Unity in Diversity: Collaborative Pre-training Across Multimodal Medical Sources

Xiaochen Wang[1], **Junyu Luo**[1], **Jiaqi Wang**[1], **Yuan Zhong**[1],
Xiaokun Zhang[2], **Yaqing Wang**[3], **Parminder Bhatia**[4], **Cao Xiao**[4], **Fenglong Ma**[1*]
[1]Pennsylvania State University, [2]Dalian University of Technology,
[3]Purdue University, [4]GE Healthcare
[1]{xcwang, junyu, jqwang, yfz5556, fenglong}@psu.edu, [2]dawnkun1993@gmail.com,
[3]wang5075@purdue.edu, [4]{parminder.bhatia, cao.xiao}@gehealthcare.com

Abstract

Although pre-training has become a prevalent approach for addressing various biomedical tasks, the current efficacy of pre-trained models is hindered by their reliance on a limited scope of medical sources. This limitation results in data scarcity during pre-training and restricts the range of applicable downstream tasks. In response to these challenges, we develop **Med**ical **C**ross-**S**ource **P**re-training (MEDCSP[1]), a new pre-training strategy designed to bridge the gap between multimodal medical sources. MEDCSP employs modality-level aggregation to unify patient data within individual sources. Additionally, leveraging temporal information and diagnosis history, MEDCSP effectively captures explicit and implicit correlations between patients across different sources. To evaluate the proposed strategy, we conduct comprehensive experiments, where the experiments are based on 6 modalities from 2 real-world medical data sources, and MEDCSP is evaluated on 4 tasks against 19 baselines, marking an initial yet essential step towards cross-source modeling in the medical domain.

1 Introduction

Pre-training, a widely adopted technique with the primary objective of enhancing the performance of downstream tasks, is a practice extensively employed in natural language processing (Kenton and Toutanova, 2019; Radford et al., 2018). In the medical domain, researchers have dedicated efforts in pretraining powerful models, including ClinicalBERT (Huang et al., 2019), ClinicalT5 (Lehman and Johnson, 2023), and MedHMP (Wang et al., 2023). While these pre-training techniques benefiting from unlabeled data have showcased superiority in diverse medical downstream tasks, they still suffer from the following challenges:

Data scarcity. Training a robust pre-trained model typically requires a substantial corpus, particularly in a multimodal approach. However, obtaining a sizable training dataset in the medical domain poses challenges owing to concerns surrounding data privacy. Hence, exploring innovative approaches for integrating more medical data into the pre-training process becomes imperative.

Limited downstream tasks. Current pre-trained models in the medical domain are often trained using data from a single source, thus limiting the spectrum of applicable downstream tasks. For example, MedHMP (Wang et al., 2023) pretrained on electronic health records (EHRs) source is only suitable for predictive modeling tasks involving EHR data. In contrast, pre-trained models in the general domain are usually applicable to various tasks. For instance, Flamingo (Alayrac et al., 2022), a visual-language model, achieves state-of-the-art performance on 16 few-shot learning tasks. Therefore, considering the multimodal nature of medical data, an ideal pre-trained model should be equipped to address as many tasks as possible.

To tackle these issues simultaneously, a promising approach involves training a model using diverse medical data sources from various datasets. This strategy not only augments the volume of training data but also broadens the spectrum of tasks. Nevertheless, achieving this objective is inherently challenging due to the following reasons:

Firstly, the number of patients who have data across multiple data sources is significantly limited. For example, this number is 14,620 between the MIMIC-IV and MIMIC-CXR databases, representing only 22.36% and 45.19% of these two databases, respectively. The scarcity of patients with data spanning multiple sources further diminishes the limited connectivity between these sources, adding complexity to cross-source integration efforts. Thus, designing an effective model that proficiently leverages the overlapped patients

[*] Corresponding author.
[1]Source codes are available at https://github.com/XiaochenWang-PSU/MedCSP.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3644–3656
August 11-16, 2024 ©2024 Association for Computational Linguistics

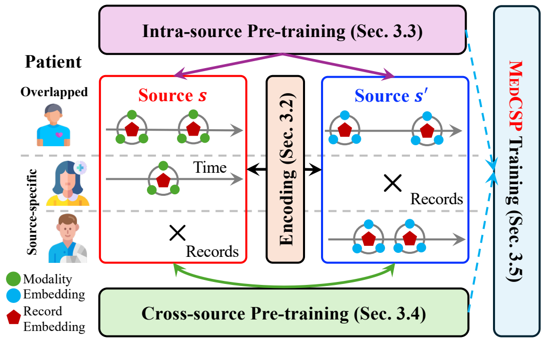


Figure 1: Overview of the proposed MEDCSP.

as a bridge to facilitate the training of other patients simultaneously is essential.

Secondly, modeling patients with data from multiple sources is challenging due to the intrinsic complexity of medical data. For example, a patient's chest X-ray images and reports may be housed in the MIMIC-CXR database, while diagnosis and treatment information are stored in the MIMIC-IV database. However, owing to the temporal nature of medical data, their recorded times of information across various sources may not align perfectly. Therefore, exploring a reasonable approach to model these relationships is urgently needed.

Finally, implicit yet informative relationships often exist among patients across different sources, requiring appropriate handling. For instance, patients with analogous conditions may exhibit similar symptoms, despite being in separate databases. Recognizing and leveraging these implicit connections is essential for facilitating cross-source training, as they hold significant potential for enhancing model performance through the aggregation of similar patient profiles.

To address the aforementioned challenges inherent in multimodal medical records from diverse sources, we introduce a pioneering pre-training framework in this paper, named **Med**ical **C**ross-**S**ource **P**re-training (MEDCSP), as shown in Figure 1. MEDCSP first encodes each modality from each source using modality-specific encoders in Section 3.2. Subsequently, it employs two distinct pre-training tasks. The first task explores modality-level relations among patients within individual sources (Section 3.3), while the second task focuses on discovering relationships among patients across different sources (Section 3.4). Specifically, MEDCSP models relations for overlapped patients across sources by considering their record times in Section 3.4.1 and establishes connections among patients in similar cohorts using their diagnosis similarities in Section 3.4.2.

Through interactive modeling, MEDCSP acquires the capability to generate informative and representative medical embeddings for diverse downstream tasks. Our exhaustive experiments across six modalities within two sources demonstrate the effectiveness of our pre-training strategy, providing an initial yet crucial solution for unifying diverse modalities across multiple medical sources.

2 Related Work

Multimodal Pre-training on Medical Data. Pre-training on multimodal medical data, although it has seen significant development in recent years, remains fragmented across various sources. The predominant approach to multimodal pre-training involves aligning images with text (Hervella et al., 2021, 2022a,b; Khare et al., 2021). Additionally, with the emergence of Large Language Models (LLMs), some pioneering studies have endeavored to integrate images into the semantic space of LLMs (Li et al., 2023; Moor et al., 2023). However, due to the constraints imposed by their pretraining data sources, applying these pretrained models to tasks devoid of images proves challenging.

Thus, research on pre-training with multimodal medical data excluding images remains relatively limited. Some researchers have achieved success by aligning numerical clinical features with diagnosis codes (Li et al., 2022, 2020), while others have explored the correlation between clinical language and codes. Recent advancements include modeling complex interactions within EHR data, incorporating multiple modalities such as diagnosis codes, demographics, clinical notes, medication codes, and clinical monitoring data (Meng et al., 2021; Wang et al., 2023). Nonetheless, these endeavors face challenges akin to those encountered in image-related pre-training, compounded by the issue of data scarcity within EHR data, which significantly restricts their broader applicability.

Multi-source Multimodal Pre-training. Conventional pre-training techniques typically leverage diverse data sources to enhance the generalizability of representations, a principle that extends to multimodal settings. Previous studies (Lu et al., 2019; Cho et al., 2021; Su et al., 2019; Lee et al., 2023a; Tan and Bansal, 2019) have demonstrated this by combining image-text pairs from multiple sources, thereby enhancing model performance across various domains. However, these models face limitations when confronted with new modalities, as they are built upon uniform data sources.

Recognizing the shortcomings of homogeneous multimodal pre-training approaches, recent endeavors have shifted focus towards harnessing more diverse and heterogeneous sources. Recent studies (Liang et al., 2022; Reed et al., 2022) have embraced data from varied modalities for joint pre-training, resulting in improved modality-specific encoders. Despite these advancements, designed to cater to general fields, these approaches struggle to capture latent medical correlations within multimodal health data, thereby impeding the generation of domain-specific representations.

3 Methodology

3.1 Model Input

Let $p \in \mathcal{P}$ represent a patient in the patient set $\mathcal{P}$. The patient's data may be distributed across multiple medical sources, as illustrated in Figure 1. We use $\mathcal{D}_s^p$ to represent data stored in the s-th source for patient p. Each patient's data from a source s may contain multiple records, i.e., $\mathcal{D}_s^p = \{\mathcal{D}_{s,r}^p\}_{r=1}^{R_s^p}$, where R_s^p represents the number of records in $\mathcal{D}_s^p$. In addition, each record usually consists of multimodal modalities. Let $\mathcal{D}_{s,r,m}^p$ denote the m-th modality in the r-th record from the s-th source for patient p. With these inputs, the subsequent step involves modality-level encoding.

3.2 Modality Encoding

Due to the significant differences among modalities in the data sources, employing a uniform encoder to embed them poses challenges. Hence, we adopt modality-specific encoders to map the modality-level data to a shared latent space, formulated as follows:

$$\mathbf{e}_{s,r,m}^p = \text{Encoder}_m(\mathcal{D}_{s,r,m}^p), \tag{1}$$

where the specifics of each encoder $\text{Encoder}_m(\cdot)$ are detailed in Appendix A. By averaging embeddings of modalities, we then obtain the record-level representation as follows:

$$\mathbf{c}_{s,r}^p = \frac{1}{M}\sum_{m=1}^{M} \mathbf{e}_{s,r,m}^p, \tag{2}$$

where M is the number of modalities.

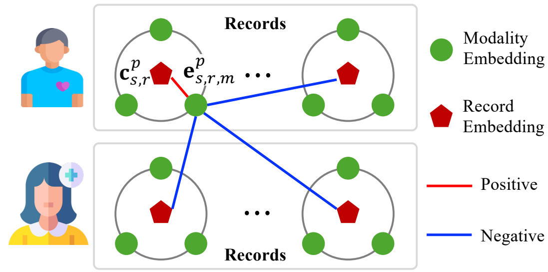


Figure 2: Illustration of intra-source pre-training.

3.3 Intra-source Pre-training

To conduct pre-training across multiple sources, the primary challenge lies in modeling the relationships among both modality-level and source-level data. Despite the different formats of modalities in sources, they inherently exhibit alignment. For instance, a chest X-ray image typically corresponds to a radiological report detailing the findings from the image, and a patient's in-hospital visit correlates with a set of diagnosis codes, procedure codes, clinical notes, and so forth. These alignments indicate that corresponding data in different modalities convey information about the same clinical event or patient admission. Consequently, it is imperative that these modality-level embeddings are mapped as closely as possible.

An ideal approach to capture the relationships among modalities is through pair-wise modality-level contrastive learning. However, the pair-wise learning paradigm encounters a drawback: the computational complexity escalates significantly with a large value of M. To address this challenge, we propose conducting record-modality-level contrastive learning. Intuitively, as illustrated in Eq. (2), the record representation $\mathbf{c}_{s,r}^p$ serves as the centroid of all modality-level representations. Ideally, each modality $\mathbf{e}_{s,r,m}^p$ should be proximate to its corresponding centroid $\mathbf{c}_{s,r}^p$ but distant from others, as shown in Figure 2.

Based on this intuition, we design our alignment-based loss. The loss is based on InfoNCE (Oord et al., 2018) and functions on a record-modality pair $(\mathbf{e}_{s,r,m}^p, \mathbf{c}_{s,r}^p)$ for intra-source pre-training as follows:

$$\mathcal{L}_a = -\log \frac{\exp(\text{sim}(\mathbf{e}_{s,r,m}^p, \mathbf{c}_{s,r}^p)/\tau)}{\sum_{\mathbf{c}_{s',r'}^{p'} \in \mathcal{N}_a} \exp(\text{sim}(\mathbf{e}_{s,r,m}^p, \mathbf{c}_{s',r'}^{p'})/\tau)}, \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity function, $\mathbf{c}_{s',r'}^{p'}$ denotes a randomly selected record within the batch, $\mathcal{N}_a$ denotes the negative set, and τ is a

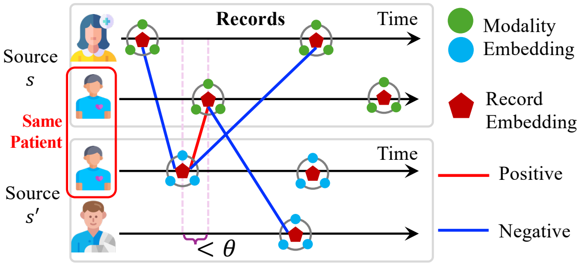


Figure 3: Illustration of pre-training for same patients across different sources.

temperature hyperparameter. Thus, the total alignment loss is defined based on Eq. (3) as follows:

$$\mathcal{L}_A = \sum_{p \in \mathcal{P}} \sum_{s=1}^{S^p} \sum_{r=1}^{R_s^p} \sum_{m=1}^{M_s} \mathcal{L}_a, \tag{4}$$

where S^p is the number of medical sources containing patient p's data, and M_s is the number of modalities within the source s.

3.4 Cross-source Pre-training

The training objectives outlined in Eq. (4) serve to direct the model in capturing explicit alignment between different modalities for the same patient within the same source comprehensively. However, an unresolved issue remains: *what if no explicit alignment is defined?* This issue becomes particularly prominent in cross-source settings, where distributed medical data often represent distinct admissions and studies. To address this issue, we propose two additional loss functions to capture relationships among patients across medical sources. The first loss term focuses on modeling relationships for patients present in different sources, while the second loss term aims to learn the relationships of patients in similar cohorts among different sources.

3.4.1 Same Patients Across Different Sources

Intuitively, the data of the same patient across different sources should exhibit similar patterns, particularly for records archived within the same time window. Let us assume that a patient's data in two distinct sources are denoted as $\mathcal{D}_{s,r}^p$ and $\mathcal{D}_{\hat{s},\hat{r}}^p$, and the timestamps of these two records satisfy $|T_{s,r}^p - T_{\hat{s},\hat{r}}^p| \leq \theta$, where θ represents a predefined time window threshold. The similarity between $\mathcal{D}_{s,r}^p$ and $\mathcal{D}_{\hat{s},\hat{r}}^p$ represented by $\text{sim}(\mathbf{c}_{s,r}^p, \mathbf{c}_{\hat{s},\hat{r}}^p)$ should be larger than that between $\mathcal{D}_{s,r}^p$ and a record $\mathcal{D}_{s',r'}^{p'}$ randomly selected from different sources, i.e, $s \neq s'$, if $p \neq p'$. As illustrated in Figure 3, we design a record-level cross-source contrastive learning loss for the same patients as follows:

$$\begin{aligned} \mathcal{L}_P &= \sum_{p \in \mathcal{P}} \sum_{s=1}^{S^p} \sum_{r=1}^{R_s^p} \mathcal{L}_p, \\ \mathcal{L}_p &= -\log \frac{\exp(\text{sim}(\mathbf{c}_{s,r}^p, \mathbf{c}_{\hat{s},\hat{r}}^p)/\tau)}{\sum_{\mathbf{c}_{s',r'}^{p'} \in \mathcal{N}_p} \exp(\text{sim}(\mathbf{c}_{s,r}^p, \mathbf{c}_{s',r'}^{p'})/\tau)}, \\ &s.t. \quad |T_{s,r}^p - T_{\hat{s},\hat{r}}^p| \leq \theta, \end{aligned} \tag{5}$$

where $\mathcal{N}_p$ is the randomly selected pairs.

3.4.2 Patients with Similar Cohorts Across Different Sources

In addition to within-patient interactions, as shown in Eq. (5), relationships between records that neither share the same patient nor belong to the same source still require appropriate analysis. When considering two data samples from distinct sources, denoted as $\mathcal{D}_{s,r}^p$ and $\mathcal{D}_{\hat{s},\hat{r}}^{\hat{p}}$, the absence of explicit similarity poses a challenge for understanding their relationship.

To ensure that no potential relationships across the medical domain are overlooked, we leverage diagnostic history from different patients to further capture implicit cross-source interactions. Intuitively, if $\mathcal{D}_{s,r}^p$ and $\mathcal{D}_{\hat{s},\hat{r}}^{\hat{p}}$ belong to patients with the same medical history — such as two patients both suffering from schizophrenia and bipolar disorder — the symptoms manifested through their records should exhibit similarity. Conversely, data without any overlap in diagnoses, i.e., $\mathcal{D}_{s,r}^p$ and $\mathcal{D}_{s',r'}^{p'}$, are unlikely to have similar recorded contents.

Let us denote the multi-hot vector representing all distinct diagnosis codes related to the patient p as $\mathbf{h}^p \in \mathbb{R}^{|\mathcal{H}|}$, where $|\mathcal{H}|$ denotes the distinct number of diagnosis codes. $\mathbf{h}^p$ serves as the diagnostic history of patient p. By calculating the cosine similarity between two patients, p and $\hat{p}$, we extend the definition of diagnostic similarity as follows:

$$\alpha_{p,\hat{p}} = \text{sim}(\mathbf{h}^p, \mathbf{h}^{\hat{p}}). \tag{6}$$

In Eqs. (3) and (5), we employ a strategy of directly choosing negative pairs with hard negative labels. This is because the positive labels exclusively originate from identical records (i.e., Eq.(3)) or patients (i.e., Eq. (5)). Consequently, pairs randomly selected in this manner exhibit a high confidence of being negative. Nevertheless, discerning positive and negative labels for similar cohorts drawn from distinct patients across diverse sources poses a challenge. To address this, we are prompted to

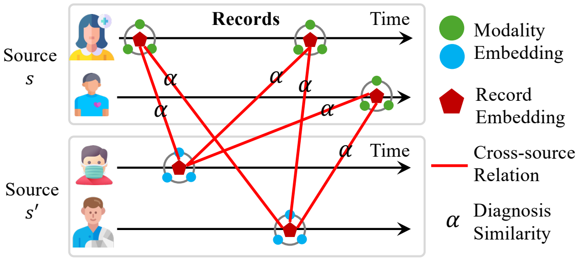


Figure 4: Illustration of pre-training for patients with similar cohorts across different sources.

adopt the similarity score calculated by Eq. (6) as a soft label for each pair, which contains more fine-grained information than a hard label.

As illustrated in Figure 4, given any cross-source pair $(\mathbf{c}_{s,r}^{p}, \mathbf{c}_{\hat{s},\hat{r}}^{p})$, we can derive its diagnostic relationship $\alpha_{p,\hat{p}}$ using Eq. (6). Building on previous research (Wu et al., 2021), we define the loss function designed for aggregating records associated with similar cohorts as follows:

$$\begin{aligned} \mathcal{L}_D &= \sum_{p\in\mathcal{P}}\sum_{s=1}^{S^p}\sum_{r=1}^{R_s^p}\mathcal{L}_d, \\ \mathcal{L}_d &= -\alpha_{p,\hat{p}}\log\frac{\exp(\mathrm{sim}(\mathbf{c}_{s,r}^{p},\mathbf{c}_{\hat{s},\hat{r}}^{\hat{p}})/\tau)}{\sum_{\mathbf{c}_{s',r'}^{p'}\in\mathcal{R}}\exp(\mathrm{sim}(\mathbf{c}_{s,r}^{p},\mathbf{c}_{s',r'}^{p'})/\tau)}, \end{aligned} \tag{7}$$

where $\mathcal{R}$ is the relation set across sources.

3.5 Training Objective of MEDCSP

The final pre-training objective of MEDCSP is the weighted summation of alignment-based, patient-based, and disease-based contrastive learning terms, expressed as:

$$\mathcal{L} = \mathcal{L}_A + \lambda_P\mathcal{L}_P + \lambda_D\mathcal{L}_D, \tag{8}$$

where λ_P and λ_D are two hyperparameters aiding in balancing the loss terms. This aggregated optimization objective balances intra- and cross-source modeling on health data, catering to sources with diverse cohorts and modalities. We will showcase the effectiveness of MEDCSP through numerous experiments in the subsequent sections.

4 Experiments

In this section, we first outline the configuration of our **pretraining process** in Section 4.1 and then demonstrate **evaluation with downstream tasks** on EHR source (Section 4.2) and medical image source (Section 4.3), respectively. Due to the space limitation, we put more experimental results in Appendix D and E.

4.1 Pretraining Setting

4.1.1 Datasets of Pretraining

For our pretraining, we engage with two distinct sources: the MIMIC-IV dataset (Johnson et al., 2023), which acts as a proxy for EHR data, and the MIMIC-CXR dataset (Johnson et al., 2019), which represents sources of medical imaging. These datasets span six modalities: text, images, temporal clinical data, demographics, diagnosis codes, and medication codes.

4.1.2 Data Processing

We adopt existing pipeline (Tang et al., 2020) for preprocessing EHR data. We follow existing work (Wang et al., 2023) to set the values of hyperparameters. To showcase the model's capability to manage non-overlapping cohorts across sources, we retain patients who do not appear in the MIMIC-CXR dataset. Regarding the CXR data source, we omit records from pre-training if their corresponding patients are not featured in the processed MIMIC-IV dataset, prioritizing efficiency. These excluded records are then utilized for zero-shot text-image retrieval tasks. This approach ensures the complete avoidance of any potential data leakage issues. From the pool of patients excluded from pre-training, we randomly select 1% and subsequently gather 1,202 records to form the evaluation set for the text-image retrieval task in Table 2. Comprehensive details on the datasets used for pretraining and fine-tuning across downstream tasks are provided in Table 1.

4.1.3 Implementation Details of Pretraining

We subject the introduced model to pretraining over 10 epochs with a learning rate of 1e-5. The batch size is configured at 128, optimized for the NVIDIA A100 GPU. Setups of modality-specific encoders are outlined in Appendix A. Throughout the training phase, we adjust all parameters, setting the balancing hyperparameters λ_P and λ_D to 0.5 and 0.2, respectively. Temperature hyperparameter τ is set to 0.1. Furthermore, we establish a time gap threshold θ of 30 days for the training process.

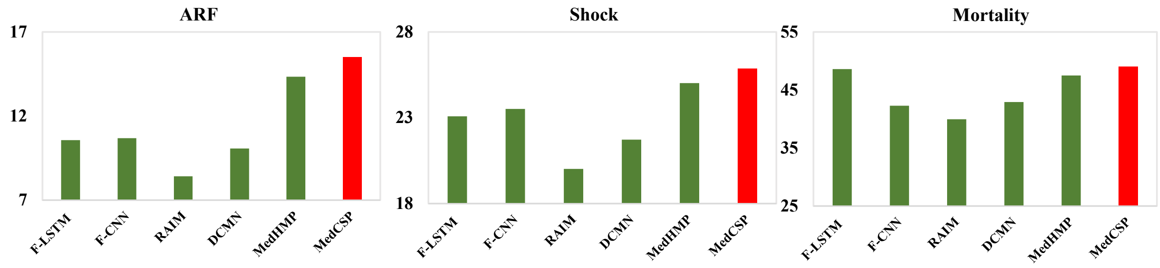
4.2 Evaluation on EHR Source

4.2.1 EHR Tasks

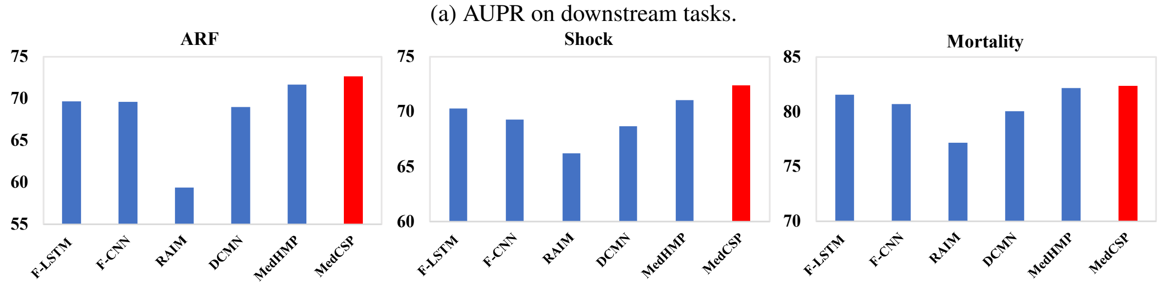
In-ICU Criticality Prediction. This experiment focuses on forecasting in-ICU activities by utilizing temporal clinical data and demographic information as inputs. We use three predictive tasks in this

Table 1: Data statistics.

Stage	Source	Dataset	# of Patients	# of Records		
Pretraining	EHR	MIMIC-IV	32,355	41,230		
	Medical Image	MIMIC-CXR	14,620	156,837		
	Source	**Dataset**	**Predictive Task**	**Total**	**Positive**	**Negative**
Downstream	EHR	MIMIC-III	ARF within 48 hours	5,038	402	4,636
			Shock within 48 hours	7,182	693	6,489
			Readmission within 30 days	11,695	1,581	10,114
	Medical Image	MIMIC-CXR	Image Text Retrieval	1,202	-	-
		COVID-19	Image Classification	13,808	3,616	10,192



(a) AUPR on downstream tasks.



(b) AUROC on downstream tasks.

Figure 5: In-ICU Criticality Prediction Tasks.

evaluation, including acute renal failure (ARF) prediction, shock prediction, and mortality prediction within a 48-hour window.

Readmission Prediction. The goal here is to forecast the likelihood of a patient's readmission within 30 days. This prediction's input includes temporal clinical data, clinical notes, demographic details, diagnosis codes, and medication codes.

4.2.2 Experimental Setups for EHR Tasks

The data used in the evaluation of these tasks are extracted from the MIMIC-III dataset (Johnson et al., 2016), which avoids the label leakage issue. We divide the dataset into training, validation, and testing subsets at an 80%/10%/10% split. The baselines include F-LSTM (Tang et al., 2020), F-CNN (Tang et al., 2020), RAIM (Xu et al., 2018), DCMN (Feng et al., 2019), and MedHMP (Wang et al., 2023) for the In-ICU Criticality Prediction task. For the Readmission Prediction task, we use eight multi-modal approaches present in existing study (Yang and Wu, 2021) and MedHMP as baselines. Note that only MedHMP and the proposed MEDCSP are pre-trained models. However, MedHMP uses both MIMIC-III and MIMIC-IV databases for the pre-training, while MEDCSP conducts the pre-training with MIMIC-IV and MIMIC-CXR databases. In other words, **MedHMP uses more training EHR data than MEDCSP for the EHR tasks.**

To evaluate the effectiveness of our pretrained encoder, we merge its output embeddings for each task and employ a Multi-layer Perceptron (MLP) module for task-specific classification. We determined the optimal learning rate and batch size through a grid search, with the batch size ranging from 32 to 256 and the learning rate from 2e-5 to 2e-2. We employ the area under the Precision-Recall (PR) curve (AUPR) and the area under the receiver operating characteristic curve (AUROC) as the evaluation metrics. The higher, the better. We obtain the final results as the mean values of five runs.

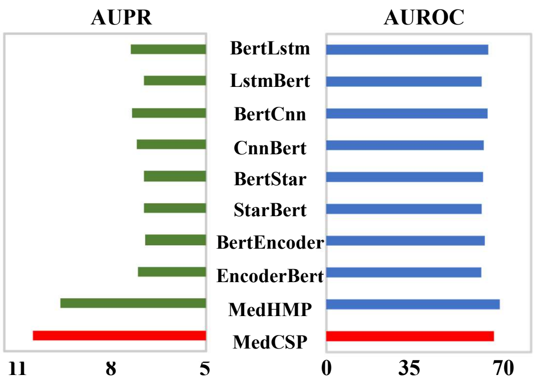


Figure 6: Results (%) of the readmission task.

4.2.3 Results of Evaluation on EHR Source

The experiment results on the in-ICU criticality prediction task are depicted in Figure 5. The pre-trained models, MedHMP and MEDCSP, outperform other non-pre-trained baselines. MEDCSP, pre-trained with less EHR data but taking the lead in all three tasks, demonstrates its superiority by using cross-sourced pre-training. This observation consolidates the correct rationale behind our design of a cross-source pre-training strategy. We can observe similar patterns for the readmission prediction task, as shown in Figure 6.

4.3 Evaluation on Radiological Source

One advantage of the proposed MEDCSP is to increase the diversity of downstream tasks by leveraging multi-source pre-training. To validate this advantage, we also conduct experiments to assess our model's proficiency in analyzing radiological images and the corresponding reports.

4.3.1 Radiological Tasks

Text-image Retrieval. This task assesses the model's ability to associate radiological images with corresponding textual descriptions correctly. It measures the model's comprehension of visual elements and textual information.

Zero-shot Image Classification. The model's accuracy in categorizing medical images into established categories without fine-tuning is evaluated. This ability is vital for healthcare applications and medical diagnostics, offering insights into the model's utility in real-world scenarios.

4.3.2 Radiological Datasets

We utilize a subset of the MIMIC-CXR dataset which is *NOT* included in the pretraining phase for the **text-image retrieval** task. Extra experiments on Open-I (Demner-Fushman et al., 2016) can be found in Appendix D. The text queries came from X-ray reports, and the corresponding X-ray images act as the ground truth for image candidates.

For **zero-shot image classification**, we use the COVID-19 dataset (Chowdhury et al., 2020; Rahman et al., 2021), consisting of COVID and non-COVID lung X-ray images, as the evaluation task. Additional experiments on CheXpert (Irvin et al., 2019) are covered in Appendix E.

4.3.3 Implementation Details

We maintain the original configuration settings for all CLIP-like baseline models, including MEDCSP. Specifically, for LLaVA-Med, we utilize models designed for pure text input to encode textual data. To process images, we employ a summarizing prompt in conjunction with the radiological image as input. The final aggregated outputs from these processes serve as the encoded embeddings for both text and image modalities. We then calculate the similarity between these modalities using the cosine distance metric, facilitating a comprehensive evaluation of the model's ability to bridge the gap between textual descriptions and visual representations. For the text-image retrieval task, we measure and report precision and recall at K scores, aligning with the methodologies established in previous studies, such as those detailed in (Wang et al., 2022) and (Zhang et al., 2023). In the image classification task, we document the precision, recall, and F1 score to evaluate model performance comprehensively.

4.3.4 Result Analysis

The findings from our Text-Image Retrieval task experiments, detailed in Table 2, indicate that our model, MEDCSP, significantly outshines CLIP-like baseline models with similar architecture in all assessed precision metrics at every k value. Impressively, MEDCSP even exceeds the performance of CXRCLIP (Lee et al., 2023b), another model pre-trained on the MIMIC-CXR dataset, evidencing the superior advantage of our multi-source pre-training approach. This advantage is particularly noteworthy because it suggests that our model's effectiveness is not merely due to its alignment with the test data's origin.

Similar to the results listed in Table 2, MEDCSP outperforms baselines on the zero-shot image classification task, as shown in Table 3. These observations highlight the robustness of MEDCSP's pre-training processes in forging strong correla-

Table 2: Results (%) on Text-image Retrieval Task

Methods	Precision @K						Recall @K					
	1	5	10	20	50	100	1	5	10	20	50	100
CLIP	0.17	0.18	0.17	0.13	0.14	0.12	0.08	0.67	1.16	1.75	4.63	7.79
MedCLIP	0.08	0.10	0.08	0.09	0.08	0.08	0.04	0.23	0.44	1.03	2.07	4.21
BiomedCLIP	0.50	0.53	0.43	0.39	0.31	0.26	0.46	2.29	3.49	5.89	11.79	18.73
PubMedCLIP	0.25	0.13	0.16	0.15	0.15	0.12	0.11	0.39	0.96	1.71	4.30	7.42
CXRCLIP	0.08	0.10	0.11	0.09	0.09	0.08	0.03	0.24	0.58	0.96	2.77	4.61
LLaVAMed	0.17	0.13	0.12	0.12	0.11	0.10	0.11	0.44	0.82	1.66	3.90	7.00
MEDCSP	**12.06**	**6.41**	**4.45**	**2.97**	**1.64**	**1.04**	**8.74**	**21.91**	**29.51**	**38.04**	**50.49**	**61.74**

Table 3: Performance(%) comparison of the zero-shot image classification task on the COVID-19 dataset.

Methods	Precision	Recall	F1
CLIP	26.01	64.91	37.14
MedCLIP	17.80	37.28	24.10
PubMedCLIP	66.67	0.11	0.22
BiomedCLIP	97.54	21.93	35.80
CXRCLIP	30.49	96.03	47.43
LLaVAMed	26.18	100.00	41.50
MEDCSP	71.98	55.00	**62.36**

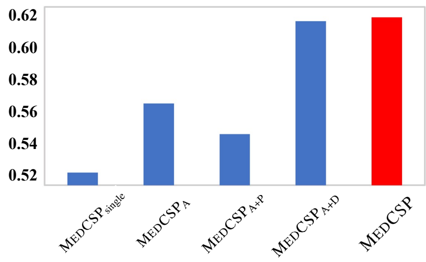


Figure 7: Results of ablation study. The X-axis denotes different settings, and the Y-axis represents the F1 score.

tions between image and text modalities. Consequently, MEDCSP emerges as a powerful asset for addressing complex medical issues, demonstrating its particular strength in the field of radiological image analysis.

4.4 Ablation Study

Our ablation study is conducted from two distinct angles: loss-wise and source-wise. This approach allows us to examine not just the impact of each individual loss term but also the benefits derived from a cross-source setting. For this purpose, we utilize the COVID-19 image classification task as a means to analyze the effectiveness of our pretraining strategy. Through this methodical examination, we aim to uncover the specific contributions of different loss components and the value added by leveraging diverse data sources to enhance our model's performance on a critical healthcare challenge. We report the F1 score of different settings in Figure 7.

Source-wise Comparison. MEDCSP$_{\text{single}}$, which represents the version of our model pretrained solely on the MIMIC-CXR dataset, exhibits a noticeable decrease in performance ($\downarrow$15.36%) compared to its multi-source pretrained counterpart, MEDCSP. This comparison starkly highlights the indispensable role of cross-source pretraining, demonstrating the substantial benefits that accrue from incorporating a variety of data sources to improve pretraining outcomes.

Loss-wise Comparison. Delving deeper into the architecture of our model, we introduce notations MEDCSP$_{\text{A}}$, MEDCSP$_{\text{A+P}}$, and, MEDCSP$_{\text{A+D}}$ to represent only keeping $\mathcal{L}_A$, $\mathcal{L}_A+\mathcal{L}_P$, and $\mathcal{L}_A+\mathcal{L}_D$ in the loss function (Eq. (8)), respectively. There are several observations: (1) The omission of any of these components results in a significant drop in performance metrics, emphasizing the essential contribution of each term to the model's comprehensive efficacy. (2) Only utilization of MEDCSP$_{\text{A+P}}$ causes the most significant drop ($\downarrow$11.56%) of the F1 score based on the experiment. The ablation study demonstrates the importance of disease-oriented metric learning terms. This analysis further elucidates the synergistic impact of these components in enhancing the model's ability to navigate complex medical data landscapes.

4.5 Case Study

To better illustrate how MEDCSP benefits from patient-centric and disease-oriented learning objectives, we visualize the case study performance in this section.

Patient-wise Modeling. The cross-source pretraining strategy aims to consolidate records from different sources linked by common patient identifiers. To evaluate the impact of alignment- and patient-centric training objectives, we focus on two

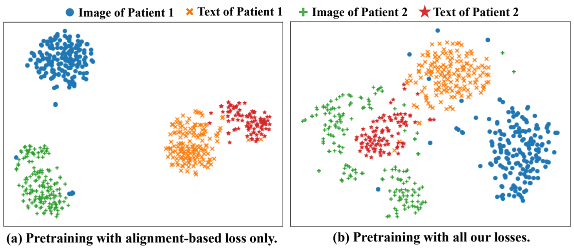


Figure 8: Case study on patient-wise modeling.

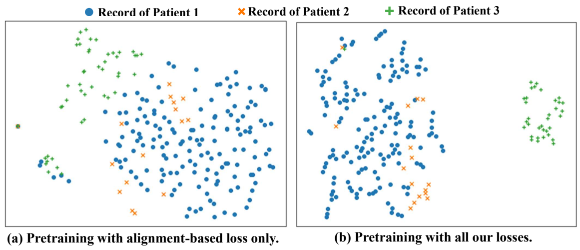


Figure 9: Case study on diagnosis-wise modeling.

patients with the highest number of radiological records in the MIMIC-CXR dataset. We then visualize the embeddings of both modalities, CXR images, and their corresponding reports to facilitate a detailed analysis, as presented in Figure 8. We can observe that data from the same patients are organized according to modality rather than patient identity when using solely the alignment-based loss function Eq. (4). This observation suggests the insufficient modeling of the unique latent medical patterns specific to each patient. In contrast, our model, designed to capture patient-level consistency, effectively clusters data by patient rather than by modality. This comparison vividly demonstrates that our model acquires an excellent understanding of patient latent medical patterns through the targeted design of our loss function.

Diagnosis-wise Modeling. We further explore our model's ability to forge connections between patients diagnosed with similar diseases. The analysis, illustrated in Figure 9, focuses on the record representations from three distinct patients. Patients 1 and 2 exhibit diagnostic similarities, whereas Patient 3 does not share any common diseases. Our findings reveal that when our model is pre-trained in the comprehensive setting, it effectively clusters records of patients with similar diagnoses. In contrast, when the model is pre-trained solely with the alignment-based loss, it faces challenges in forming consistent connections within disease-specific cohorts. This outcome underscores MEDCSP's proficiency in capturing the relationships between patients with similar diagnostic profiles, thereby generating meaningful representations for diverse downstream tasks.

5 Conclusion

This paper introduces a novel pre-training framework, MEDCSP, specifically designed for the complexities of diverse and highly heterogeneous medical sources. MEDCSP aggregates patient data within individual sources by aligning different modalities and subsequently captures patient relationships across multiple medical sources by leveraging temporal information and diagnosis history. Our experiments across a range of medical tasks and sources demonstrate that MEDCSP achieves superior performance. The observations are further supported by ablation studies and case analyses, underscoring the potential of MEDCSP in advancing medical cross-source modeling.

6 Ethic Consideration

The data utilized in our study have been appropriately de-identified according to Health Insurance Portability and Accountability Act (HIPAA) standards, which mandate the removal of all sensitive information as outlined in the HIPAA guidelines. As such, privacy concerns regarding the data we employ are mitigated. Additionally, the pretrained checkpoints of MEDCSP will be released following a thorough assessment of privacy, ethnicity, and security considerations.

7 Limitations

This study is constrained by computational resources, leading to the inclusion of only two medical databases during the pre-training phase. Recognizing the importance of diverse data for comprehensive learning, we aim to incorporate a wider array of medical sources in future research endeavors. Furthermore, we are considering an upgrade of our text encoding system by integrating advanced large language models (LLMs), as detailed in Appendix C. This strategic enhancement is expected to augment the learning capabilities of our framework, paving the way for more sophisticated analyses and applications in the medical domain.

Acknowledgements

This work is partially supported by the National Science Foundation under Grant No. 2238275 and the National Institutes of Health under Grant No. R01AG077016.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Muhammad EH Chowdhury, Tawsifur Rahman, Amith Khandakar, Rashid Mazhar, Muhammad Abdul Kadir, Zaid Bin Mahbub, Khandakar Reajul Islam, Muhammad Salman Khan, Atif Iqbal, Nasser Al Emadi, et al. 2020. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Sedigheh Eslami, Christoph Meinel, and Gerard De Melo. 2023. Pubmedclip: How much does clip benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1151–1163.

Yujuan Feng, Zhenxing Xu, Lin Gan, Ning Chen, Bin Yu, Ting Chen, and Fei Wang. 2019. Dcmn: Double core memory network for patient outcome prediction with multimodal data. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 200–209. IEEE.

Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. 2021. Self-supervised multimodal reconstruction pre-training for retinal computer-aided diagnosis. *Expert Systems with Applications*, 185:115598.

Alvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. 2022a. Multimodal image encoding pre-training for diabetic retinopathy grading. *Computers in Biology and Medicine*, 143:105302.

Álvaro S Hervella, José Rouco, Jorge Novo, and Marcos Ortega. 2022b. Retinal microaneurysms detection using adversarial pre-training with unlabeled multimodal images. *Information Fusion*, 79:146–161.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. 2021. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE.

Kwanhyung Lee, Soojeong Lee, Sangchul Hahn, Heejung Hyun, Edward Choi, Byungeun Ahn, and Joohyung Lee. 2023a. Learning missing modal electronic health records with unified multi-modal data embedding and modality-aware attention. *arXiv preprint arXiv:2305.02504*.

Seowoo Lee, Jiwon Youn, Mansu Kim, and Soon Ho Yoon. 2023b. Cxr-llava: Multimodal large language model for interpreting chest x-ray images. *arXiv preprint arXiv:2310.18341*.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv preprint arXiv:2306.00890*.

Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. 2022. Hi-behrt: Hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE Journal of Biomedical and Health Informatics*.

Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. 2020. Behrt: transformer for electronic health records. *Scientific reports*, 10(1):1–12.

Paul Pu Liang, Yiwei Lyu, Xiang Fan, Jeffrey Tsaw, Yudong Liu, Shentong Mo, Dani Yogatama, Louis-Philippe Morency, and Russ Salakhutdinov. 2022. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Yiwen Meng, William Speier, Michael K Ong, and Corey W Arnold. 2021. Bidirectional representation learning from transformers using multimodal electronic health record data to predict depression. *IEEE journal of biomedical and health informatics*, 25(8):3121–3129.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. 2023. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Obioma Pelka, Sven Koitka, Johannes Rückert, Felix Nensa, and Christoph M Friedrich. 2018. Radiology objects in context (roco): a multimodal image dataset. In *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings 3*, pages 180–189. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Tawsifur Rahman, Amith Khandakar, Yazan Qiblawey, Anas Tahir, Serkan Kiranyaz, Saad Bin Abul Kashem, Mohammad Tariqul Islam, Somaya Al Maadeed, Susu M Zughaier, Muhammad Salman Khan, et al. 2021. Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. *Computers in biology and medicine*, 132:104319.

Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gómez Colmenarejo, Alexander Novikov, Gabriel Barth-maron, Mai Giménez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *Transactions on Machine Learning Research*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111.

Shengpu Tang, Parmida Davarmanesh, Yanmeng Song, Danai Koutra, Michael W Sjoding, and Jenna Wiens. 2020. Democratizing ehr analyses with fiddle: a flexible data-driven preprocessing pipeline for structured clinical data. *Journal of the American Medical Informatics Association*, 27(12):1921–1934.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Xiaochen Wang, Junyu Luo, Jiaqi Wang, Ziyi Yin, Suhan Cui, Yuan Zhong, Yaqing Wang, and Fenglong Ma. 2023. Hierarchical pretraining on multimodal electronic health records. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2852.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887.

Bichen Wu, Ruizhe Cheng, Peizhao Zhang, Tianren Gao, Joseph E Gonzalez, and Peter Vajda. 2021. Data efficient language-supervised zero-shot recognition with optimal transport distillation. In *International Conference on Learning Representations*.

Yanbo Xu, Siddharth Biswal, Shriprasad R Deshpande, Kevin O Maher, and Jimeng Sun. 2018. Raim: Recurrent attentive and intensive model of multimodal patient monitoring data. In *Proceedings of the 24th ACM SIGKDD international conference on Knowledge Discovery & Data Mining*, pages 2565–2573.

Bo Yang and Lijun Wu. 2021. How to leverage the multimodal EHR data for better medical prediction? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4029–4038, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. 2023. Cxr-clip: Toward large scale chest x-ray language-image pre-training. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 101–111. Springer.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, et al. 2023. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*.

A Details of Encoders

As detailed in Section 3.2, we utilize modality-specific encoders to handle data from different modalities. Specific details about these encoders and the modalities they correspond to are provided in Table 4. For initialization, we employ Biomed-CLIP checkpoints for both the image and language encoders. Throughout the pretraining phase, the language encoders designated for clinical notes and radiological reports are set up to share parameters.

Table 4: Modalities leveraged in our experiments, along with their corresponding encoders.

Sources	Modalities	Encoders
EHR	ICD Codes	Multi-Layer Perceptron
	Drug Codes	Multi-Layer Perceptron
	Clinical Notes	PubMedBERT_256
	Demograhics	Multi-Layer Perceptron
	Clinical Temporal Readings	Long-short Term Memory
CXR	Radiological Images	VIT_base_patch16_224
	Radiological Reports	PubMedBERT_256

B Baselines

B.1 Baselines for EHR Tasks

The following multimodal approaches designed to handle clinical tasks serve as our baselines in EHR-related evaluation: **F-LSTM** (Tang et al., 2020) is a Long-short Term Memory (LSTM) architecture that processes inputs consisting of concatenated demographic and clinical temporal features. **F-CNN** (Tang et al., 2020) is a conventional Convolutional Neural Network (CNN) operating on the concatenation of clinical time series and demographic information for prediction on downstream tasks. **Raim** (Xu et al., 2018) is an advanced architecture engineered to process clinical information with a multi-channel attention mechanism. **DCMN** (Feng et al., 2019) is a combination of two distinct memory networks, with one focusing on temporal information and the other on static demographic data, allowing for comprehensive analysis. **MedHMP** (Wang et al., 2023) leverages a hierarchical pretraining strategy for boosting the model's performance in medical downstream tasks. Representations of modalities are aggregated through an attention mechanism for pretraining and fine-tuning. **BertLstm et al.** (Yang and Wu, 2021) contains different combinations of modality-specific encoders, including BERT, Star-Transformer, LSTM, and MLP. Multimodal representations are aggregated through summation for prediction tasks.

B.2 Baselines for Radiological Tasks

We adopt the following baselines for our evaluation of the radiological source: **CLIP** (Radford et al., 2021) is the backbone architecture developed by OpenAI. By performing contrastive learning between aligned image-text pairs, the model marks a significant step towards the unification of vision and language domains. **MedCLIP** (Wang et al., 2022) is pretrained on multiple datasets in a multi-tasking pattern.It relies on labeled images for extracting medical knowledge, thus performing contrastive learning without leveraging alignment between image and text. **BiomedCLIP** (Zhang et al., 2023) leverages PMC-15M dataset for deepening CLIP's adaptation in the biomedical domain, pretraining with InfoNCE loss (Radford et al., 2021). **PubMedCLIP** (Eslami et al., 2023) performs pairwise pretraining based on ROCO dataset (Pelka et al., 2018), following the conventional CLIP design. **CXRCLIP** (You et al., 2023) is pretrained on MIMIC-CXR dataset. The authors utilize contrastive learning loss between image and text, as well as multi-view of images, to achieve competitive performance. **LLaVAMed**(Li et al., 2023) is a multimodal Large Language Model (LLM) built upon pretrained LLaVA(Liu et al., 2023). It leverages the PMC-15M dataset for additional pretraining in a generative pattern.

We adopt image processors and tokenizers corresponding to each baseline for a fair comparison, as introduced in the original papers.

Table 5: Text-Image Retrieval Results (%) on the Open-I Dataset.

Methods	Precision @K						Recall @K					
	1	5	10	20	50	100	1	5	10	20	50	100
CLIP	0.03	0.05	0.04	0.04	0.04	0.04	0.03	0.13	0.21	0.45	1.12	2.08
MedCLIP	0.18	0.09	0.10	0.08	0.07	0.06	0.09	0.23	0.51	0.78	1.79	2.90
BiomedCLIP	0.21	0.10	0.14	0.11	0.09	0.08	0.12	0.26	0.70	1.10	2.24	3.95
PubMedCLIP	0.00	0.03	0.04	0.04	0.03	0.03	0.00	0.06	0.20	0.41	0.89	1.62
CXRCLIP	0.03	0.04	0.03	0.02	0.03	0.02	0.01	0.09	0.12	0.21	0.62	1.12
LLaVAMed	0.03	0.03	0.03	0.03	0.03	0.03	0.01	0.05	0.13	0.30	0.66	1.43
MEDCSP	**0.91**	**0.63**	**0.48**	**0.37**	**0.25**	**0.19**	**0.49**	**1.74**	**2.62**	**4.09**	**6.92**	**10.25**

C Implementing MEDCSP with Large Language Model (LLM)

Inspired by recent findings demonstrating the efficacy of applying LLM in the medical domain (Li et al., 2023), we sought to integrate our pre-training strategies with proficient LLM. Specifically, for modalities other than text, we employ the modality-specific encoders detailed in Table 4 to generate uniform embeddings, which are then concatenated with modality-specific prompts as input for LLaMA (Touvron et al., 2023). Textual contents are directly encoded alongside the prompt. Our pre-training approach, encompassing various data modalities, records, and patient information, aligns with the methodologies outlined in Sections 3.3, 3.4, and 3.5. We adopt the LLaVA-med model (Li et al., 2023) as the structural foundation for our exploration.

During this exploration, we solely fine-tune the projection layer between the frozen visual encoder and the fixed LLM, comprising only 3.15 million parameters. We observed a notable performance enhancement in the text-image retrieval task. This improvement is particularly evident when comparing our results to those obtained with LLaVA-med, achieving a significant increase in Recall@100 on the MIMIC-CXR dataset (16.19% versus 7.00%). This exploratory investigation underscores the effectiveness of MEDCSP's pre-training strategy and hints at its potential for integration with various backbone architectures.

D Extra Text-image Retrieval Results

Although we meticulously executed data splitting that absolutely eliminates data leakage concerns in our text-image retrieval task experiments, there might still be skepticism regarding whether MEDCSP truly surpasses baseline models on the MIMIC-CXR dataset, given its pre-training on the same source. To address this and demonstrate that the robust performance of MEDCSP is attributed to our strategically crafted pre-training approach, we conducted additional experiments on the Open-I dataset (Demner-Fushman et al., 2016). The results are presented in Table 5. Echoing the findings detailed in Table 2, MEDCSP consistently exceeds the performance of all comparison models across various metrics, further validating the effectiveness and soundness of our well-designed pre-training strategies.

Table 6: Performance(%) comparison of the zero-shot image classification task on the CheXpert dataset.

Methods	Precision	Recall	F1
CLIP	55.42	42.20	47.92
MedCLIP	31.52	26.61	28.86
PubMedCLIP	36.61	37.61	37.10
BiomedCLIP	68.42	11.92	20.31
CXRCLIP	42.11	44.04	43.05
LLaVAMed	46.58	100.00	63.56
MEDCSP	62.93	66.97	**64.89**

E Extra Experiments for Zero-shot Image Classification

To further demonstrate the effectiveness of MEDCSP in the zero-shot image classification task, we conducted experiments on the CheXpert dataset (Irvin et al., 2019). In these experiments, MEDCSP and baseline models are tasked with predicting the presence of an enlarged cardiomediastinum in images without any fine-tuning. The results are presented in Table 6. Consistent with our findings from the COVID-19 dataset experiments (Table 3), MEDCSP surpasses both CLIP-like baselines and the Large Language Model (LLaVAMed), underscoring its superior performance resulting from well-devised pretraining strategies.