MEDBind: Unifying Language and Multimodal Medical Data Embeddings

Yuan Gao[1,3,4,7,*], Sangwook Kim[1,3,7,*], David E Austin[1,7], and Chris McIntosh[1,2,3,4,5,6,7]

[1] Peter Munk Cardiac Centre, University Health Network (UHN), Toronto, Canada
[2] Department of Computer Science, University of Toronto (UofT), Toronto, Canada
[3] Department of Medical Biophysics, UofT, Toronto, Canada
[4] Ted Rogers Centre for Heart Research, UHN, Toronto, Canada
[5] Toronto General Hospital Research Institute, UHN, Toronto, Canada
[6] Department of Medical Imaging, UofT, Toronto, Canada
[7] Vector Institute, Toronto, Canada
{yuan.gao, sangwook.kim, chris.mcintosh}@uhn.ca

Abstract. Medical vision-language pretraining models (VLPM) have achieved remarkable progress in fusing chest X-rays (CXR) with clinical texts, introducing image-text data binding approaches that enable zero-shot learning and downstream clinical tasks. However, the current landscape lacks the holistic integration of additional medical modalities, such as electrocardiograms (ECG). We present MEDBind (**M**edical **E**lectronic patient recor**D** Bind), which learns joint embeddings across CXR, ECG, and text. Using text data as the central anchor, MEDBind features tri-modality binding, delivering competitive performance in top-K retrieval, zero-shot, and few-shot benchmarks against established VLPM, and the ability for CXR-to-ECG zero-shot classification and retrieval. This seamless integration is achieved by combining contrastive loss on modality-text pairs with our proposed contrastive loss function, Edge-Modality Contrastive Loss, fostering a cohesive embedding space for CXR, ECG, and text. Finally, we demonstrate that MEDBind can improve downstream tasks by directly integrating CXR and ECG embeddings into a large-language model for multimodal prompt tuning.

Keywords: Vision-Language Pretraining · Contrastive Learning · Multimodal Deep Learning · Self-Supervised Learning

1 Introduction

Vision-language pre-training models (VLPM) have advanced the integration of medical texts with imaging data, facilitating the convergence of diverse modalities into a unified representation space. This fusion deepens the understanding of text-image relationships and enhances their zero-shot learning capabilities.

VLPMs have revolutionized the interpretation of chest X-ray (CXR) by effectively aligning CXR with connected radiological reports through self-supervised

[*] Equal contribution.

contrastive learning. GloRIA [9] and BioVIL [2] showcased the potential to discern local and global visual features in CXR through textual analysis. Further, MedCLIP [29] and CXR-CLIP [30] elevated training efficacy by improving image-text specific loss functions. However, the scope of multimodal pre-training within the medical domain has been predominantly limited to image-text pairs, overlooking the potential integration of other clinical data types.

Incorporating more modalities from different domains is emerging as a critical research frontier. ImageBind [5] represents a significant stride in this direction by extending the VLPM contrastive learning approaches to accommodate more than two modalities within a unified embedding space, using images as the focal modality. ImageBind also broadened previous multimodal representations to additional tasks, including multimodal information retrieval and cross-modality zero-shot classification. Similarly, *all in one* [31] achieved alignment by integrating video and text into a transformer for joint feature extraction across different modalities. Med-PaLM M [26] recently advanced medical multimodal models by instruction prompt tuning PaLM-E, a large language model (LLM). Unlike contrastive learning approaches, Med-PaLM M incorporated multimodal data with text without explicit binding via LLM prompt tuning.

However, self-supervised contrastive learning in binding more than two medical modalities has yet to be explored. We present MEDBind (**M**edical **E**lectronic patient recor**D** Bind), a contrastive learning model that explicitly binds CXR, electrocardiograms (ECG), and medical texts into a unified embedding space. We chose text as the central anchor for binding CXR and ECG since many medical modalities are interpreted and given clinical narrative summaries.

Contributions: MEDBind is the first tri-modality framework that employs contrastive learning to fuse CXR, ECG, and medical texts into a unified representation space. We introduce a non-text edge-modality contrastive loss (EMCL), which strengthens the binding of CXR and ECG and is adept at handling varying numbers of cross-modality pairs in datasets. MEDBind pretrained with EMCL improves information retrieval, zero-shot, and few-shot performance. We utilize MEDBind in downstream clinical tasks, where ECG and CXR embeddings are integrated with LLM to predict readmission and in-hospital mortality.

2 Methods and Materials

2.1 Model Architecture

MEDBind is designed to process and analyze data from three distinct modalities: CXR, ECG, and medical text. Inspired by ImageBind [5], our model employs dedicated encoders for each modality to extract representations (Fig. 1). **Modality Encoder:** For the **CXR encoder**, we used Swin Transformer [19] as our backbone following [29, 30]. For the **ECG encoder**, we employed a vanilla transformer backbone [27]. We loaded ECG into a transformer by converting the time-series data into sequences of embeddings where each time point is tokenized using a linear embedding. We utilized BioBERT [17], a BERT [14] variant fine-tuned on medical texts to capture biomedical semantics for the **Text encoder**.

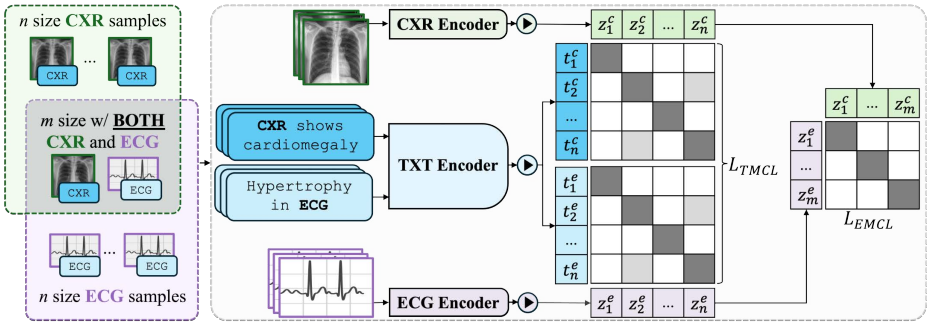

Fig. 1. Proposed method. Batch size n: CXR (green), ECG (purple), and paired text (blue). Subset size m: paired ECG/CXR. Inputs are embedded and normalized (▶). We used two losses: 1) Text-Modality Contrastive Loss (TMCL); 2) Edge-Modality Contrastive Loss (EMCL). Grey is positive-pair; light grey is additional related pairs.

We opted not to apply BioClinicalBert [1] for the text encoder to preserve the integrity of our training datasets for downstream tasks and to avoid potential bias since BioClinicalBert was fine-tuned on MIMIC-III. Finally, we used class token embeddings for all modality encoders because they are a critical component in transformer-based models that aggregate the global context of the input. **Projection/Normalization**: We projected and normalized CXR, ECG, and text embeddings to 256 dimensions using modality-specific linear layer and L2 normalization. This ensures final embeddings are comparable across encoders.

2.2 Loss function

We trained MEDBind using Text-Modality Contrastive Loss (TMCL) for text-modality binding and Edge-Modality Contrastive Loss (EMCL), a novel loss function we propose for improving cross-modality binding.

CLIP [24] showed that noise-contrastive estimation information (infoNCE) loss can bind image-text pairs, where a single positive-paired text is attracted for each image while the remaining texts are repelled. However, infoNCE loss does not account for cases where two patients have the same clinical text, incorrectly repelling their associated images. We implemented **TMCL**, similar to infoNCE, to link text with other modalities, but we considered identical paired texts as additional positive pairs. This is highlighted in the TMCL matrix of Fig. 1, where light grey pairs are additional related pairs with the same clinical text (for example, we encourage CXR with the same report to bind together).

Adopting Supervised Contrastive Learning (SCL) from [16, 21], we applied SCL for VLPM training in TMCL, reflected in Eq. 1, where z^j and t^j denote embeddings for the non-text modality and text, respectively with $j \in \{c : CXR,\ e : ECG\}$. We denote $i, l \in n$ as the i^{th}, l^{th} element in a batch size of n, yielding:

$$L_{TMCL}^{t^j \to z^j} = -\sum_{i=1}^{n} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(t_i^j \cdot z_p^j / \tau)}{\sum_{l=1}^{n} \exp(t_i^j \cdot z_l^j / \tau)} \tag{1}$$

where $p \in P(i)$ is the set of all positive pairs for text t_i^j and modality z^j. The temperature parameter τ modulates the scale of distribution over embeddings. We use a symmetric loss for L_{TMCL}, so $L_{TMCL} = \sum_{j \in \{c,e\}} (L_{TMCL}^{t^j \to z^j} + L_{TMCL}^{z^j \to t^j})$, where $t_j \to z_j$ denotes text-to-modality, e.g. text-to-ECG, and $z_j \to t_j$ denotes modality-to-text, thereby enforcing consistency in both modality directions.

We introduce **EMCL**, a novel contrastive loss that refines binding between non-text modalities. Unlike ImageBind [5], EMCL explicitly binds CXR to ECG and can dynamically adapt to different CXR-ECG pair counts in a batch. We defined positive pairs of CXR-ECG when a patient's CXR and ECG are taken during the same clinical visit, which pairs non-text modalities at the patient and temporal level. We sub-sampled paired CXR-ECG instances from n to optimize the usage for all training data. Thus, not all samples have corresponding CXR-ECG pairs (Fig. 1). Note that sub-sampling leads to a varying subset of size m in each batch. We define EMCL as:

$$L_{EMCL}^{z^c \to z^e} = -\sum_{u=1}^{m} \log \frac{\exp(z_u^c \cdot z_u^e / \tau)}{\frac{n}{m} \sum_{q=1}^{m} \exp(z_u^c \cdot z_q^e / \tau)} \tag{2}$$

where $u, q \in m$ are the u^{th}, q^{th} element of the batch, and the embeddings of CXR, z^c, and ECG, z^e, from the same patient case, are aligned to a unified embedding space. EMCL stabilizes the fluctuating cardinality of this subset by normalizing the denominator with a factor of $\frac{n}{m}$ across different batch iterations. Similar to TMCL, we employed symmetric loss on L_{EMCL} for bidirectional consistency, where $L_{EMCL} = L_{EMCL}^{z_c \to z_e} + L_{EMCL}^{z_e \to z_c}$. Our overall loss is defined as $L_{TMCL} + L_{EMCL}$, where L_{EMCL} equips MEDBind for cross-modality binding.

2.3 ECG-CLIP and Tri-modality Evaluations

To our knowledge, no VLPM has bound ECG and text, making direct comparisons with existing models challenging. Thus, we devised a novel **ECG-CLIP** as a baseline using our ECG and text encoders, trained with L_{TMCL} where $j = e$.

To assess the impact of tri-modality binding and EMCL, we introduce MEDBind; $\textbf{MEDBind}_{\textbf{BD}}$ (bound) with $L_{TMCL} + L_{EMCL}$ and $\textbf{MEDBind}_{\textbf{NM}}$ (normal) with only L_{TMCL} as an ablation. Moreover, we assessed if separately trained CXR and ECG VLPM could perform similarly to MEDBind, given that all VLPM bind the text modality. For tasks needing CXR and ECG encoders, we assessed various CXR VLPM paired with ECG-CLIP as the ECG encoder. This multiple single-paired VLPM approach is analogous to "**encoder zoo**" in [22].

Implementation Details: For training, we normalized CXR followed by augmentations [29]. We normalized ECG and applied Gaussian noise augmentation. Input dimensions were 224×224 for CXR and 12×1000 for ECG. For efficiency,

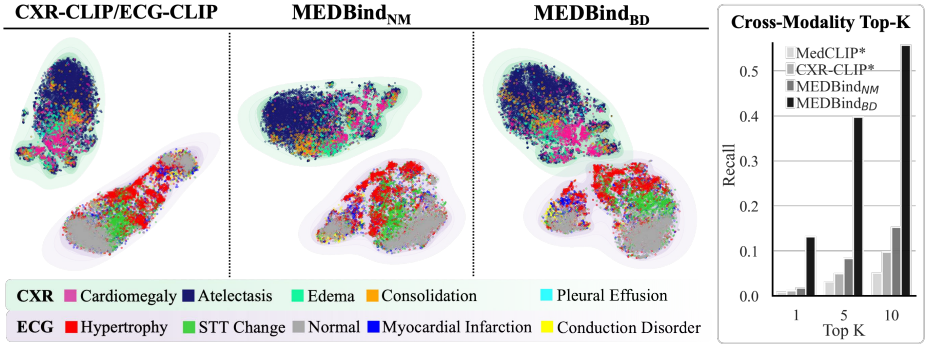

Fig. 2. Embedding visualization and CXR-to-ECG cross-modality retrieval. (Left) t-SNE plots of CXR and ECG embeddings for various models. (Right) Cross-modality retrieval Top-K recall. MEDBind$_{BD}$ brings CXR and ECG clusters closer in t-SNE and tops cross-modality recall@{1,5,10}. *CXR VLPM with ECG-CLIP as encoder zoo.

we truncated text to the first 100 words during pre-training, as 97% of CXR and ECG reports were under 100 words. We set the final embedding size to 256 and temperature τ to 0.07. We trained models for 150 epochs with batch size 128 and used AdamW [20] with weight decay 1e-1, learning rate 4e-4 adjusted via cosine annealing. We used PyTorch on an NVIDIA A100 GPU.

3 Experiments and Results

3.1 Datasets

We present datasets and details in Table 1. We pretrained MEDBind on MIMIC-CXR and MIMIC-ECG, including MIMIC-PAIR subset. To avoid training contamination, we maintained the same patient-level splits for all MIMIC datasets.

Starting with CXR datasets, **MIMIC-CXR** [13] consists of CXR with their paired reports and labels [6]. We pre-processed CXR and text using methods from MedCLIP [29]. In this study, we only included AP and PA view CXR. **CheXpert** [10] consists of a large number of CXR. Like [29, 30], we formed **CheXpert5x200** with 200 randomly selected CXR from 5 classes in [10]. We generated prompts for CXR-text retrieval tasks, as proposed in [29]. **COVID** [3] is a public dataset with binary COVID-19 labels. We generated prompts as suggested in [29] (details in Appendix). **RSNA** [25] contains pneumonia cases from CXR, publicly available in the National Institutes of Health database.

MIMIC-ECG [7] has 10-second 12-lead ECG at 500Hz, downsampled to 100Hz using a low-pass filter [15]. ECG has machine reports and links (`cart_id`) to free-form text. We used free-form text where available or machine reports to generate ECG text. We created labels (Hypertrophy, STT Change, Normal, Myocardial Infarction, Conduction Disorder) from text using a rule-based method inspired by [10], excluding ECG with undetectable labels (Appendix). **PTB-**

Table 1. Overview of datasets, tasks, class count (CLS), and training split. LINK shows connected datasets. *Subset from MIMIC-CXR and ECG with both CXR/ECG.

Dataset	Task	LINK	CLS	Train	Valid	Test
MIMIC-CXR [13]	Pretrain/Retrieval/LLM-Prompt	✓	12	86,853	12,059	24,799
Open-I [4]	Retrieval	✗	-	-	-	3,269
CheXpert [10]	Retrieval	✗	5	-	-	1,000
COVID [3]	Zero-Shot/Few-Shot	✗	2	11,028	-	2,780
RSNA [25]	Zero-Shot/Few-Shot	✗	2	18,678	-	5,338
MIMIC-ECG [7]	Pretrain/Retrieval/LLM-Prompt	✓	5	88,291	12,065	24,644
PTB-XL [28]	Retrieval/Zero-Shot/Few-Shot	✗	5	17,415	2,183	2,198
ICBEB [18]	Zero-Shot/Few-Shot	✗	9	5,501	-	1,376
MIMIC-IV [12]	LLM-Prompt	✓	-	218,787	40,995	72,473
MIMIC-PAIR*	Pretrain/Retrieval/LLM-Prompt	✓	-	22,397	3,292	6,664

XL [28] has 10 second 12-lead ECG at 100 Hz, and superclass labels [6, 28]. **ICBEB** [18] has 6-60 seconds 12 leads ECG at 100 Hz and class labels. We used the first 10 seconds for all ECG and zero-padded shorter ECG to 10 seconds.

MIMIC-IV [12] contains health records from patients in MIMIC, including discharge notes. We derived in-hospital mortality labels with `discharge_loc` and 30-day readmission from if a patient had a subsequent visit within a 30-day window, using patient's `subject_id` and admission time `admittime`. We linked CXR and ECG to MIMIC-IV by `subject_id` and modality recording times within 24 hours. Using this pairing strategy, we linked MIMIC-CXR and MIMIC-ECG, referred to as **MIMIC-PAIR**, by linking visit identifiers (`hadm_id`) in MIMIC-CXR and MIMIC-ECG if available. Without `hadm_id`, we paired cases on `subject_id` and if CXR and ECG recording times were within 24 hours.

3.2 Modality-to-Text and Cross-Modality Retrieval

Modality-to-Text Retrieval: MEDBind preserved retrieval integrity for CXR and ECG compared to single-paired VLPM. We tested CXR and ECG *modality*-to-text retrieval accuracy using recall for the top-K correct clinical reports. In Table 2, we benchmarked MEDBind against MedCLIP [29] and CXR-CLIP$_{SwinT}$ [30] for CXR-to-text retrieval, and ECG-CLIP for ECG-to-text retrieval. MEDBind outperformed all separately trained VLPM in total RSUM *modality*-to-text retrieval. However, in Open-I dataset, MedBind$_{\text{NM}}$ outperformed MedBind$_{\text{BD}}$, indicating that binding may lose some task depth in exchange for its breadth.

Cross-Modality Retrieval: In addition to *modality*-to-text retrieval, we also evaluated cross-modality retrieval between CXR and ECG on MIMIC-PAIR test set (Section 3.1) and compared to encoder zoo (CXR-CLIP or MedCLIP with ECG-CLIP). In Fig. 2, top-K recall for cross-modality retrieval highlights MEDBind$_{\text{BD}}$, with EMCL, outperforms MEDBind$_{\text{NM}}$ and other models. We visualize t-SNE plots to qualitatively prove that MEDBind$_{\text{BD}}$ brings CXR and ECG clusters closer within a joint space while maintaining class clustering, com-

Table 2. Results of CXR and ECG *modality*-to-text retrieval. Recall@K ={1,10} (R_K). C5x200 is CheXpert5x200. Total RSUM is sum of R_K per modality. **Bold**=best, underline=Second best. *data splits differed from CXR-CLIP, so results taken from [30].

Model	MIMIC-CXR		C5x200		Open-I		MIMIC-ECG		PTB-XL		Total RSUM	
	R_1	R_{10}	R_1	R_{10}	R_1	R_{10}	R_1	R_{10}	R_1	R_{10}	CXR	ECG
CLIP	1.0	10.5	1.1	13.9	0.8	8.1	-	-	-	-	35.4	-
MedCLIP	2.8	18.0	**2.9**	**31.2**	0.9	8.6	-	-	-	-	64.4	-
CXR-CLIP	21.6*	60.2*	2.2	20.3	14.1	39.3	-	-	-	-	157.7	-
ECG-CLIP	-	-	-	-	-	-	51.5	**95.5**	**2.1**	17.4	-	166.5
$MEDBind_{NM}$	43.8	88.5	2.4	22.1	**14.3**	**41.1**	50.2	93.9	1.9	18.2	**212.2**	164.2
$MEDBind_{BD}$	**44.7**	**91.0**	2.4	20.0	13.6	39.7	**53.6**	94.5	1.6	**19.2**	211.4	**168.9**

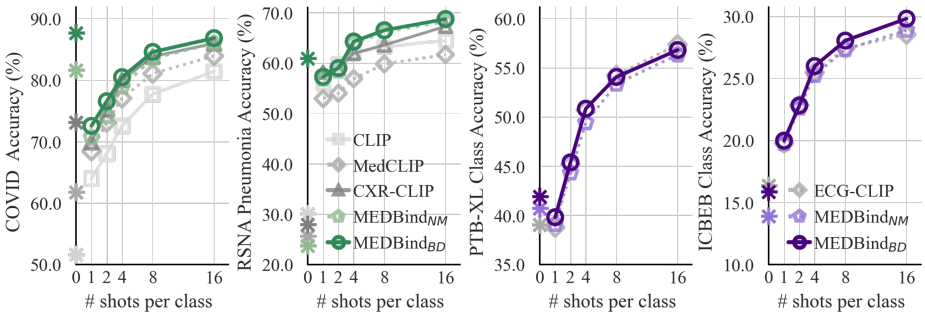

Fig. 3. Results of zero-shot (denoted as astericks (*) on y-axis) and few-shot (K={1,2,4,8,16}) classification using balanced accuracy (%) on CXR (green): COVID and RSNA datasets, and ECG (purple): PTB-XL and ICBEB datasets.

pared to other models. These results demonstrate the ability of $MEDBind_{BD}$ to match CXR to ECG and project modalities within a unified space.

3.3 Zero/Few-shot and Cross-Modality Classification

Modality-to-Text Zero/Few-shot: We evaluated zero-shot performance by calculating the cosine distance between text embeddings and non-text modality embeddings following [29]. Few-shot classification was assessed using embeddings from the frozen CXR or ECG encoders via the linear probing method [5]. We reported the average balanced accuracy for each shot over 300 different support sets. In Fig. 3, we compared MEDBind with other state-of-the-art models. In the zero-shot task, $MEDBind_{BD}$ beat $MEDBind_{NM}$ across all datasets and outperformed other models in three out of four datasets. $MEDBind_{BD}$ also maintained strong performance in all few-shot scenarios. These results demonstrate that EMCL boosted performance without compromising CXR or ECG zero and few-shot capabilities. Notably, $MEDBind_{BD}$'s zero-shot exceeded few-shot performance in the COVID dataset, highlighting its robustness on unseen classes.

Table 3. Cross-modality zero-shot performance, ACC refers to balanced accuracy (%).

Cross-Modality Zero-shot Task	Inputs	Query	Support	ACC
Hypertrophy vs. Other	CXR	MedCLIP	ECG-CLIP	60.9
Given CXR (query),		CXR-CLIP	ECG-CLIP	72.9
predict its ECG class using		$\text{MEDBind}_{\text{NM}}$	$\text{MEDBind}_{\text{NM}}$	<u>73.7</u>
ECG support set		$\text{MEDBind}_{\text{BD}}$	$\text{MEDBind}_{\text{BD}}$	**82.1**
Cardiomegaly vs. Other	ECG	ECG-CLIP	MedCLIP	<u>73.5</u>
Given ECG (query),		ECG-CLIP	CXR-CLIP	70.2
predict its CXR class using		$\text{MEDBind}_{\text{NM}}$	$\text{MEDBind}_{\text{NM}}$	69.1
CXR support set		$\text{MEDBind}_{\text{BD}}$	$\text{MEDBind}_{\text{BD}}$	**84.6**

Table 4. LLM prompt tuning task. Inputs include: medical text (TXT_{MD}); CXR/ECG text (TXT_{C}/ TXT_{E}); CXR/ECG embedding (EMB_{C}/EMB_{E}). *discharge and admission text used for 30-day readmission (Readmit.) and in-hospital mortality (In Hosp.). **Bold** and <u>underline</u> denote best and second best mixed input models (✓), respectively.

Method	Mixed Input	CXR/ECG Interpreter	LLM Inputs	Readmit. ACC	In Hosp. ACC
Text-Only	✗	*Clinical Expert*	$\text{TXT}_{\text{MD}^*,\text{C,E}}$	65.0	74.5
Encoder zoo	✓	MedCLIP/ECG-CLIP	$\text{TXT}_{\text{MD}}{}^*$+$\text{EMB}_{\text{C,E}}$	60.5	72.0
	✓	CXR-CLIP/ECG-CLIP		59.9	71.6
MEDBind	✓	$\text{MEDBind}_{\text{NM}}$	$\text{TXT}_{\text{MD}}{}^*$+$\text{EMB}_{\text{C,E}}$	<u>60.5</u>	<u>73.6</u>
	✓	$\text{MEDBind}_{\text{BD}}$		**64.3**	**74.8**

Cross-Modality Zero-Shot: Cardiomegaly and hypertrophy are commonly diagnosed from CXR and ECG, respectively. Both diseases can manifest pathophysiological signs detectable in CXR and ECG [23]. Thus, we introduce a novel cross-modality zero-shot classification task, assessing if we can detect hypertrophy via CXR and cardiomegaly via ECG, on MIMIC-PAIR test set. We calculated cosine distances between query and support embeddings. For example, we used CXR as query with ECG as support to predict hypertrophy. In Table 3, results showed that $\text{MEDBind}_{\text{BD}}$ outperformed $\text{MEDBind}_{\text{NM}}$ and encoder zoo. $\text{MEDBind}_{\text{BD}}$'s strong cross-modality zero-shot performance implies its ability to integrate CXR and ECG into a unified space–a unique advantage of EMCL.

3.4 Multimodal LLM Integration

To assess the efficacy of MEDBind in integrating cross-modality data directly into an LLM, we conducted experiments on predicting 30-day hospital readmission and in-hospital mortality [11]. We used BioBERT due to its compatibility with BERT-based VLPM. While not generative, BioBERT, as a non-casual masked model, is well suited for classification [11]. Using MIMIC-IV, we provided BioBERT with discharge summaries for readmission and patient demographics notes for mortality predictions (Appendix). We excluded discharge texts to prevent bias in mortality information. If CXR or ECG were connected to a clinical visit, we provided the following inputs after the text: 1) CXR and ECG

clinical interpretation from experts (**Text-only**), 2) embeddings from **encoder zoo**, or 3) embeddings from **MEDBind**. We extracted embeddings from frozen non-text modality encoders with a trainable linear projection layer for LLM integration following [22]. We used Low-Rank Adaption for efficient LLM prompt-tuning [8]. Table 4 highlights the performance of $\text{MEDBind}_{\text{BD}}$ for prompt-tuning LLM, BioBERT, compared to $\text{MEDBind}_{\text{NM}}$ and encoder zoo. $\text{MEDBind}_{\text{BD}}$ outperforms its counterparts by binding CXR and ECG pairs using our proposed EMCL. While the text-only LLM performs similarly on downstream tasks, it relies on clinician-generated texts. Instead, $\text{MEDBind}_{\text{BD}}$ is more automated as it can directly process CXR and ECG–increasing clinical workflow efficiency.

4 Conclusion

We introduced MEDBind, a tri-modality binding framework integrating multimodal medical data of CXR, ECG, and text. We demonstrated its benefits in binding different modalities into a unified space via EMCL, which enhanced zero-shot and downstream task performance over single-paired VLPM. Our method is scalable and open for future expansion to include additional modalities.

Acknowledgments. Study was funded by NSERC RGPIN-2022-05117. CM holds the Chair in Medical Imaging at the Joint Department of Medical Imaging (UHN & UofT).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Alsentzer, E., Murphy, J., Boag, W., Weng, W.H., Jindi, D., Naumann, T., McDermott, M.: Publicly available clinical bert embeddings. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop (2019)
2. 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: European conference on computer vision. pp. 1–21. Springer (2022)
3. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? Ieee Access **8**, 132665–132676 (2020)
4. Demner-Fushman, D., Kohli, M.D., Rosenman, M.B., Shooshan, S.E., Rodriguez, L., Antani, S., Thoma, G.R., McDonald, C.J.: Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association **23**(2), 304–310 (2016)
5. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
6. Goldberger, A.L., Amaral, L.A., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. circulation **101**(23), e215–e220 (2000)

7. Gow, B., Pollard, T., Nathanson, L.A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Berkowitz, S., Moukheiber, D., Eslami, P., et al.: Mimic-iv-ecg-diagnostic electrocardiogram matched subset (2023)
8. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2021)
9. Huang, S.C., Shen, L., Lungren, M.P., 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 (ICCV). pp. 3942–3951 (2021)
10. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
11. Jiang, L.Y., Liu, X.C., Nejatian, N.P., Nasir-Moin, M., Wang, D., Abidin, A., Eaton, K., Riina, H.A., Laufer, I., Punjabi, P., et al.: Health system-scale language models are all-purpose prediction engines. Nature pp. 1–6 (2023)
12. Johnson, A.E., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T.J., Hao, S., Moody, B., Gow, B., et al.: Mimic-iv, a freely accessible electronic health record dataset. Scientific data **10**(1), 1 (2023)
13. 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., Horng, S.: MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
14. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT. vol. 1, p. 2 (2019)
15. Kher, R., et al.: Signal processing techniques for removing noise from ecg signals. J. Biomed. Eng. Res **3**(101), 1–9 (2019)
16. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. Advances in neural information processing systems (NeurIPS) **33**, 18661–18673 (2020)
17. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)
18. Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al.: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics **8**(7), 1368–1373 (2018)
19. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF ICCV (2021)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2018)
21. Mo, S., Kim, M., Lee, K., Shin, J.: S-clip: Semi-supervised vision-language learning using few specialist captions. Advances in NeurIPS **36** (2024)
22. Moon, S., Madotto, A., Lin, Z., Nagarajan, T., Smith, M., Jain, S., Yeh, C.F., Murugesan, P., Heidari, P., Liu, Y., et al.: Anymal: An efficient and scalable any-modality augmented language model. arXiv preprint arXiv:2309.16058 (2023)
23. Nakamura, M., Sadoshima, J.: Mechanisms of physiological and pathological cardiac hypertrophy. Nature Reviews Cardiology **15**(7), 387–407 (2018)

24. 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)
25. Shih, G., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019)
26. Tu, T., Azizi, S., Driess, D., Schaekermann, M., Amin, M., Chang, P.C., Carroll, A., Lau, C., Tanno, R., Ktena, I., et al.: Towards generalist biomedical ai. NEJM AI **1**(3), AIoa2300138 (2024)
27. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
28. Wagner, P., Strodthoff, N., Bousseljot, R.D., Kreiseler, D., Lunze, F.I., Samek, W., Schaeffter, T.: Ptb-xl, a large publicly available electrocardiography dataset. Scientific data **7**(1), 154 (2020)
29. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. pp. 3876–3887 (2022)
30. You, K., Gu, J., Ham, J., Park, B., Kim, J., Hong, E.K., Baek, W., Roh, B.: Cxr-clip: Toward large scale chest x-ray language-image pre-training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2023)
31. Zhang, C., Sun, X., Yang, Y., Liu, L., Liu, Q., Zhou, X., Wang, Y.: All in one: Exploring unified vision-language tracking with multi-modal alignment. In: Proceedings of the 31st ACM Multimedia. pp. 5552–5561 (2023)