CXR-T (361 pairs)		**RadNLI** (145 pairs)	
Text Model	**Accuracy**	**ROC-AUC**	**Accuracy**	**ROC-AUC**
PubMedBERT [29]	60.39	.542	81.38	.727
CXR-BERT-G [9]	62.60	.601	87.59	.902
CXR-BERT-S [9]	78.12	.837	89.66	.932
BioViL-T	**87.77** ± 0.5	**.933** ± **.003**	90.52 ± 1.0	**.947** ± **.003**

The latest work

Unlocking the Power of Spatial and Temporal Information in Medical Multimodal Pre-training

Jinxia Yang [1] Bing Su [1 2] Wayne Xin Zhao [1 2] Ji-Rong Wen [1 2 3]



Contributions

The contributions of our study are outlined as follows:

- We thoroughly explore the information in medical multimodal datasets without additional manual labeling. Beyond text-image pairing, we **leverage multi-view spatial data and historical temporal data**, yielding a richer set of supervision signals.
- Our spatial modeling utilizes the **MoVE architecture** to tackle both frontal and lateral views with specialized experts, and introduces **modality-weighted local alignment** to establish fine-grained contrastive learning between spatial image regions and semantic tokens.
- For temporal modeling, we propose a novel **crossmodal bidirectional cycle consistency objective** that progresses from simple to complex. By forward mapping classification and reverse mapping regression, our model becomes capable of perceiving the context of sequences.
- We evaluate the performance of our method in temporal tasks and medical image classification tasks. The results demonstrate the effectiveness of our method.

Framework

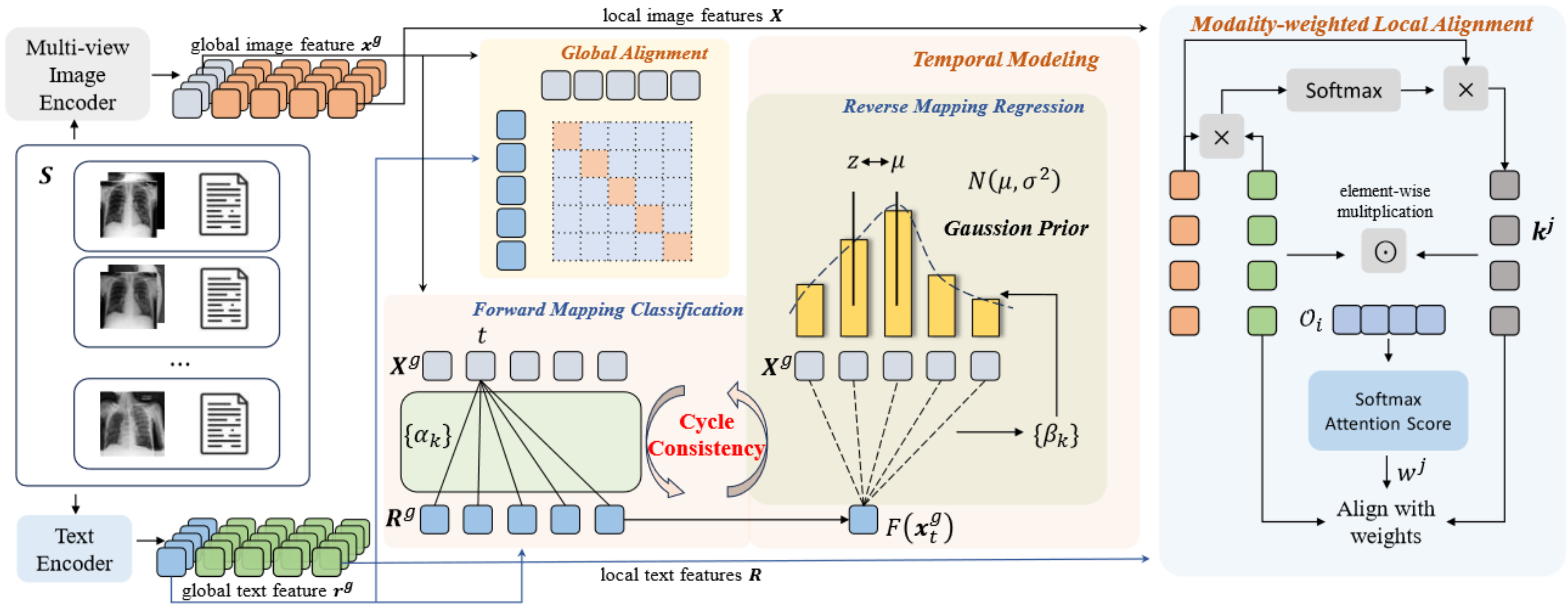


Figure 2: The framework of Med-ST. For spatial modeling, Med-ST extracts integrated multi-view visual representations, performs global alignment between paired global features, and introduces modality-weighted local alignment between paired local features. For temporal modeling, global image and text features in all pairs form two sequences respectively, and Med-ST imposes cross-modality bidirectional cycle consistency constraints between them.

MoVE architecture

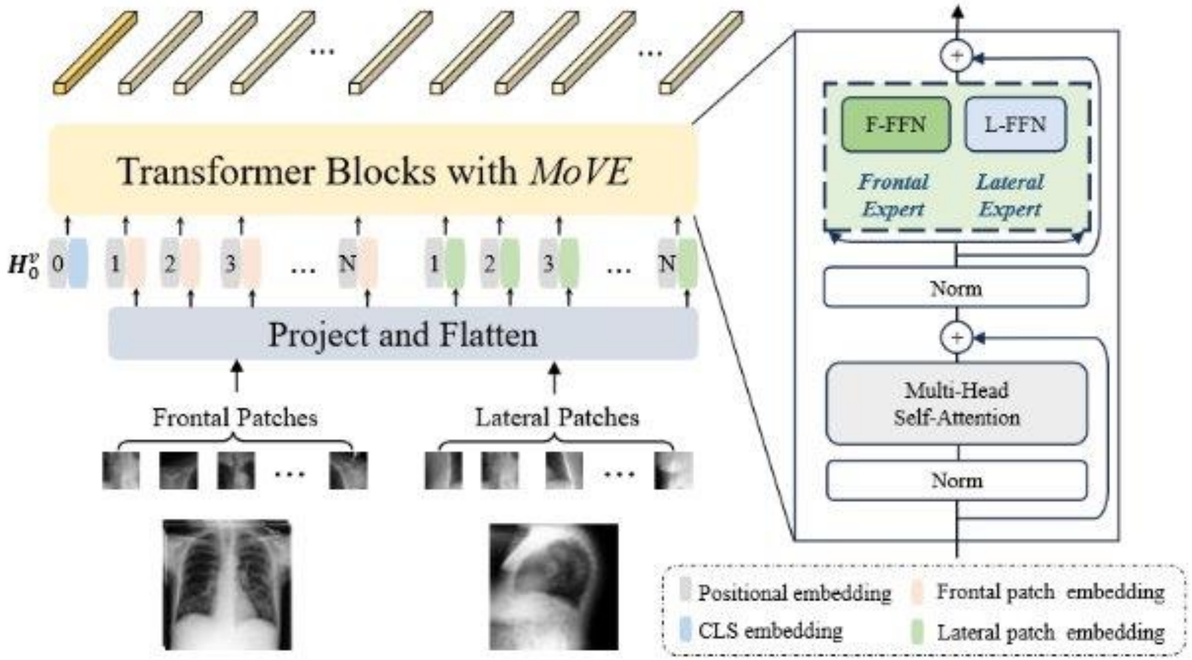


Figure 3: The multi-view image encoder. We input both the frontal and lateral views into *MoVE* blocks. The right side shows the structure of *MoVE*.

Modality-weighted local alignment

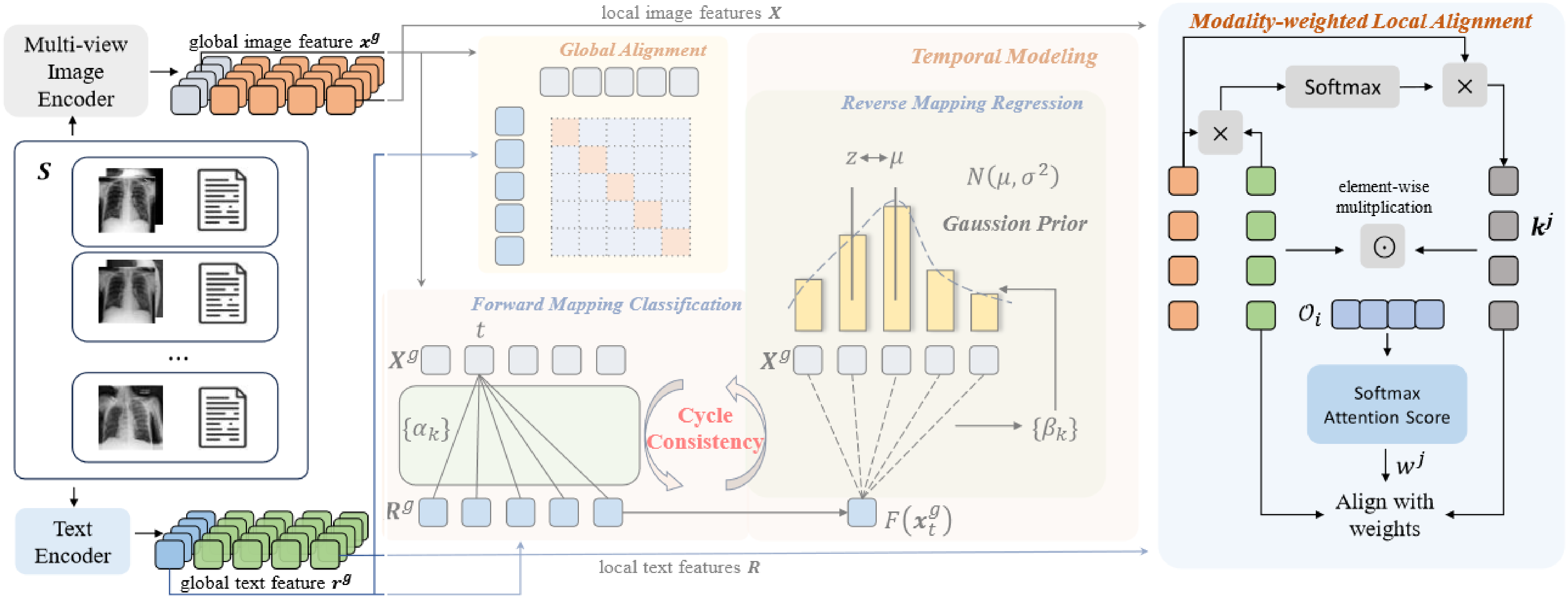


Figure 2: The framework of Med-ST. For spatial modeling, Med-ST extracts integrated multi-view visual representations, performs global alignment between paired global features, and introduces modality-weighted local alignment between paired local features. For temporal modeling, global image and text features in all pairs form two sequences respectively, and Med-ST imposes cross-modality bidirectional cycle consistency constraints between them.

Temporal modeling

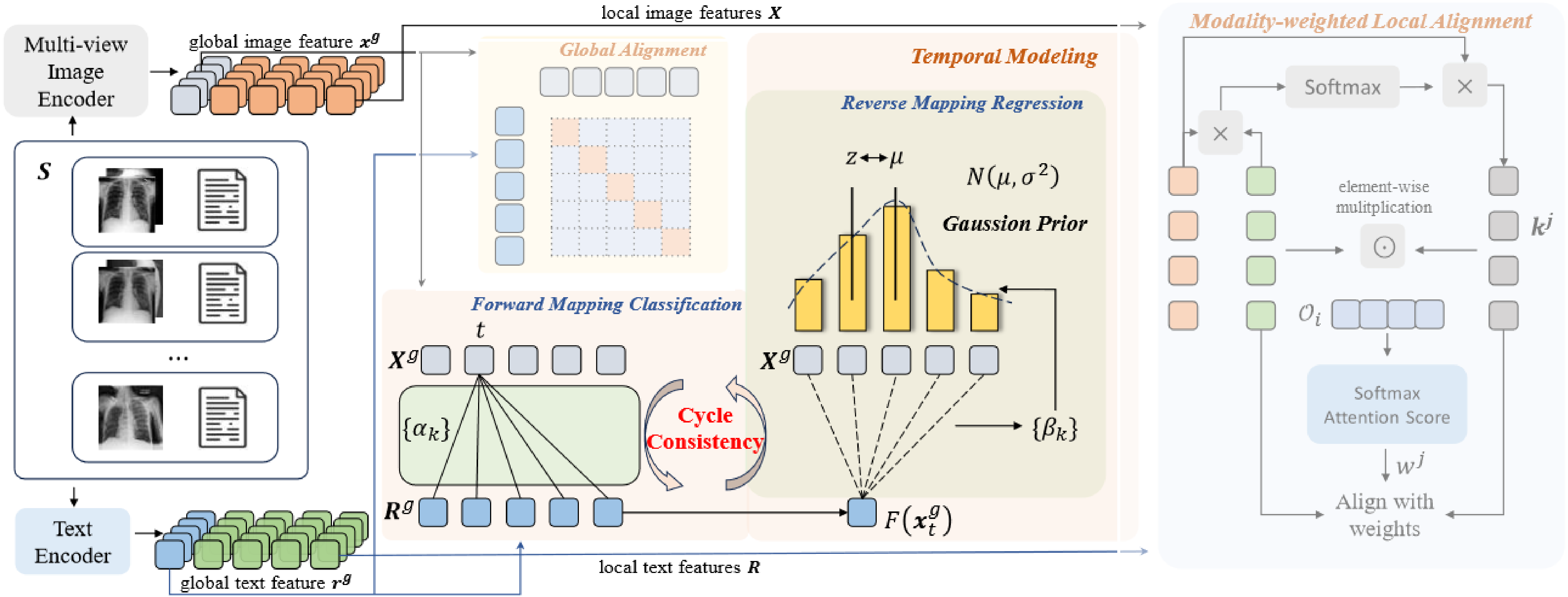


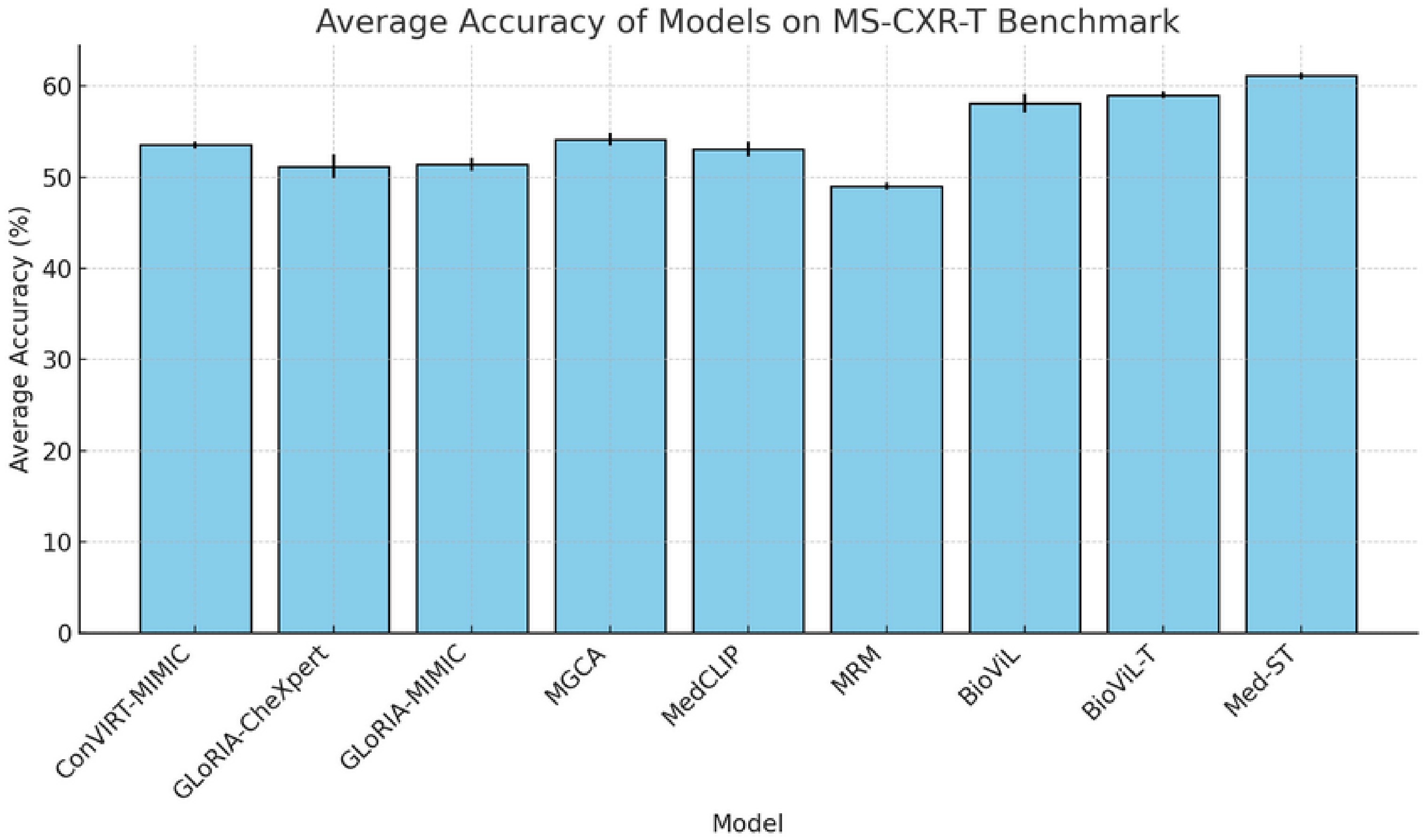
Figure 2: The framework of Med-ST. For spatial modeling, Med-ST extracts integrated multi-view visual representations, performs global alignment between paired global features, and introduces modality-weighted local alignment between paired local features. For temporal modeling, global image and text features in all pairs form two sequences respectively, and Med-ST imposes cross-modality bidirectional cycle consistency constraints between them.

Results

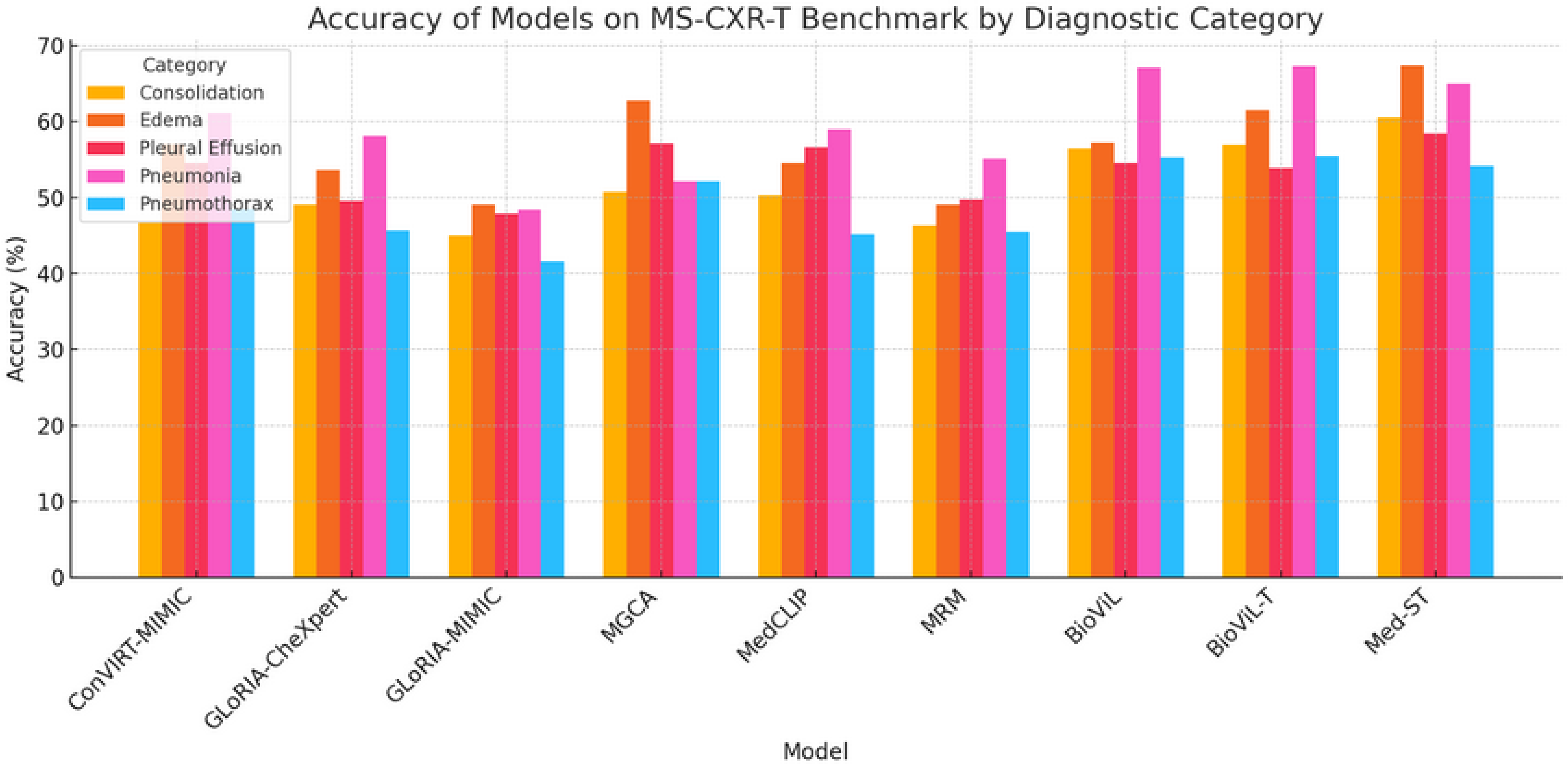
Table 1: Temporal image classification results on the MS-CXR-T benchmark across five findings under 10-fold cross-validation setting. We run with different seeds three times and calculate the mean and standard deviations. **Avg. Acc.** stands for average accuracy. PL.effusion denotes pleural effusion. We report the accuracy [%]. Best results are in boldface.

Model	Consolidation	Edema	Pl.effusion	Pneumonia	Pneumothorax	Avg. Acc.
ConVIRT-MIMIC (Zhang et al., 2022)	46.62± 1.03	57.04± 1.00	54.50± 0.19	61.09± 1.54	48.49± 0.22	53.55± 0.36
GLoRIA-CheXpert (Huang et al., 2021)	49.13± 1.52	53.67± 0.61	49.56± 1.84	58.11± 2.21	45.64± 2.49	51.22± 1.35
GLoRIA-MIMIC (Huang et al., 2021)	44.99± 0.47	49.13± 1.54	47.85± 3.08	60.18± 1.11	41.53± 0.92	48.74± 0.84
MGCA (Wang et al., 2022a)	50.79± 0.44	62.71± 0.24	57.17± 0.69	63.73± 0.89	52.16± 1.02	57.31± 0.34
MedCLIP (Wang et al., 2022c)	50.32± 0.65	54.53± 2.82	56.61± 0.65	58.94± 2.62	45.19± 1.47	53.12± 0.13
MRM (Wang et al., 2022c)	46.33± 2.89	49.09± 5.70	49.13± 0.69	55.14± 1.42	45.48± 2.97	49.03± 0.80
BioViL (Boecking et al., 2022)	56.40± 0.24	57.26± 0.77	54.51± 0.39	67.10± 0.01	55.30± 0.21	58.11± 0.02
Temporal-based						
BioViL-T (Bannur et al., 2023)	56.93± 1.77	61.55± 0.95	53.94± 0.89	**67.24**± 0.20	**55.46**± 0.01	59.02± 0.34
Med-ST	**60.57**± 1.18	**67.35**± 0.32	**58.47**± 1.50	65.00± 0.34	54.18± 0.81	**61.12**± 0.34

Results


Average Accuracy of Models on MS-CXR-T Benchmark
Average Accuracy (%)
0
10
20
30
40
50
60
ConVIRT-MIMIC
GLoRIA-CheXpert
GLoRIA-MIMIC
MGCA
MedCLIP
MRM
BioViL
BioViL-T
Med-ST
Model

Results



Accuracy of Models on MS-CXR-T Benchmark by Diagnostic Category
Category
Consolidation
Edema
Pleural Effusion
Pneumonia
Pneumothorax
Accuracy (%)
0
10
20
30
40
50
60
70
ConVIRT-MIMIC
GLoRIA-CheXpert
GLoRIA-MIMIC
MGCA
MedCLIP
MRM
BioViL
BioViL-T
Med-ST
Model

Results

Table 2: Temporal sentence similarity classification results on RadGraph subset of MS-CXR-T sentence similarity benchmark. Accuracy and AUROC scores are reported. Best results are in boldface.

Model	MLM	Acc.	AUROC
MedCLIP	×	66.41	52.66
MGCA	×	75.42	76.38
BioViL	✓	69.49	68.92
BioViL-T	✓	78.81	81.39
Med-ST	×	**83.76**	**84.60**

Table 3: Zero-shot classification results on RSNA. We report Accuracy, F1 and AUROC scores. Best results are highlighted in bold.

Model	Acc.	F1	AUROC
MGCA	67.25	54.87	73.97
BioViL	64.10	55.19	75.27
BioViL-T	63.23	54.90	75.12
Med-ST	**68.37**	**57.63**	**77.14**

Table 4: Medical image classification results on COVIDx datasets with 1%, 10% and 100% training data.

Method	1%	10%	100%
GLoRIA	67.3	77.8	89.0
ConVIRT	72.5	82.5	92.0
GLoRIA-MIMIC	66.5	80.5	88.8
MedKLIP	74.5	85.2	90.3
MGCA	**74.8**	84.8	92.3
Med-ST	71.2	**87.7**	**93.0**

Summary

- Multimodal vision-language models are example of effort towards generalization of computer-aided radiology solutions
- They achieve moderate-to-good accuracy in downstream tasks and extend capabilities of simpler models
- The temporal and spatial context provides benefits (yet limited and architecture-dependent)
- Phrase grounding is a step towards more explainable models