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]

Abstract

Medical vision-language pre-training methods mainly leverage the correspondence between paired medical images and radiological reports. Although multi-view spatial images and temporal sequences of image-report pairs are available in off-the-shelf multi-modal medical datasets, most existing methods have not thoroughly tapped into such extensive supervision signals. In this paper, we introduce the Med-ST framework for fine-grained spatial and temporal modeling to exploit information from multiple spatial views of chest radiographs and temporal historical records. For spatial modeling, Med-ST employs the *Mixture of View Expert (MoVE)* architecture to integrate different visual features from both frontal and lateral views. To achieve a more comprehensive alignment, Med-ST not only establishes the global alignment between whole images and texts but also introduces modality-weighted local alignment between text tokens and spatial regions of images. For temporal modeling, we propose a novel cross-modal bidirectional cycle consistency objective by forward mapping classification (FMC) and reverse mapping regression (RMR). By perceiving temporal information from simple to complex, Med-ST can learn temporal semantics. Experimental results across four distinct tasks demonstrate the effectiveness of Med-ST, especially in temporal classification tasks. Our code and model are available at https://github.com/SVT-Yang/MedST.

[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. Correspondence to: Bing Su <subingats@gmail.com>, Wayne Xin Zhao <batmanfly@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

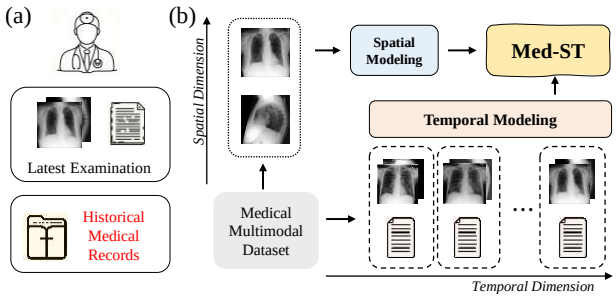


Figure 1: The motivation and our framework: (a) The practice of physicians in a clinical setting. (b) We propose the Med-ST framework, which explicitly designs spatial and temporal modeling by mining spatio-temporal supervision signals from the dataset.

1. Introduction

Medical vision-language pre-training aims to learn visual and textual representations from paired radiological reports and images. The learned representations can be adapted to various downstream tasks, which have great potential value for clinical auxiliary diagnosis and so on. Existing medical vision language pre-training methods generally focus on utilizing the correspondences between paired images and texts (Huang et al., 2021; Wang et al., 2022a; Zhou et al., 2022; Wan et al., 2023; Zhou et al., 2023). The radiological report is aligned with the frontal medical image through contrastive learning or masked modality reconstruction (Cheng et al., 2023; Zhou et al., 2023; Moon et al., 2022). Although the learned visual representations can encode medical semantic information, they cannot fully capture fine-grained spatial information in different views, nor can they distinguish fine-grained temporal differences between image-text pairs of the same patient at different times.

The applicability of medical vision-language pre-training models significantly depends on their capacity to mimic human cognitive process in interpreting radiological reports and images. Typically, in a clinical setting as shown in Figure 1(a), the doctor's diagnosis relies on the comprehensive examination results, considering frontal views and lateral views. This process also involves reviewing the patient's his-

torical medical records to comprehend past symptoms and trends. Fortunately, such comprehensive spatio-temporal information is available in off-the-shelf medical multimodal datasets. Besides image-text pairs, these datasets naturally embody parts of two types of supervision signals: spatial information (frontal and lateral views) and temporal information (chronological order of diagnoses).

However, these supervision signals have not been fully explored by existing medical pre-training methods. For spatial information, images of lateral views are either ignored or treated identically to those of frontal views (Huang et al., 2021; Wang et al., 2022a; Zhou et al., 2023; Dawidowicz et al., 2023; Shu et al., 2024). Regarding temporal information, existing work only utilizes a single prior image and does not design objectives specifically for this self-supervised signal (Bannur et al., 2023). The temporal information in historical sequences of image-report pairs has not yet been explored.

In this work, we introduce the Med-ST framework, which jointly exploits the comprehensive spatial and temporal information within off-the-shelf medical datasets to supervise the pre-training of visual and textual representations. As visualized in Fig. 1(b), Med-ST comprises two main components: spatial modeling and temporal modeling. In spatial modeling, we introduce the *Mixture of View Expert (MoVE)* architecture to construct the multi-view image encoder. We feed both frontal and lateral views into the encoder, which utilizes two experts to extract complementary information from different spatial perspectives. Features generated by both experts are integrated to form a joint visual representation of these varied spatial angles. Visual representations are then aligned with textual representations through contrastive learning (Oord et al., 2018). Recognizing that pathological areas occupy only a part of the images or texts, we propose modality-weighted local alignment, which assigns different weights to different local image patches and text tokens pairs based on the information they contain, achieving a fine-grained local alignment. For temporal modeling, we encourage the learned image-text feature sequences to express the same semantic changes, allowing the pre-training model to gain more supervision signals. Motivated by Dwibedi et al. (2019), we perform bidirectional cycle consistency between sequences of different modalities. Since the model has never explicitly perceived temporal sequence information, directly learning such information is challenging. Therefore, we adopt a novel bidirectional learning approach from simple to complex. In the forward process, we design a relatively simple classification loss to initially perceive information in the sequence. In the reverse process, we add a Gaussian prior for regression. We penalize inconsistencies by transforming the mismatch measurements into classification and regression targets. Through a bidirectional process from simple to complex, our model perceives the information of the sequence context, thus being able to capture changes. By incorporating spatial and temporal modeling, the global multi-view visual representation and textual representation interact in both spatial and temporal dimensions. In this way, our Med-ST gains the ability to jointly encode fine-grained spatiotemporal information into multi-modal representations.

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 cross-modal 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.

2. Related Work

Vision-Language Pre-training Models. Multimodal vision-language pre-training (VLP) is a field of research that aims to jointly learn representations of both visual and textual data. It trains deep neural networks on large-scale datasets that contain both image and text data, mainly image-text pairs (Radford et al., 2021; Li et al., 2021; 2023; Wang et al., 2022b; Alayrac et al., 2022; Chen et al., 2022; Bao et al., 2022; Zhang et al., 2023; Lu et al., 2023). By learning to represent images and text in a shared embedding space, multimodal pre-training can generate cross-modal representations that capture the semantic meaning of both modalities. Due to the impressive performance of VLP in the natural domain, it has also gained significant popularity in other fields, including the medical domain. It is important to note that directly transferring VLP models to the medical domain may not be feasible due to differences between the two domains (Chambon et al., 2022a;b).

Representation Learning in Medical Domain. The research on vision-language pre-training in medical domain can be divided into two branches: one is leveraging external knowledge bases (Wu et al., 2023; Wang et al., 2022c; Qin et al., 2022) and the other is fully utilizing existing datasets through specific model design (Huang et al., 2021;

Wang et al., 2022a; Zhou et al., 2022; 2023; Cheng et al., 2023). In the first branch, MedKLIP and MedCLIP (Wu et al., 2023; Wang et al., 2022c) leverage entity extraction or descriptive information to enhance the model. Currently, several large language models are being applied in medical domain (Wang et al., 2023; Singhal et al., 2022; Shaib et al., 2023; Yunxiang et al., 2023), primarily by using medical instruction datasets to generate more accurate answers. However, utilizing external knowledge may be limited by the availability and completeness of the external knowledge base. In the other branch, MGCA (Wang et al., 2022a) learns generalized medical visual representations through leveraging the inherent semantic correlations and MRM (Zhou et al., 2023) learns knowledge-enhanced semantic representations by reconstructing both masked image patches and masked report tokens. However, these works only utilize the frontal view of the chest X-ray, discarding the lateral views (Huang et al., 2021; Wang et al., 2022a), or treat them the same as frontal views without exploring the relationship between them (Zhou et al., 2023; Dawidowicz et al., 2023). We believe there is an association between the frontal and lateral views that could be helpful for downstream tasks, and medical evidence has shown the value of lateral views in diagnosis (Raoof et al., 2012). Therefore, we attempt to better utilize the lateral view through spatial modeling.

Temporal Self-Supervised Learning in Medical Domain. Effective representations can be learned from time series data through self-supervision, eliminating the need for manual annotation. Specifically, medical time series data often contain abundant semantic information. BioViL-T (Bannur et al., 2023) exploits time-related information by using prior images, thereby gaining additional supervision. However, BioViL-T only utilizes a single prior image and lacks a temporal optimization objective. Our approach considers both text and image pairs in a temporal context and designs a time-related objective, i.e., cross-modality bidirectional cycle consistency, to better leverage the information provided by time series.

3. Method

3.1. Overview

Existing medical datasets contain some inherent spatial and temporal supervision signals that have been previously overlooked, obviating the need for supplementary manual annotations. A collection of images associated with a single report is referred to as a *study*[1]. Given a dataset comprising multiple *studies*, each *study* typically includes a frontal view image along with its corresponding radiological report. In some cases, lateral views may be present. All *studies* of a patient have a chronological order. As a result, we can extract matched multi-view data and time-series data from the dataset, which provide spatial and temporal supervision signals for pre-training multi-modal representations.

Formally, we reorganize the data in a medical dataset $\mathcal{D}$ into a number of sequences, where each sequence contains a series of image-text pairs and typically corresponds to the historical diagnosis records of a patient. Some image-text pairs do not belong to any sequences, e.g., patients have only one diagnosis record. We regard such individual image-text pairs as sequences with a length of 1. Therefore, $\mathcal{D} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_{|\mathcal{D}|}\}$, where $|\mathcal{D}|$ is the number of sequences in the dataset. For the i-th sequence, $\boldsymbol{S}_i = \{(\boldsymbol{I}^f_{i,t}, \boldsymbol{I}^l_{i,t}, \boldsymbol{T}_{i,t})\}_{t=1}^{|\boldsymbol{S}_i|}$, where $\boldsymbol{I}^f_{i,t}$, $\boldsymbol{I}^l_{i,t}$, and $\boldsymbol{T}_{i,t}$ are the frontal image, the corresponding lateral image, and the related text at the t-th timestep, and $|\boldsymbol{S}_i|$ is the number of image-text pairs in this sequence. Not every image-text pair has a lateral view, i.e., for some i and t, $\boldsymbol{I}^l_{i,t}$ does not exist, and we set it to a tensor with all zeros in this case.

Our proposed model Med-ST leverages both spatial and temporal modeling to learn multi-modal representations from such sequence data of multi-view image-text pairs. Med-ST consists of a multi-view image encoder f_I and a text encoder f_T. The image encoder employs the proposed *MoVE* architecture to extract integrated representations from frontal and lateral images, i.e., $[\boldsymbol{x}^g_{i,t}, \boldsymbol{X}_{i,t}] = f_I([\boldsymbol{I}^f_{i,t}, \boldsymbol{I}^l_{i,t}])$, where $\boldsymbol{x}^g_{i,t} \in \mathbb{R}^d$ is the global image feature, $\boldsymbol{X}_{i,t} \in \mathbb{R}^{M \times d}$ is the sequence of patch-wise features and M is the number of tokens. The text encoder extracts the text representation, i.e., $[\boldsymbol{r}^g_{i,t}, \boldsymbol{R}_{i,t}] = f_T(\boldsymbol{T}^f_{i,t})$, where $\boldsymbol{r}^g_{i,t} \in \mathbb{R}^d$ is the global text feature and $\boldsymbol{R}_{i,t} \in \mathbb{R}^{W \times d}$ is the sequence of token-wise features and W is the number of tokens. As shown in Fig. 2, in the training phase, for each input sequence $\boldsymbol{S}$, Med-ST first extracts the image and text representations for all time steps: $\{\boldsymbol{x}^g_t, \boldsymbol{X}_t, \boldsymbol{r}^g_t, \boldsymbol{R}_t\}_{t=1}^{|\boldsymbol{S}|}$. By viewing image-text pairs in all sequences as independent, Med-ST performs global alignment between matched global image and text features $\{\boldsymbol{x}^g_t, \boldsymbol{r}^g_t\}$, and introduces modality-weighted local alignment between matched sequences of patch-wise and token-wise features $\{\boldsymbol{X}_t, \boldsymbol{R}_t\}$. For temporal modeling, the global image and text features form two feature sequences, $\boldsymbol{X}^g = [\boldsymbol{x}^g_1, \boldsymbol{x}^g_2, \cdots, \boldsymbol{x}^g_{|\boldsymbol{S}|}]$ and $\boldsymbol{R}^g = [\boldsymbol{r}^g_1, \boldsymbol{r}^g_2, \cdots, \boldsymbol{r}^g_{|\boldsymbol{S}|}]$, respectively, and Med-ST imposes the cross-modality cycle consistency constraints to align $\boldsymbol{X}^g$ and $\boldsymbol{R}^g$.

3.2. Spatial Modeling

3.2.1. Image Feature Extraction

Based on Vision Transformer (ViT) (Dosovitskiy et al., 2020), we develop a *Mixture of View Experts (MoVE)* architecture as f_I to encode the multi-view images. As shown in Fig. 3, for a pair of frontal and lateral images $(\boldsymbol{I}^f, \boldsymbol{I}^l)$ in a sequence sample $\boldsymbol{S}$, we initially divide them into patches,

[1] https://physionet.org/content/mimic-cxr/2.0.0/

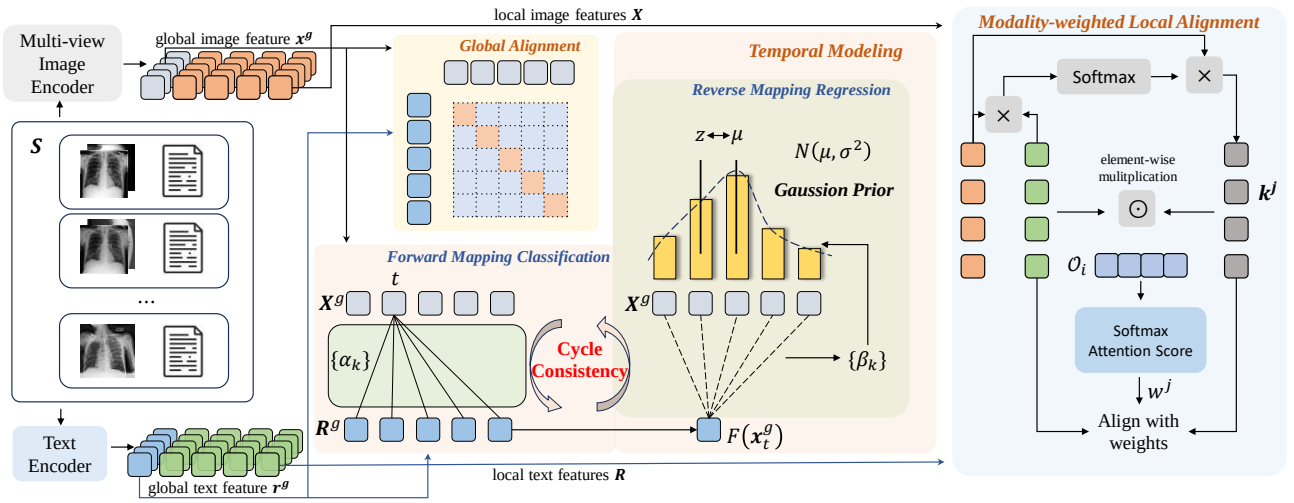


Figure 2: The framework of Med-ST. For spatial modeling, Med-ST extracts integrated multi-view visual representations, performs global alignment between paired global features, and introduces modality-weighted local alignment between paired local features. For temporal modeling, global image and text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

respectively, which are then flattened through a projector to obtain a series of patch embeddings $\{\boldsymbol{p}_j^f\}_{j=1}^{n_p}, \{\boldsymbol{p}_j^l\}_{j=1}^{n_p}$ with dimension D. We add a learnable embedding $\boldsymbol{p}_{\text{CLS}}$. To preserve the positional information of patches in each image, we introduce position encoding $\boldsymbol{E}_{pos} \in \mathbb{R}^{(n_p+1)\times D}$ to both frontal and lateral patches, respectively. All patch embeddings from both views are concatenated to form the embedding sequence $\boldsymbol{H}_0^v$ of the image encoder as follows, which serves as the input of the image encoder.

$$\boldsymbol{H}_0^v = (\boldsymbol{p}_{\text{CLS}} + \boldsymbol{E}_{pos}[0,:]) \oplus (\{\boldsymbol{p}_j^f\}_{j=1}^{n_p} + \boldsymbol{E}_{pos}[1:,:]) \oplus (\{\boldsymbol{p}_j^l\}_{j=1}^{n_p} + \boldsymbol{E}_{pos}[1:,:]) \tag{1}$$

where $\oplus$ is the concatenation operation and $\boldsymbol{H}_0^v \in \mathbb{R}^{(2n_p+1)\times D}$.

To enable the model to perceive distinctions between the two views, we employ our *MoVE* architecture to replace the standard transformer's block in ViT. Specifically, after obtaining the output $\boldsymbol{H}_{l-1}$ from the previous block, multi-head self-attention layers are performed to capture shared spatial dependencies among all patch embeddings. Then the first n_p+1 vectors are processed through a Frontal Feedforward Network (F-FFN), while the latter n_p vectors go through a Lateral Feedforward Network (L-FFN). These two FFNs act as distinct experts for capturing complementary information from different perspectives. The outputs of the two experts are concatenated to form the output $\boldsymbol{H}_l$ of this block. The image encoder consists of L *MoVE*-based blocks. For the output $\boldsymbol{H}_L$ of the final block, $\boldsymbol{x}^g = \boldsymbol{H}_L[0,:]$ is the global multi-view image feature and $\boldsymbol{X} = \boldsymbol{H}_L[0,:]$ is the sequence of patch-wise features.

3.2.2. Cross Modal Alignment

Global Alignment. For the t-th pair $(\boldsymbol{I}_{i,t}^f, \boldsymbol{I}_{i,t}^l, \boldsymbol{T}_{i,t})$ of the i-th sequence in the batch, we obtain $[\boldsymbol{x}_{i,t}^g, \boldsymbol{X}_{i,t}] = f_I([\boldsymbol{I}_{i,t}^f, \boldsymbol{I}_{i,t}^l])$. We also use the text encoder to extract the global text feature $\boldsymbol{r}_{i,t}^g$ and the sequence of token features $\boldsymbol{R}_{i,t}$. Following CLIP (Radford et al., 2021), we employ contrastive learning to bring global features of matching images and radiological reports closer while pushing global features of non-matching pairs apart.

$$l_{i,t}^{v2t} = -\log \frac{\exp(sim(\boldsymbol{x}_{i,t}^g, \boldsymbol{r}_{i,t}^g)/\tau_1)}{\sum_{b=1}^{B}\sum_{k=1}^{|\boldsymbol{S}_k|}\exp(sim(\boldsymbol{x}_{i,t}^g, \boldsymbol{r}_{b,k}^g)/\tau_1)}, \tag{2}$$

$$l_{i,t}^{t2v} = -\log \frac{\exp(sim(\boldsymbol{r}_{i,t}^g, \boldsymbol{x}_{i,t}^g)/\tau_1)}{\sum_{b=1}^{B}\sum_{k=1}^{|\boldsymbol{S}_k|}\exp(sim(\boldsymbol{r}_{i,t}^g, \boldsymbol{x}_{b,k}^g)/\tau_1)}. \tag{3}$$

Here, B is the batch size, $sim(\cdot,\cdot)$ is the similarity function and is measured by dot product. The temperature variable τ_1 is used to scale the logits. So the objective of Global Alignment is the average of the two losses:

$$\mathcal{L}_{global} = \frac{1}{2|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\frac{1}{|\boldsymbol{S}_i|}\sum_{t=1}^{|\boldsymbol{S}_i|}(l_{i,t}^{v2t} + l_{i,t}^{t2v}). \tag{4}$$

Modality-Weighted Local Alignment. Considering that pathological regions may only occupy specific portions of both the image and text, we introduce a novel modality-weighted local alignment between $\boldsymbol{X}_{i,t}$ and $\boldsymbol{R}_{i,t}$, i.e., paired sequences of patch and token features, to perform fine-grained spatial modeling.

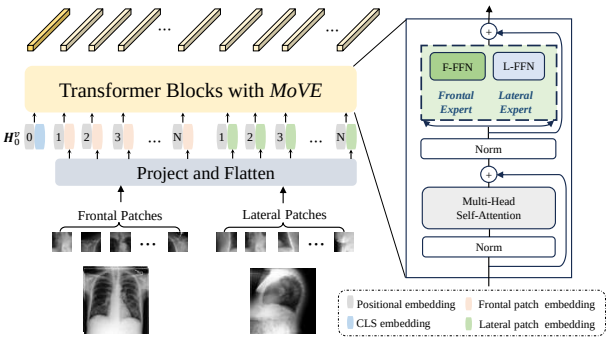


Figure 3: The multi-view image encoder. We input both the frontal and lateral views into *MoVE* blocks. The right side shows the structure of *MoVE*.

For the j-th text token feature $\boldsymbol{r}_{i,t,j}$ in $\boldsymbol{R}_{i,t}$, we compute the cosine similarity between $\boldsymbol{r}_{i,t,j}$ and all image patch features in $\boldsymbol{X}_{i,t}$, then we apply softmax function to the similarity scores. Thus we get $s = \text{softmax}(\boldsymbol{r}_{i,t,j}\boldsymbol{X}_{i,t}^T)$. These scores are then multiplied by the corresponding image patch features and aggregated. As a result, we generate a new textual-attended visual representation $\boldsymbol{k}_{i,t,j}$. Our goal of local alignment is to align $\boldsymbol{r}_{i,t,j}$ and its corresponding textual-attended visual representation $\boldsymbol{k}_{i,t,j}$.

$$l_{i,t,j}^{v2t} = -\log \frac{\exp(sim(\boldsymbol{k}_{i,t,j}, \boldsymbol{r}_{i,t,j})/\tau_2)}{\sum_{c=1}^{W} \exp(sim(\boldsymbol{k}_{i,t,j}, \boldsymbol{r}_{i,t,c})/\tau_2)}, \quad (5)$$

$$l_{i,t,j}^{t2v} = -\log \frac{\exp(sim(\boldsymbol{r}_{i,t,j}, \boldsymbol{k}_{i,t,j})/\tau_2)}{\sum_{c=1}^{W} \exp(sim(\boldsymbol{r}_{i,t,j}, \boldsymbol{k}_{i,t,c}/\tau_2)}. \quad (6)$$

Acknowledging that different local pairs contain varying amounts of information, those representing pathological semantics should be assigned greater importance. Different from Wang et al. (2022a) using averaged last-layer attention weight in a single modality, we considered the information contained in both visual and textual modalities. For the pair $(\boldsymbol{k}_{i,t,j}, \boldsymbol{r}_{i,t,j})$, the weights are determined by comparing the value of this pair against the average value of all local pairs in $\boldsymbol{X}_{i,t}, \boldsymbol{R}_{i,t}$. Firstly, we perform an element-wise multiplication of $\boldsymbol{r}_{i,t,j}$ and $\boldsymbol{k}_{i,t,j}$, i.e., $\boldsymbol{o}_{i,t,j} = \boldsymbol{r}_{i,t,j} \odot \boldsymbol{k}_{i,t,j}$. These vectors capture relationships and importance across different dimensions and we form them into a matrix $\boldsymbol{\mathcal{O}}_{i,t} = [\boldsymbol{o}_{i,t,1}, \boldsymbol{o}_{i,t,2}, \ldots, \boldsymbol{o}_{i,t,W}]$. Then we calculate the mean: $\bar{o}_{i,t} = \frac{1}{W}\sum_{j=1}^{W} \boldsymbol{o}_{i,t,j}$, from which we derive the weight through softmax attention score.

$$w_{i,t,j} = \left(\text{softmax}\left(\frac{\bar{o}_{i,t}\boldsymbol{Q}(\boldsymbol{\mathcal{O}}_{i,t}\boldsymbol{K})^T}{\sqrt{d}}\right)\right)_j, \quad (7)$$

where $\boldsymbol{Q} \in \mathbb{R}^{d\times d}, \boldsymbol{K} \in \mathbb{R}^{d\times d}$ are learnable matrices. We apply $w_{i,t,j}$ as the weight of $l_{i,t,j}$ and derive the Visual modality-weighted Local Alignment loss $\mathcal{L}_{\text{VLA}}$ as:

$$\mathcal{L}_{\text{VLA}} = \frac{1}{|\mathcal{D}|W}\sum_{i=1}^{|\mathcal{D}|}\frac{1}{|\boldsymbol{S}_t|}\sum_{t=1}^{|\boldsymbol{S}_t|}\sum_{j=1}^{W} w_{i,t,j}(l_{i,t,j}^{v2t} + l_{i,t,j}^{t2v}). \quad (8)$$

Correspondingly, we can derive the Texual Modality-Weighted Local Alignment Loss $\mathcal{L}_{\text{TLA}}$. The objective of Local Alignment is:

$$\mathcal{L}_{local} = \mathcal{L}_{\text{VLA}} + \mathcal{L}_{\text{TLA}}. \quad (9)$$

3.3. Temporal Modeling

We have observed that there is inherent temporal information in the dataset, which can provide additional supervision signals without the need for extra manual labeling. The objective of our temporal modeling is to use this information to enhance the alignment of image and text sequences that convey the same temporal semantics. As mentioned in Sec. 3.1, for sequence $\boldsymbol{S}$, the global image and text features form two feature sequences, $\boldsymbol{X}^g = [\boldsymbol{x}_1^g, \boldsymbol{x}_2^g, \cdots, \boldsymbol{x}_{|\boldsymbol{S}|}^g]$ and $\boldsymbol{R}^g = [\boldsymbol{r}_1^g, \boldsymbol{r}_2^g, \cdots, \boldsymbol{r}_{|\boldsymbol{S}|}^g]$, respectively, and Med-ST imposes the cross-modality bidirectional cycle consistency constraints to align $\boldsymbol{X}^g$ and $\boldsymbol{R}^g$.

Cycle Consistency. As illustrated in Figure 2, we want to establish bidirectional perception in the sequence. Given an image point $\boldsymbol{x}_t^g$ from $\boldsymbol{X}^g$, we first map it to a textual point using function F. Subsequently, the result of this mapping is transformed back to an image point through function G, thus we get $\boldsymbol{x}_{t'}^g = G(F(\boldsymbol{x}_t^g))$. If $t = t'$, the point $\boldsymbol{x}_t^g$ is cycle-consistent. The bidirectional matching mechanism of cycle consistency can perceive fine-grained contextual differences, thereby establishing a cross-modal match. Since our model has never explicitly perceived temporal sequence information, directly learning such information is challenging. Therefore, we adopt a novel bidirectional learning approach from simple to complex.

Forward Mapping Classification (FMC). We first calculate the distance between $\boldsymbol{x}_t^g$ and all points in $\boldsymbol{R}^g$ and then we get the soft nearest neighbor $F(\boldsymbol{x}_t^g)$ of $\boldsymbol{x}_t^g$:

$$F(\boldsymbol{x}_t^g) = \sum_{z=1}^{|\boldsymbol{S}|}\alpha_z \boldsymbol{r}_z^g, \ \alpha_z = \frac{\exp(-\langle \boldsymbol{x}_t^g, \boldsymbol{r}_z^g\rangle)}{\sum_{k=1}^{|\boldsymbol{S}|}\exp(-\langle \boldsymbol{x}_t^g, \boldsymbol{r}_k^g\rangle)}. \quad (10)$$

Since our sequence is paired, unlike Dwibedi et al. (2019), we can introduce constraints to the forward process. As our current model lacks temporal awareness, we incorporate a relatively simple classification loss in the forward process. We treat all features in sequence $\boldsymbol{R}^g$ as different classes and classify $\boldsymbol{x}_t^g$ accordingly. So the predicted output $\hat{y}_t = \alpha_t$. The ground truth $\boldsymbol{y}$ is represented by a one-hot vector, in which the t-th position is set to 1.

$$\mathcal{L}_t^{cls} = -\sum_{z=1}^{|\boldsymbol{S}|} y_z \log(\hat{y_z}). \quad (11)$$

As for sequence $\boldsymbol{S}$, the objective of FMC is as follows:

$$\mathcal{L}^{I}_{\text{FMC}} = \sum_{t}^{|\boldsymbol{S}|} \mathcal{L}^{cls}_{t}. \tag{12}$$

Reverse Mapping Regression (RMR). Since FMC enables cycle consistency from a relatively easy classification perspective, we want to incorporate a more complex objective in the process of reverse mapping. We impose greater penalties on predictions that are temporally distant according to the distance between $F(\boldsymbol{x}^g_t)$ and $\boldsymbol{x}^g_t$. To address this, we propose RMR. Initially, we obtain the distribution of similarities $\{\beta_z\}$ between $F(\boldsymbol{x}^g_t)$ and all points in $\boldsymbol{X}^g$.

$$\beta_z = \frac{\exp\left(-\langle F(\boldsymbol{x}^g_t), \boldsymbol{x}^g_z\rangle\right)}{\sum_{k=1}^{|\boldsymbol{S}|} \exp\left(-\langle F(\boldsymbol{x}^g_t), \boldsymbol{x}^g_k\rangle\right)}. \tag{13}$$

Representations that are temporally closer tend to be more similar. Thus the further the distance from timestep t, the smaller the similarity should be. So we impose a Gaussian prior on β. Our goal is to maximize β_t and ensure that the similarity distribution peaks at t. Therefore, we optimize $\frac{|t-\mu|^2}{\sigma^2}$. A regularization term $\log(\sigma)$ is added to ensure the peak of the distribution remains intact and avoid a reduction in variance resulting from this optimization. Since the temporal information contained in the sequence may have a large span, for cases with larger errors, we optimize $|t-\mu|$. So the objective of RMR for sequence $\boldsymbol{S}$ is as follows:

$$\mathcal{L}^{I}_{\text{RMR}} = \begin{cases} \dfrac{|t-\mu|^2}{\sigma^2} + \lambda \log(\sigma) & , |t-\mu| \leq \delta, \\ \delta|t-\mu| - \dfrac{\delta^2}{2} & , \text{otherwise}, \end{cases} \tag{14}$$

where $\mu = \sum_{z=1}^{|\boldsymbol{S}|} \beta_z \cdot z$ and $\sigma^2 = \sum_{z=1}^{|\boldsymbol{S}|} \beta_z \cdot (z-\mu)^2$, λ denotes the regularization weight, and δ is a hyperparameter.

Therefore, for all sequences in $\mathcal{D}$, we sum these two objectives to obtain $\mathcal{L}^{I}_{temporal} = \mathcal{L}^{I}_{\text{RMR}} + \mathcal{L}^{I}_{\text{FMC}}$. Similarly, starting from a point in the text sequence, we can get $\mathcal{L}^{T}_{temporal}$. So the objective of temporal modeling is:

$$\mathcal{L}_{temporal} = \mathcal{L}^{I}_{temporal} + \mathcal{L}^{T}_{temporal}. \tag{15}$$

3.4. Overall Objective

Therefore, our overall objective is denoted as $\mathcal{L}$ in Eq. 16. By optimizing this objective, our model can perceive spatial and temporal information with greater granularity.

$$\mathcal{L} = \mathcal{L}_{global} + \lambda_1 \mathcal{L}_{local} + \lambda_2 \mathcal{L}_{temporal}, \tag{16}$$

where λ_1 and λ_2 are hyperparameters that control the weights of the loss functions.

4. Experiments

4.1. Pre-training Setup

Dataset. We pretrain our Med-ST framework on **MIMIC-CXR** dataset (Johnson et al., 2019). It is a publicly available dataset that includes paired chest X-ray images and radiology reports. Sometimes images incorporate both frontal and lateral views. All pairs have attributes of `studydate` and `studytime`, which allow us to construct sequences. We pre-process the dataset following Wang et al. (2022a), including image transformation and text tokenization, resulting in $232k$ image-text pairs. We preserve lateral views and thus $114k$ pairs include lateral images.

Implementation Details. Our code is implemented using PyTorch (Paszke et al., 2019). The pre-training is performed using two GeForce RTX 3090 GPUs. Due to the effectiveness of pre-trained language models, we utilized BioClinicalBERT (Alsentzer et al., 2019) as the text encoder. For unified modal architecture design, we defaulted to using ViT-B/16 (Dosovitskiy et al., 2020) as the image encoder backbone. We use the AdamW optimizer (Loshchilov & Hutter, 2017) with a learning rate of 4e-5 and weight decay of 0.05. We employed a linear warmup with cosine annealing scheduler (Loshchilov & Hutter, 2016), initializing the learning rate to 1e-8 and the warmup epoch to 20. The τ_1 is 0.07 and τ_2 is 0.1. We set λ_1, λ_2 to 1, 1 in Eq. 16 and δ=2, λ=0.001 in Eq. 14. For more details, please refer to Appendix A.

4.2. Experimental Setup

4.2.1. Dataset

MS-CXR-T benchmark (Bannur et al., 2023) is a multimodal benchmark dataset aimed at evaluating biomedical vision-language processing models in radiology, specifically through two temporal tasks: image classification of chest X-rays and analysis of sentence similarity.

RSNA dataset (Shih et al., 2019) is a publicly available medical imaging dataset. It contains chest radiographs with annotations for the presence or absence of pneumonia.

COVIDx (Wang et al., 2020) is a dataset with over 30,000 chest X-ray (CXR) images from over 16,600 patients, including 16,490 positive COVID-19 cases. It's used for a three-class classification task: categorizing radiographs as COVID-19, non-COVID pneumonia, or normal.

4.2.2. Downstream Tasks

Temporal Image Classification. It includes 1,326 labeled chest X-rays across five findings (Consolidation, Edema, Pleural Effusion, Pneumonia, and Pneumothorax), categorized into three stages of disease progression: `Improving`,

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

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**

`Stable, Worsening`. Due to the limited data size, we choose an SVM (Boser et al., 1992) classifier and employ cross-validation. We report the accuracy for each findings and the average classification accuracy under the setting of 5-fold and 10-fold.

Temporal Sentence Similarity. It includes 361 sentence pairs for evaluating temporal-semantic similarity. The task focuses on binary classification of pairs as paraphrases or contradictions regarding disease progression. This dataset includes two subsets: RadGraph and Swaps.We report our results on the RadGraph subset since it is more challenging. For more experimental settings and results on the Swaps subset, please refer to the Appendix B.2.

Zero-shot Image Classification. We test our zero-shot classification results on RSNA. We use the dataset division from previous work (Wang et al., 2022a; Huang et al., 2021) and also employ the prompt construction of BioViL-T. We report our classification results on the test set.

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**

Medical Image Classification. Our model is tested on the COVIDx dataset to evaluate its efficacy in medical image classification. As with prior methods (Wang et al., 2022a), we fine-tune a classification head using different proportions of training data and then evaluate the classification performance on the test set.

4.3. Results

Temporal Image Classification Results. In Table 1, we report our performance on the task of temporal image classification under a 10-fold cross-validation setup. Our method achieves the highest average accuracy, surpassing BioViL-T (Bannur et al., 2023), which also utilizes temporal information, by 2.10%. Across five different findings, our method achieves the highest scores in three categories, particularly in the edema category, where our approach exceedes BioViL-T by 5.80%. Table 7 also validates our performance. We also achieve the best average accuracy in the experimental setup with a fold number of 5 in Table 7. Our method consistently achieves the best results in different cross-validation setups. This demonstrates that through

Table 5: Ablation study on objectives and lateral view use.

$\mathcal{L}_{global}$	$\mathcal{L}_{temporal}$	$\mathcal{L}_{local}$	Lateral	Consolidation	Edema	Pl.effusion	Pneumonia	Pneumothorax	Avg. Acc.
✓		✓	✓	55.75± 2.52	62.20± 1.00	57.91± 1.20	63.88± 0.79	54.53± 0.99	58.85± 0.61
✓	✓		✓	54.89± 0.51	65.30± 0.91	57.42± 0.39	66.40± 0.19	53.40± 0.96	59.49± 0.30
✓	✓	✓		55.09± 1.34	63.32± 0.78	58.06± 1.40	53.58± 0.40	50.25± 0.99	58.06± 0.29
✓	✓	✓	✓	60.57± 1.18	67.35± 0.32	58.47± 1.50	65.00± 0.34	54.18± 0.81	61.12± 0.34

our temporal modeling, the model is able to capture the semantics of temporal changes and integrate this information into the model, making it more suitable for temporal tasks.

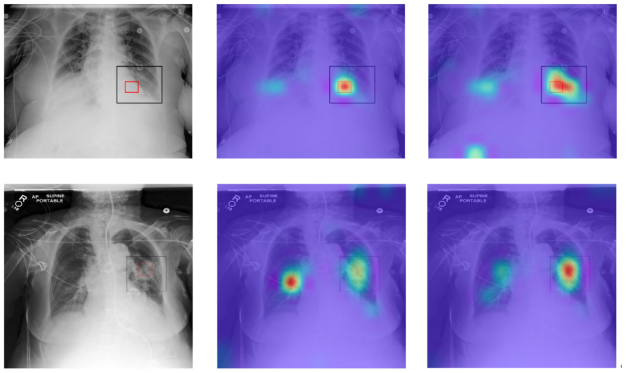
Temporal Sentence Similarity Classification Results. In Table 2, we report our performance on the temporal text classification task. Our method achieves the best results in both accuracy and AUROC metrics. It surpasses BioViL-T by 4.95% in accuracy and by 3.21% in the AUROC metric. The BioViL-T method not only utilizes temporal information but also performs specialized mask language modeling. Additionally, the RadGraph dataset is a relatively complex text classification task, hence our results in this experiment thoroughly demonstrate the effectiveness of our approach. Through temporal modeling, our method acquires the capability to perceive changes, and with fine-grained local alignment, it more comprehensively understands the semantic context of the text.

Zero-shot Image Classification Results. Table 3 shows our results of zero-shot medical image classification task on RSNA dataset. Our method also achieves the best performance. This indicates that our performance improvement is not limited to temporal tasks but also benefits static classification tasks. Our model outperforms previous methods by 1%, 2%, and 1% in the accuracy, F1 score, and AUROC metrics, respectively. These consistent improvements across metrics suggest that spatial and temporal modeling can enhance the model's representational capabilities. Since the logits in zero-shot classification are based on the similarity between image features and text prompts, it also indicates that our image and text encoders have learned more fine-grained and consistent representations.

Medical Image Classification Results. We present our classification results on COVIDx in Table 4. Our method achieved commendable results using both 10% and 100% of the training data. The outcomes in the classification task also demonstrate that the gains from our approach are evident in static tasks. Because our model absorbs a wider range of supervision signals, it may fail to capture the relevant information with only 1% of the training data available.

4.4. Analysis of Our Framework

Ablation Study. Table 5 displays the ablation results of our various components on the temporal image classification task. It can be observed that the best results are obtained when all objectives are combined with the use of lateral views. The first two rows indicate the effectiveness of our temporal modeling and local alignment objectives. The comparison between the third and fourth rows highlights the efficacy of using lateral views, and when all objectives are combined, we can achieve further improvements.



(a) GT (b) Ours w/o LA (c) Ours

Figure 4: Attention visualization with reference patch (red box) in the bounding box (black box). LA stands for local alignment. The bounding box represents the main pathological area annotated manually. It can be observed that our method achieves better correspondence with the bounding box after local alignment.

Visualization. Figure 4 presents the attention visualization of our model based on the reference patch (red box). This patch, contained within the bounding box (black box), encapsulates the corresponding textual information. The text corresponding to the first row is "Left basilar consolidation seen" and the second row corresponds to "patchy consolidation in the mid left lung". It can be observed that our method, without local alignment, can only perceive certain areas. However, after applying modality-weighted local alignment, our method is able to recognize the pathological areas in the context, which correspond well with the bounding box. This demonstrates that our fine-grained modeling can indeed learn local alignment based on the amount of information contained in local pairs.

5. Conclusion

In this study, we present the Med-ST framework, which employs explicit spatial and temporal modeling along with comprehensive strategies for global and local alignment, and cross-modal bidirectional cycle consistency. This approach not only achieves superior performance in tasks including temporal sequences but also enhances the representation learning for medical image classification tasks. Med-ST emerges as a promising approach for medical multimodal pretraining, capitalizing on the extensive information present in multimodal datasets to acquire nuanced spatial and temporal insights.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China No. 62376277, No. 61976206 and No. 62222215, and Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098.

Impact Statement

Although the dataset we used contains temporal information, the specific dates are anonymized, eliminating the risk of privacy infringement. The Med-ST framework introduced in our paper models from both spatial and temporal perspectives, effectively simulating the behavior of clinicians in practice. We offer a promising approach that not only diagnoses abnormalities in chest radiographs but also analyzes historical data to make decisions. This can help reduce the workload on doctors and advance the development of medical diagnostics, making diagnosis more intelligent and accessible, especially in underdeveloped areas.

References

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Alsentzer, E., Murphy, J. R., Boag, W., Weng, W.-H., Jin, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Bannur, S., Hyland, S., Liu, Q., Perez-Garcia, F., Ilse, M., Castro, D. C., Boecking, B., Sharma, H., Bouzid, K., Thieme, A., et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15016–15027, 2023.

Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O. K., Aggarwal, K., Som, S., Piao, S., and Wei, F. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

Boecking, B., Usuyama, N., Bannur, S., Castro, D. C., Schwaighofer, A., Hyland, S., Wetscherek, M., Naumann, T., Nori, A., Alvarez-Valle, J., et al. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pp. 1–21. Springer, 2022.

Boser, B. E., Guyon, I. M., and Vapnik, V. N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pp. 144–152, 1992.

Chambon, P., Bluethgen, C., Delbrouck, J.-B., Van der Sluijs, R., Połacin, M., Chaves, J. M. Z., Abraham, T. M., Purohit, S., Langlotz, C. P., and Chaudhari, A. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022a.

Chambon, P., Bluethgen, C., Langlotz, C. P., and Chaudhari, A. Adapting pretrained vision-language foundational models to medical imaging domains. *arXiv preprint arXiv:2210.04133*, 2022b.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

Cheng, P., Lin, L., Lyu, J., Huang, Y., Luo, W., and Tang, X. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21361–21371, 2023.

Dawidowicz, G., Hirsch, E., and Tal, A. Limitr: Leveraging local information for medical image-text representation. *arXiv preprint arXiv:2303.11755*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P., and Zisserman, A. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1801–1810, 2019.

Huang, S.-C., Shen, L., Lungren, M. P., and Yeung, S. 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*, pp. 3942–3951, 2021.

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., and Hoi, S. C. H. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705, 2021.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Lu, H., Huo, Y., Ding, M., Fei, N., and Lu, Z. Cross-modal contrastive learning for generalizable and efficient image-text retrieval. *Machine Intelligence Research*, 20 (4):569–582, 2023.

Moon, J. H., Lee, H., Shin, W., Kim, Y.-H., and Choi, E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12): 6070–6080, 2022.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Qin, Z., Yi, H., Lao, Q., and Li, K. Medical image understanding with pretrained vision language models: A comprehensive study. *arXiv preprint arXiv:2209.15517*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Raoof, S., Feigin, D., Sung, A., Raoof, S., Irugulpati, L., and Rosenow III, E. C. Interpretation of plain chest roentgenogram. *Chest*, 141(2):545–558, 2012.

Shaib, C., Li, M. L., Joseph, S., Marshall, I. J., Li, J. J., and Wallace, B. C. Summarizing, simplifying, and synthesizing medical evidence using gpt-3 (with varying success). *arXiv preprint arXiv:2305.06299*, 2023.

Shih, G., Wu, C. C., Halabi, S. S., Kohli, M. D., Prevedello, L. M., Cook, T. S., Sharma, A., Amorosa, J. K., Arteaga, V., Galperin-Aizenberg, M., et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1):e180041, 2019.

Shu, C., Zhu, Y., Tang, X., Xiao, J., Chen, Y., Li, X., Zhang, Q., and Lu, Z. Miter: Medical image–text joint adaptive pretraining with multi-level contrastive learning. *Expert Systems with Applications*, 238:121526, 2024.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Wan, Z., Liu, C., Zhang, M., Fu, J., Wang, B., Cheng, S., Ma, L., Quilodrán-Casas, C., and Arcucci, R. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *arXiv preprint arXiv:2305.19894*, 2023.

Wang, F., Zhou, Y., Wang, S., Vardhanabhuti, V., and Yu, L. Multi-granularity cross-modal alignment for generalized medical visual representation learning. *arXiv preprint arXiv:2210.06044*, 2022a.

Wang, L., Lin, Z. Q., and Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific reports*, 10(1):19549, 2020.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022b.

Wang, Y., Zhao, Y., and Petzold, L. Are large language models ready for healthcare? a comparative study on clinical language understanding. *arXiv preprint arXiv:2304.05368*, 2023.

Wang, Z., Wu, Z., Agarwal, D., and Sun, J. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022c.

Wu, C., Zhang, X., Zhang, Y., Wang, Y., and Xie, W. Medklip: Medical knowledge enhanced language-image pretraining. *medRxiv*, pp. 2023–01, 2023.

Yunxiang, L., Zihan, L., Kai, Z., Ruilong, D., and You, Z. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge. *arXiv preprint arXiv:2303.14070*, 2023.

Zhang, L., Ruan, L., Hu, A., and Jin, Q. Multimodal pretraining from monolingual to multilingual. *Machine Intelligence Research*, 20(2):220–232, 2023.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Zhou, H.-Y., Chen, X., Zhang, Y., Luo, R., Wang, L., and Yu, Y. Generalized radiograph representation learning via cross-supervision between images and free-text radiology reports. *Nature Machine Intelligence*, 4(1):32–40, 2022.

Zhou, H.-Y., Lian, C., Wang, L., and Yu, Y. Advancing radiograph representation learning with masked record modeling. *arXiv preprint arXiv:2301.13155*, 2023.

A. Pre-training Details

We sort all cases of the same patient according to the studyDate attribute in chronological order. Then we divide them into sequences of length 4, treating any remaining part that is less than 4 as a sequence of length 1. In other words, our dataset contains sequences of either four image-text pairs or a single image-text pair. During training, a batch randomly includes sequences of length 4 or 1. Global alignment and local alignment use all image-text pairs in all sequences, while sequence modeling employs all sequences of length 4.

Table 6 outlines the hyperparameters used during pre-training. We opt for MGCA weights initialization, leveraging their high performance to ensure faster convergence, particularly aiming for time efficiency. We use PyTorch Lightning and an early stopping mechanism to optimize training duration.

Table 6: Hyperparameters of Pre-training

Hyperparameters	
Pre-training epochs	10
Batch size per GPU	10
Number of GPUs	2
Learning Rate	4e-5
Learning Rate Optimizer	CosineAnnealing
Weight Decay	0.05

B. Downstream Tasks

B.1. Temporal Image classification Similarity

B.1.1. Experimental Details

We split the dataset according to the annotated disease types, i.e., Consolidation, Edema, Pleural Effusion, Pneumonia, and Pneumothorax. And then, for each subtask, we first obtain the features of two images, concatenate them, and use an SVM for classification, employing cross-validation to get the accuracy for each category. The task is to categorize the concatenated feature into three stages of disease progression: `Improving, Stable, Worsening`. Subsequently, we calculate the average accuracy across five types of findings. We use publicly available pre-trained models from others and determine their accuracy on this task in the same manner.

B.2. Temporal Sentence Similarity

B.2.1. Experimental Details

The task has two subsets. The first subset, generated by RadGraph, identifies paraphrase and contradiction sentence pairs through analyzing graph representations of sentences. The second subset is created using the Swaps method, which involves changing temporal keywords within sentences to produce paraphrases and contradictions. Swaps approach creates pairs that differ subtly in their temporal semantics while pairs obtained through the RadGraph method exhibit greater variation. Table B.2 are two examples of sentence pairs from each subset. We divide the dataset into two subsets based on the construction of text pairs into RadGraph and Swaps and conduct tests on each subset. Then, firstly, we randomize the dataset and then obtain the AUROC results based on the similarity between text pairs. Subsequently, through ten-fold cross-validation, we select the similarity threshold corresponding to the highest accuracy on the validation set. We then test on the entire set to obtain the accuracy. We use publicly available pre-trained models from others and follow the same process to determine their performance on this task.

B.2.2. Results

Tables 2 and Table 8 respectively show the experimental results on the two subsets. Our method achieves better results on the more complex dataset. For the Swaps dataset, methods BioViL and BioViL-T achieve better results due to the use of masked language modeling.

Table 7: Temporal image classification results on the MS-CXR-T benchmark under 5-fold cross-validation setting. MLM stands for masked language modeling objective. Best results are highlighted in bold.

Model	Consolidation	Edema	Pl.effusion	Pneumonia	Pneumothorax	**Avg. Acc.**
ConVIRT-MIMIC	46.97± 1.68	56.03± 1.07	54.98± 2.75	60.62± 2.51	46.61± 2.51	53.04± 0.62
GLoRIA-CheXpert	50.10± 1.94	51.63± 0.46	51.09± 0.72	60.49± 1.05	44.10± 1.18	51.48± 0.84
GLoRIA-MIMIC	44.32± 2.51	49.64± 1.74	48.74± 1.69	58.63± 2.52	42.17± 1.66	48.70± 0.86
MGCA	52.43± 1.82	63.30± 1.07	56.13± 1.97	64.83± 0.21	52.78± 1.37	57.90± 0.30
MedCLIP	49.44± 2.51	52.27± 3.98	53.79± 0.59	58.92± 1.43	41.87± 2.38	51.25± 0.65
MRM	50.10± 1.16	49.37± 3.84	48.97± 0.61	55.12± 1.09	48.23± 0.47	50.36± 1.01
BioViL	57.56± 1.17	57.27± 0.72	54.10± 0.65	66.95± 0.53	54.82± 0.22	58.14± 0.41
Temporal-based						
BioViL-T	56.73± 1.76	59.52± 1.09	52.80± 0.90	**66.67**± 0.60	**55.61**± 0.23	58.26± 0.42
Med-ST	**60.72**± 2.02	**65.56**± 1.24	**57.99**± 0.30	64.98± 0.60	53.71± 1.83	**60.59**± 0.72

Table 8: Temporal sentence similarity classification results on Swaps subset.

Model	Accuracy	ROC-AUC
MedCLIP	59.59	48.33
MGCA	68.16	71.97
BioViL	81.63	90.00
BioViL-T	94.29	97.59
Med-ST	76.73	86.12

B.3. Zero-shot Image Classification

B.3.1. Text Prompt

We use the same text prompt as BioViL-T for `pneumonia`.

```
pos_query = [ 'Findings consistent with pneumonia', 'Findings suggesting
pneumonia', 'This opacity can represent pneumonia', 'Findings are most compatible
with pneumonia', ],
neg_query = [ 'There is no pneumonia', 'No evidence of pneumonia', 'No evidence
of acute pneumonia', 'No signs of pneumonia', ]
```

Table 9: Examples of sentence pairs from the MS-CXR-T temporal sentence similarity benchmark.

	Label	Sentence 1	Sentence 2
Swaps	Paraphrase	"Nearly resolved bilateral pleural effusions."	"Nearly cleared bilateral pleural effusions."
	Contradiction	"Bibasal atelectasis is unchanged."	"Bibasal atelectasis is new."
RadGraph	Paraphrase	"Moderate pulmonary edema is exaggerated by low lung volumes, but also worsened."	"Lung volumes are lower exaggerating what is at least worsened moderate pulmonary edema."
	Contradiction	"Small right pneumothorax has developed at the base of the right lung."	"Small pneumothorax at the base of the right lung is unchanged."

Table 10: Ablation study on the proportion of lateral views.

Lateral proportion (%)	Consolidation	Edema	Pl.effusion	Pneumonia	Pneumothorax	**Avg. Acc.**
0	55.09± 1.34	63.32± 0.78	58.06± 1.40	53.58± 0.40	50.25± 0.99	58.06± 0.29
60	55.39± 1.67	67.32± 0.88	58.63± 1.00	65.98± 0.91	51.04± 0.87	59.67± 0.41
100	60.57± 1.18	67.35± 0.32	58.47± 1.50	65.00± 0.34	54.18± 0.81	61.12± 0.34

Table 11: Ablation study on FMC and RMR.

$\mathcal{L}_{\text{FMC}}$	$\mathcal{L}_{\text{MRM}}$	Consolidation	Edema	Pl.effusion	Pneumonia	Pneumothorax	Avg. Acc.
✓		53.72	66.46	56.93	63.27	51.66	58.33
	✓	53.26	64.30	55.02	65.02	55.04	58.53
✓	✓	60.71	67.31	58.89	65.02	55.48	61.48

B.3.2. Experimental Details

We use the dataset split from previous work (Huang et al., 2021; Wang et al., 2022a) and test on the test set. The text embedding is the average of four prompt embeddings, thus we obtain two text features representing the presence and absence of pneumonia. After obtaining the image features, we determine the predicted category based on the similarity between the image features and the two text features.

C. More Visualization

C.1. t-SNE visualizations results of image features

Figure 5 shows the visualization results of image features at different time steps for 200 sequences. The left side is from MGCA, and the right side is ours. Each category represents different time points in the sequence. From the figure, it can be seen that the feature distribution of MGCA is relatively uniform, with no discernible trend of change. In contrast, the trends between the four time points in our diagram are consistent, indicating that our model has learned time-related information.

C.2. Visualization of the learned distribution of β

In Figure 6, we randomly select 800 sample points and obtained their β distributions before training (first row) and after training (second row). The first column represents the distribution of samples with a ground truth index of 1, while the second column represents those with a ground truth index of 3. The red dashed line indicates the ground truth, and the blue dashed line represents the mean of β of the samples. Our objective is to bring the blue dashed line closer to the red dashed line.

It can be observed that, compared to the row above, the β distributions obtained after training exhibit closer proximity to the ground truth, and they also demonstrate better fitting to Gaussian distributions. This suggests that our model has learned temporal semantics to a certain extent, as the features of samples closer in time are more similar.

D. More Ablation

D.1. Ablation study on the proportion of lateral views

Table 10 illustrates the results of ablation experiments using different proportions of lateral views. It can be observed that even without the use of lateral views, our mean performance is 58.06, surpassing nearly all baselines. This underscores the effectiveness of local alignment objectives and temporal modeling. As the proportion of lateral views increases, performance also improves, indicating that lateral views provide an additional boost to performance.

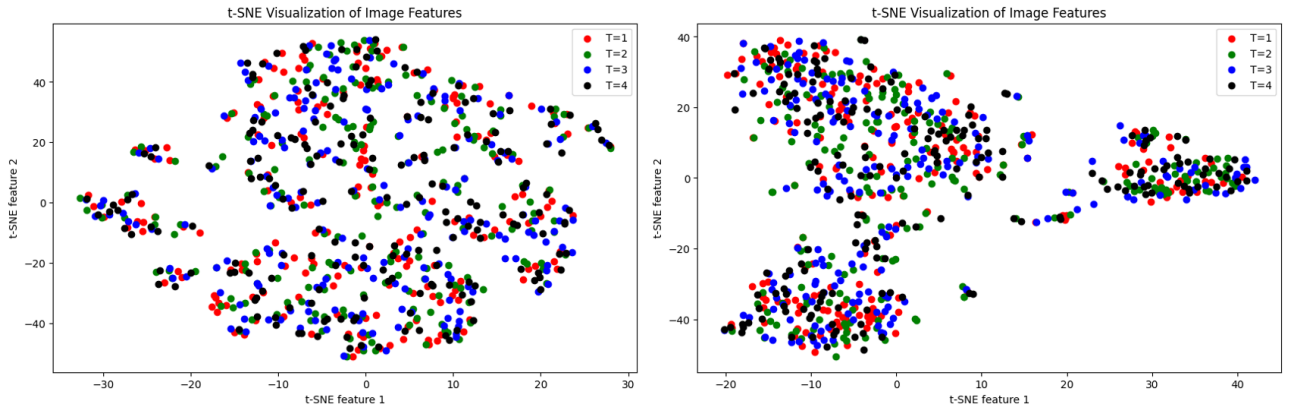


Figure 5: t-SNE visualizations results of image features at different time steps for 200 sequences. The left side is from MGCA, and the right side is ours.

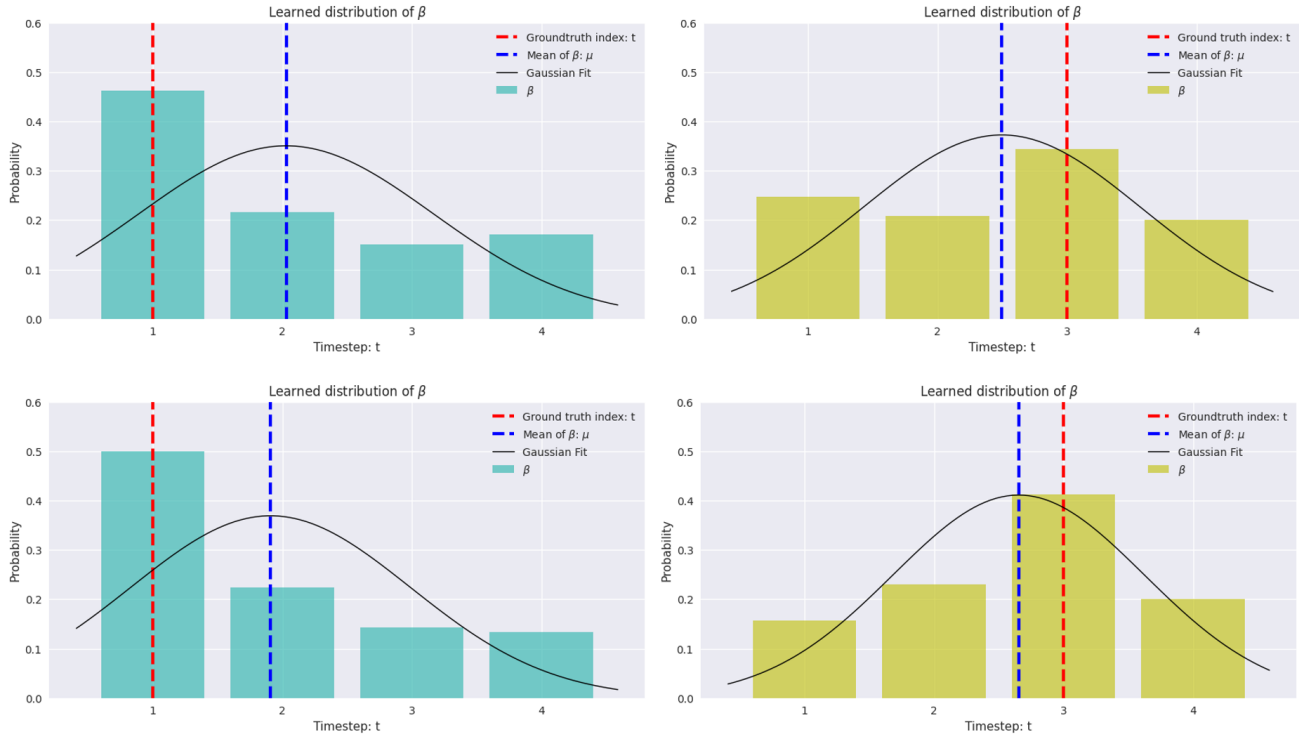


Figure 6: Visualization of the learned distribution of β.

D.2. Ablation study on FMC and RMR

Table 11 presents the ablation experiments conducted on FMC and MRM. The experimental results indicate that MRM contributes to greater benefits, suggesting that more complex regression objectives enable the model to better capture contextual information within sequences. When both are applied, the performance on sequential tasks is optimal, indicating that combining these two objectives, from simple to challenging, allows the model to maximize its understanding of sequence semantics.