# 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

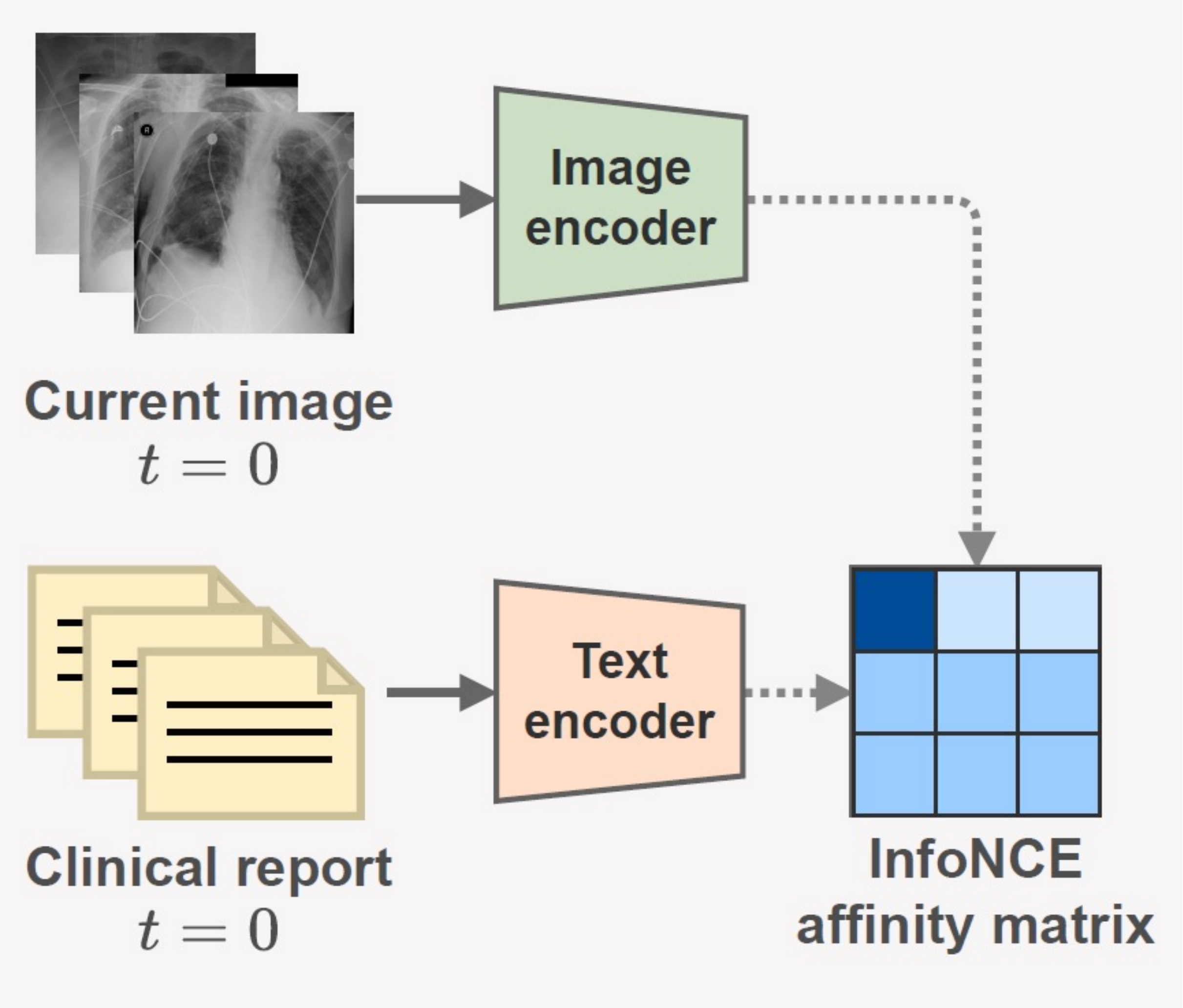
Presenter: Linyang He

# Overview

- This paper proposes a new self-supervised vision–language processing (VLP) framework, called BioViL-T, that leverages the temporal relationship between medical images and reports to enhance the cross-modal semantic alignment.
- BioViL-T uses a hybrid CNN-Transformer multi-image encoder that can handle missing prior images and spatial misalignment in longitudinal image sequences.
- BioViL-T achieves state-of-the-art performance on multiple downstream tasks, including progression classification, phrase grounding, and report generation, and provides a new multi-modal temporal benchmark dataset MS-CXR-T to evaluate the temporal semantic quality of chest X-ray VLP models.

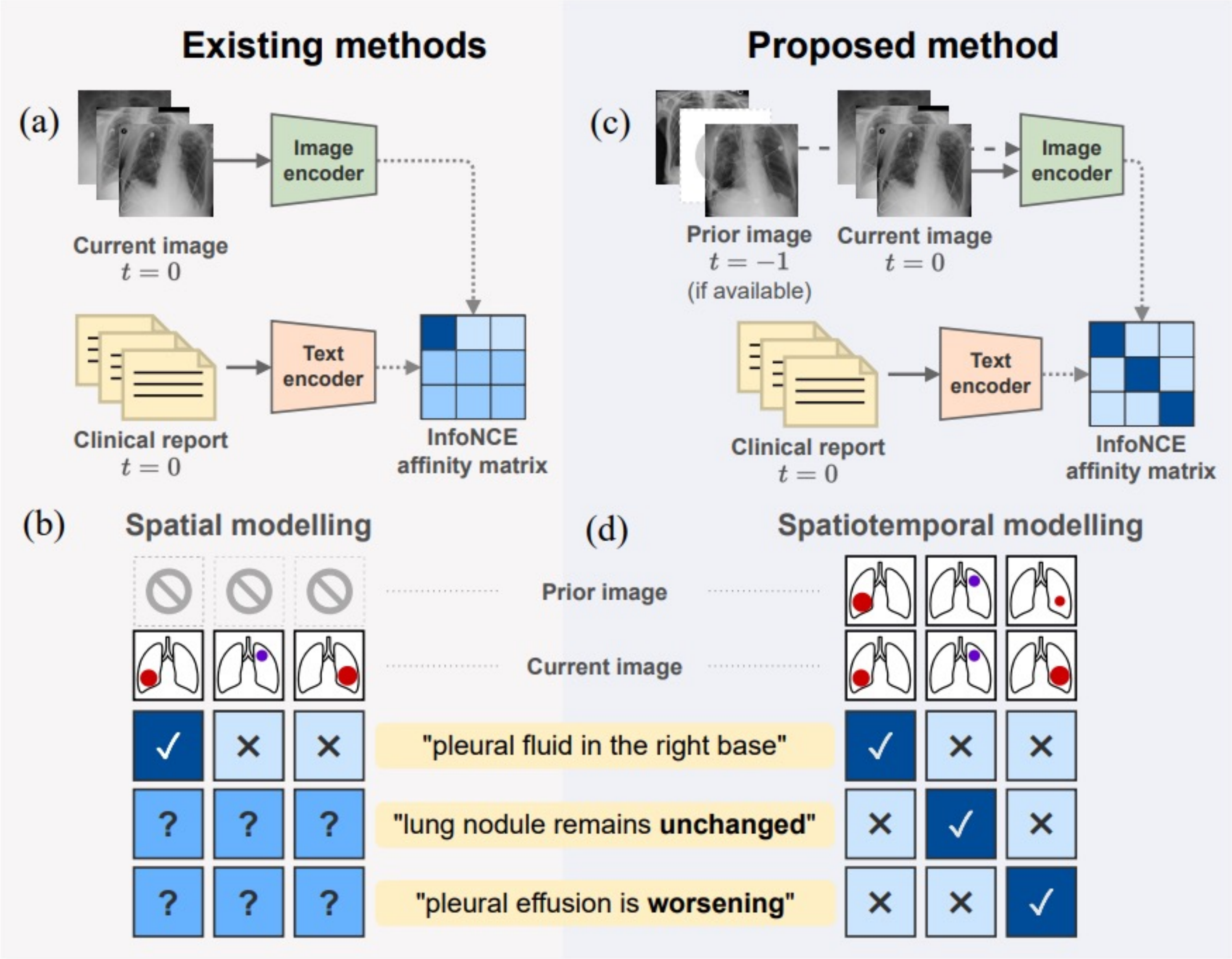
# Background

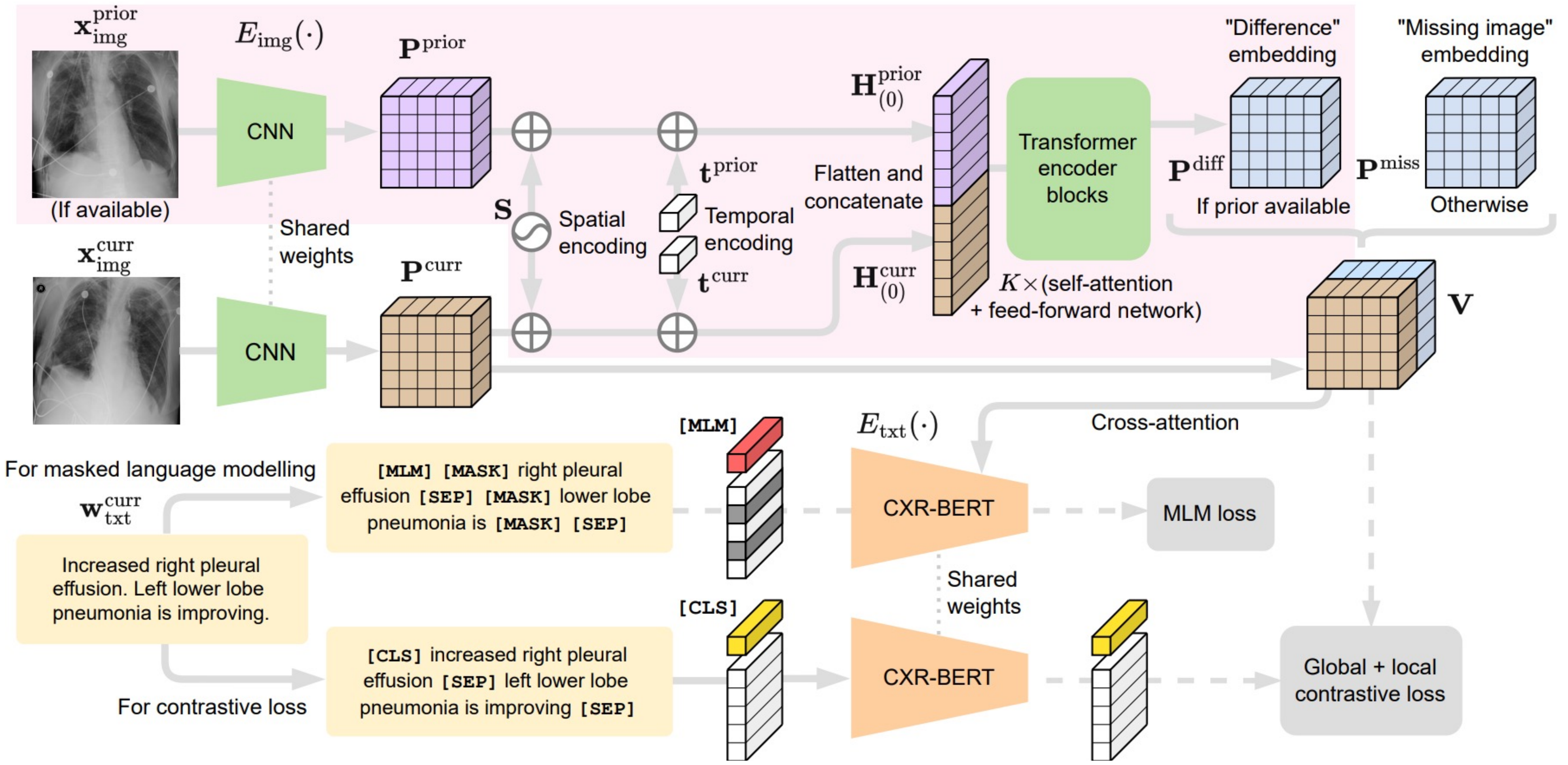
## Vision-Language Processing



# Limitation of Previous Biomedical VLP

They didn't consider *temporal information*.





$\mathbf{x}_{\mathrm{img}}^{\mathrm{prior}}$
(If available)
$E_{\mathrm{img}}(\cdot)$
CNN
Shared weights
$\mathbf{x}_{\mathrm{img}}^{\mathrm{curr}}$
CNN
$\mathbf{P}^{\mathrm{prior}}$
$\mathbf{P}^{\mathrm{curr}}$
$\mathbf{S}$
Spatial encoding
$\mathbf{t}^{\mathrm{prior}}$
Temporal encoding
$\mathbf{t}^{\mathrm{curr}}$
Flatten and concatenate
$\mathbf{H}_{(0)}^{\mathrm{prior}}$
$\mathbf{H}_{(0)}^{\mathrm{curr}}$
Transformer encoder blocks
$K\times$(self-attention + feed-forward network)
"Difference" embedding
$\mathbf{P}^{\mathrm{diff}}$
If prior available
"Missing image" embedding
$\mathbf{P}^{\mathrm{miss}}$
Otherwise
$\mathbf{V}$
Cross-attention
$E_{\mathrm{txt}}(\cdot)$
For masked language modelling
$\mathbf{w}_{\mathrm{txt}}^{\mathrm{curr}}$
[MLM] [MASK] right pleural effusion [SEP] [MASK] lower lobe pneumonia is [MASK] [SEP]
[MLM]
CXR-BERT
MLM loss
Increased right pleural effusion. Left lower lobe pneumonia is improving.
Shared weights
[CLS]
[CLS] increased right pleural effusion [SEP] left lower lobe pneumonia is improving [SEP]
CXR-BERT
For contrastive loss
Global + local contrastive loss

# Model:BioVIL-T

- Image Representation
- Transformer Encoding
- Text Representation
- Cross-Attention
- Loss Computation

# Datasets & Experiments

Pre-training

- MIMIC-CXR v2
  - longitudinal chest X-ray images and radiology reports

Evaluated

- Several downstream tasks datasets
  - MS-CXR
  - RSNA Pnenumonia Detection
  - Chest ImaGenome
  - MS-CXR-T(New benchmark proposed in this paper)
    - Temporal Image Classification
    - Sentence similarity

# 4 tasks

- Report generation
- Temporal image classification
- Phrase grounding
- Sentence similarity (New)

# Results

## Report Generation Task

| | Method | Pre-training | PI / PR | BLEU-4 | ROUGE | CHEXBERT | TEM |
|---|---|---|---|---|---|---|---|
| NN | CXR-RePaiR-2 [25] | BioViL | ✗ / ✗ | 2.1 | 14.3 | 28.1 | 12.5 |
| NN | Baseline (NN) [9] | BioViL | ✗ / ✗ | 3.7 | 20.0 | 28.3 | 11.1 |
| NN | Proposed (NN) | BioViL-T | ✓/ ✗ | 4.5 | 20.5 | 29.0 | 13.0 |
| AR | Baseline (AR) [9] | BioViL | ✗ / ✗ | 7.5 ± 0.1 | 27.9 ± 0.1 | 29.3 ± 0.3 | 13.8 ± 0.1 |
| AR | Proposed | BioViL-T | ✓/ ✗ | 8.2 ± 0.1 | 28.7 ± 0.1 | 30.2 ± 0.7 | 16.0 ± 0.3 |
| AR | Proposed | BioViL-T | ✓/ ✓ | **9.2 ± 0.3** | **29.6 ± 0.1** | **31.7 ± 1.0** | **17.5 ± 0.1** |

# Results

## Temporal Image Classification Task (MS-CXR-T)

| | Method (% of labels) | Pre-train | Consolidation | Pl. effusion | Pneumonia | Pneumothorax | Edema |
|---|---|---|---|---|---|---|---|
| Z&F | BioViL-T prompt (0%) | Temporal | 53.6 ± 1.9 | 59.7 ± 2.1 | 58.0 ± 3.9 | 34.9 ± 1.0 | 64.2 ± 1.5 |
| | BioViL-T (10%) | Temporal | 59.7 ± 2.4 | 62.4 ± 1.4 | 60.1 ± 2.1 | 35.3 ± 2.6 | 62.6 ± 1.7 |
| Supervised | CNN + Transformer | ImageNet | 44.0 ± 2.0 | 61.3 ± 1.6 | 45.1 ± 3.5 | 31.5 ± 3.1 | 65.5 ± 1.1 |
| | CheXRelNet [37] | ImageNet | 47 | 47 | 47 | 36 | 49 |
| | BioViL [9] | Static | 56.1 ± 1.5 | 62.3 ± 1.1 | 59.4 ± 1.0 | 41.7 ± 2.8 | 67.5 ± 0.8 |
| | BioViL w/reg [9] | Static | 56.0 ± 1.5 | 63.0 ± 0.9 | 60.2 ± 0.7 | 42.5 ± 2.7 | 67.5 ± 0.9 |
| | BioViL-T wout curation | Temporal | 58.9 ± 1.7 | 65.5 ± 0.7 | 61.5 ± 2.2 | 44.4 ± 2.1 | 67.4 ± 0.8 |
| | BioViL-T | Temporal | **61.1 ± 2.4** | **67.0 ± 0.8** | **61.9 ± 1.9** | 42.6 ± 1.6 | **68.5 ± 0.8** |

# Results

## Phrase Grounding Task (MS-CXR)

| **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** |

# Results

## Sentence similarity (MS-CXR-T)

| Text Model | MS-CXR-T (361 pairs) | | RadNLI (145 pairs) | |
|---|---|---|---|---|
| | 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** |

# Results Summary

- BioViL-T achieves state-of-the-art results on chest X-ray report generation, temporal image classification, and phrase grounding tasks. It also outperforms domain-specific BERT models on sentence similarity tasks.

# Conclusion

- This paper presents BioViL-T, a novel self-supervised VLP framework that exploits the temporal structure of biomedical data to learn better cross-modal representations.
- BioViL-T demonstrates its versatility and effectiveness on various downstream tasks, both static and temporal, achieving state-of-the-art performance.
- BioViL-T also introduces a new dataset MS-CXR-T to benchmark the temporal semantic quality of VLP models.

# Future Work

- Extend BioViL-T to other modalities such as MRI or CT scans, incorporating more prior images or reports for richer temporal information, and exploring other self-supervised objectives for VLP.

# Quiz

What kind of encoder does BioViL-T use to extract spatio-temporal features from a series of images?

How does BioViL-T utilize prior reports as a prompt in the report generation task?

# Thank You!
# Q&A