ProbMED: A Probabilistic Framework for Medical Multimodal Binding

Yuan Gao[1,2,3,5,6*] Sangwook Kim[1,3,4,5,6*] Jianzhong You[1,2,3,5,6] Chris McIntosh[1,2,3,4,5,6]

[1]Peter Munk Cardiac Centre [2]Ted Rogers Centre for Heart Research [3]University Health Network
[4]Joint Department of Medical Imaging [5]University of Toronto [6]Vector Institute

{yuan.gao, sangwook.kim, jianzhong.you, chris.mcintosh}@uhn.ca

Abstract

*Medical decision-making requires integrating diverse medical information, from imaging to clinical narratives. These medical modalities are often acquired in a many-to-many manner. However, current medical vision-language pretraining models (Med-VLPMs) fail to directly account for this many-to-many mapping in their model training and embeddings. To address this, we present **P**robabilistic **M**odality-**E**nhanced **D**iagnosis (ProbMED), a multimodal Med-VLPM that employs probabilistic contrastive learning to model distributions over embeddings rather than deterministic estimates. ProbMED aligns four distinct modalities—chest X-rays, electrocardiograms, echocardiograms, and clinical text—into a unified probabilistic embedding space. We use InfoNCE loss with Hellinger distance to integrate inter-modality distributions. We introduce a probabilistic synthetic sampling loss that captures modality-specific mean and variance to improve intra-modality binding. Extensive experiments across 13 medical datasets demonstrate that our model outperforms current Med-VLPMs in cross-modality retrieval, zero-shot, and few-shot classification. We also demonstrate the robust integration of multiple modalities for prognostication, showing improved intra- and inter-medical modality binding. Code is available: https://github.com/mcintoshML/probMED.*

1. Introduction

Medical decision-making is inherently multimodal and requires integrating diverse information ranging from imaging modalities to clinical reports. Despite the growing success of medical vision language pretraining models (Med-VLPM) in extracting embeddings from paired modalities, typically chest radiographs (CXR) with corresponding reports, these approaches operate primarily under a deterministic embedding framework that enforces one-to-one mappings [8, 16, 22, 50, 61, 62]. Existing approaches face two key limitations: **1)** Deterministic models may struggle to capture the inherent variability and complex many-to-many relationships in medical data, **2)** Majority of existing models focus exclusively on CXR-text alignment, overlooking broader multimodal nature of medical care.

Regarding the *first* problem, there are many medical cases where many-to-many relationships exist. For example, a CXR of a patient in respiratory distress can have multiple valid interpretations: "`A cloudy patch in the lower lung`" or "`CXR has pneumonia`". Although phrased differently, these examples inherently convey the same meaning, where one **describes** pneumonia and the other **states** the disease. Thus, these two examples should both relate to a pneumonia CXR, but not all cloudy patches are pneumonia (e.g., cloudy patches could be lung cancer). Next, consider a patient visit; they may require multiple electrocardiograms (ECGs) and CXRs to confirm prognosis–the relationships between ECGs and CXRs are inherently **many-to-many**. These examples highlight the limitations of deterministic methods that force embeddings into discrete positive/negative labels, making them ill-suited to modeling the ambiguity in the pairings. Recent advances in contrastive learning have highlighted the importance of capturing these relationships in learned representations [10]. Probabilistic contrastive learning extends traditional methods by modeling distributions over embeddings rather than single-point estimates [11, 30]; thus, each instance is represented as a distribution, which can enable better semantic overlap and *resolve* ambiguity during training.

For the *second* problem, real-world diagnostics integrate multiple modalities, for a more comprehensive clinical picture [16, 51]. However, as modalities multiply, cross-modal pairings grow quadratically, requiring a probabilistic contrastive approach. Thus, our approach unifies multiple medical data pairs through probabilistic contrastive learning by randomly sampling one modality pair per gradient update.

In this work, we introduce **Prob**abilistic **M**odality-**E**nhanced **D**iagnosis (PROBMED), a probabilistic multimodal Med-VLPM that bridges the gap between multiple medical modalities. This is the first study to leverage prob-

*Equal contribution to this work.

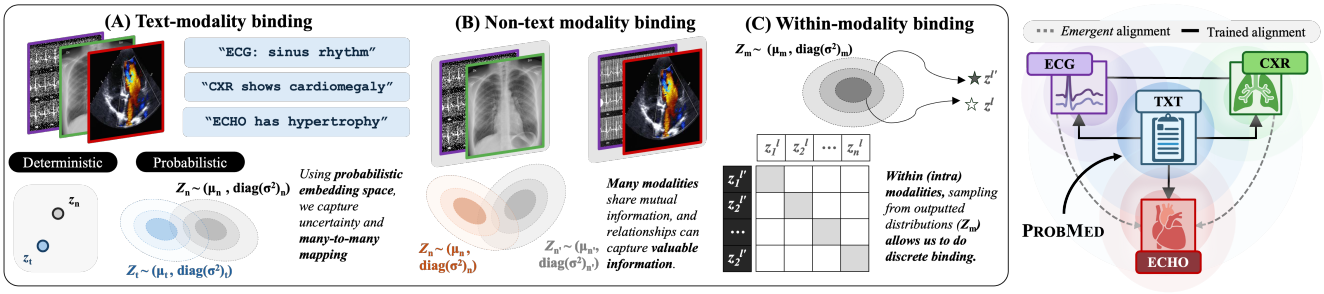

Figure 1. The overview of **PROBMED**. (Left) **(A) Text-modality:** Given multiple medical modalities, we align each with its corresponding textual description using a contrastive learning framework. Traditional *deterministic* approaches represent each modality-text pair as fixed points in latent space (z_n and z_t), whereas *probabilistic* ones model each modality embedding as a Gaussian distribution (Z_n and Z_t), detailed in §3.3. **(B) Non-text modality:** We also bind *between non-text modalities* and model overlapping related modalities, where n and n' are different non-text modalities (§3.3). **(C) Within-modality:** We introduce a Synthetic Instance Sampling Loss for improved *within*-modality binding. Given any modality Z_m, we use the learned distribution to sample additional samples z^l and $z^{l'}$, detailed in §3.4. (Right) **PROBMED** links ECG, CXR, ECHO, and text into a unified probabilistic embedding space, we trained (***black solid lines***) on all text-modality pairs and ECG-CXR (non-text modality) pairs. We observe *emergent* alignment (*grey dashed lines*) after training our model.

abilistic modeling for such extensive multimodal integration in the medical domain. Unlike prior approaches [10, 12], PROBMED uses a probabilistic contrastive framework to cross-integrate *four* distinct medical modalities: CXR, ECG, echocardiogram (ECHO), and text.

Contributions: 1) We propose **PROBMED**, which integrates multimodal medical data through probabilistic mappings. It aligns learned probability distributions across modalities within a shared embedding space while leveraging a novel loss function, Synthetic Instance Sampling Loss, to enhance intra-modal representation. **2)** We evaluated PROBMED on 13 distinct medical datasets, demonstrating superior performance in cross-modality retrieval, zero-shot, and few-shot classification. **3)** We further showed its multimodal capability by integrating CXR and ECG for improved prognostication of diseases often misdiagnosed using a single modality [43]. **4)** We showcased use-cases of the probabilistic modeling enabled by PROBMED: *Uncertainty-based prompt filtering* to enhance robustness against ambiguous data pairs, and *Distribution-based sampling* to improve classification in few-shot scenarios.

2. Related Work

Contrastive Learning. We focus on cross-modal contrastive learning, which integrates multimodal data like images and text [46]. Cross-modal learning has proven especially valuable in the medical domain [8, 50, 56, 61–63]. The objective is to maximize the similarity between modality pairs, e.g., aligning a CXR with its corresponding radiology report. These foundational approaches generate semantically rich representations that can be fine-tuned for downstream tasks, such as disease classification, even with limited annotated datasets. The ability to learn from data-scarce scenarios is a significant focus of this study since it is especially valuable in medical research, where acquiring large-scale annotated datasets is limited.

Probabilistic Multimodal Embeddings. Probabilistic contrastive learning enhances cross-modal learning in natural images by explicitly modeling the ambiguity in mapping visual features to textual descriptions [10, 11, 30, 52]. This ambiguity arises primarily from the prevalence of false-negative pairs during training [12]. Early methods, such as Probabilistic Cross-Modal Embedding (PCME) [11], introduced probabilistic embeddings to move beyond fixed representations but suffered from high computational costs and loss saturation. PCME++ [10] mitigates these challenges by introducing a closed-form probabilistic distance (CSD) and pseudo-positive samples with mixed-sample augmentations, enhancing learning of many-to-many relationships.

As described, the many-to-many challenge is also evident in the medical domain. Probabilistic embeddings can thus offer a dual benefit: they enable flexible cross-modal correspondences and explicitly capture uncertainty (via mean and variance), leading to richer representations for downstream tasks in data-scarce scenarios. Extending these models to bind more than three or more modalities remains an open challenge–one we address here.

Multimodal Learning. Multimodal learning extends many learning principles beyond simple image-text pairs to integrate additional modalities. ImageBind [17] introduced an effective way to integrate multiple modality pairs into a single unified model, improving the understanding of cross-modal mappings. Although the study was conducted on natural images, the potential of multimodal training in the medical domain is significant [39, 45, 47, 51, 58]. For instance, the use of medical data such as computed tomography [20, 60], sensor signals [16, 37], endoscopic videos [3], and even genomic information [59] are emerging. As such,

it can be seen that the multi-pair training paradigm capitalizes on each modality's complementary strengths, allowing for models that generalize across diverse clinical scenarios and tasks, especially for disease assessments that require examination of multiple modalities [43].

3. Methods

In this section, we introduce PROBMED, which learns a unified joint probabilistic embedding space for multiple medical modalities by utilizing all possible data pairs, with clinical notes to connect them (functioning as the primary binder). Here, each modality's embeddings are aligned with their corresponding text embeddings (e.g., CXR to radiology text) and/or across modalities (e.g., CXR to its corresponding ECG) as shown in Fig. 1. Inspired by ImageBind [17], PROBMED does not require complete pairs of modalities (that is, the four modalities of a single patient) during training, making it more practical for real-world datasets. We hypothesize that the resulting probabilistic embedding space across multiple modalities can better deal with the many-to-many mapping typically found in medical data. For the rest of this section, we explain how we trained PROBMED to integrate multiple medical modalities using a probabilistic approach.

3.1. Aligning Specific Pairs of Data.

Probabilistic contrastive learning is a technique for representing embeddings as distributions in an embedding space. Like traditional contrastive learning, the premise is based on using pairs of related examples (positives) and unrelated examples (negatives) to align data optimally. The key difference lies in the representation: instead of a fixed-point deterministic representation, the feature extractor outputs a distribution parameterized by mean, μ, and covariance matrix, $\mathbf{\Sigma}$. As proposed in [10], we stipulate that the covariance matrix is strictly the **diagonal covariance** matrix:

$$\mathbf{\Sigma} = \mathrm{diag}(\sigma^2), \tag{1}$$

where σ^2 represents the variance in each dimension. As such, we define our distributions as:

$$Z \sim \mathcal{N}(\mu,\, \mathrm{diag}(\sigma^2)), \tag{2}$$

where $\mu,\ \sigma^2 \in \mathbb{R}^D$ for a D-dimensional embedding. The pair similarities are computed using probabilistic similarities (§3.3), which optimize the joint embedding of different data pairs represented as distributions, allowing for greater flexibility to address many-to-many relationships better.

3.2. Model Framework

PROBMED processes four medical modalities—CXR, ECG, ECHO, and medical text—using dedicated modality-specific encoders inspired by [17]. Supplementary Fig. 4 presents the architecture overview; however, by design, the model framework is simple to implement. As in [10], our model relies on pretrained weights to facilitate the transition from deterministic to a probabilistic embedding space. Specifically, for the CXR encoder, we employ Swin Transformer backbone pretrained with ImageNet-1k [34, 38]. For the ECG encoder, we adopt an XResNet-1d-101, a widely used ECG encoder [49]. For the text encoder, we use BioBERT [31], a variant of BERT fine-tuned on biomedical corpora. Finally, for the ECHO encoder, we employ pre-trained ECHO-CLIP [8].

Generalized Pooling Operator. To generate probabilistic embeddings, we added two additional layers in parallel on top of each encoder, one for predicting the mean μ (initialized from the pre-trained backbone) and one for predicting $\log \sigma^2$ (randomly initialized). For computational efficiency, the $\log \sigma^2$ head is a single-layer network, and features are aggregated using Generalized Pooling Operator (GPO) following [5, 10].

Batch Normalization. Before GPO for feature aggregation, we incorporate Batch Normalization (BN). BN enforces a zero mean and unit variance across mini-batch, which may stabilize training and mitigate internal covariate shifts [23]. This may be beneficial in probabilistic settings, where BN could ensure estimated distribution parameters remain **less sensitive** to fluctuations in individual samples.

3.3. Binding Multimodal Probabilistic Embeddings

PROBMED integrates CXR–TEXT, ECG–TEXT, ECHO–TEXT, and CXR–ECG pairs in the MIMIC datasets to learn a unified joint embedding space, the available modality pairs is defined as $B \in \{(\mathrm{CXR}, \mathrm{TEXT}), (\mathrm{ECG}, \mathrm{TEXT}), (\mathrm{ECHO}, \mathrm{TEXT}), (\mathrm{CXR}, \mathrm{ECG})\}$. CXR–ECG was the only non-text modality pair we trained on as our training datasets had < 1000 samples of other non-text modality pairs. The model is trained using a meta-learning strategy akin to [17], where each gradient update is derived from a distinct objective for each available pair (described in §3.6).

We encode an input for each modality m into a probabilistic embedding using Eq. 2, as $\mu_m,\ \sigma_m \in \mathbb{R}^D$, and

$$Z_m \sim \mathcal{N}\Big(\mu_m,\, \mathrm{diag}(\sigma_m^2)\Big), \tag{3}$$

where $m \in \{\mathrm{CXR}, \mathrm{ECG}, \mathrm{ECHO}, \mathrm{TEXT}\}$.

Modality-Text Alignment. For modality-text alignment (e.g., CXR–TEXT), we use InfoNCE loss [41] on the probabilistic embeddings. Let q_n be the output of the **non-text** modality $n \in \{\mathrm{CXR}, \mathrm{ECG}, \mathrm{ECHO}\}$, and k_t be the output of the **corresponding paired** text. Based on Eq. (3):

$$q_n \sim \mathcal{N}\Big(\mu_n,\, \mathrm{diag}(\sigma_n^2)\Big) \text{ and } k_t \sim \mathcal{N}\Big(\mu_t,\, \mathrm{diag}(\sigma_t^2)\Big) \tag{4}$$

where t refers to the text representation of modality n. Then, given a batch of N modality–text pairs $\{(q_{n,i}, k_{t,i})\}_{i=1}^{N}$, we can calculate the InfoNCE loss as:

$$\mathcal{L}_{\text{MOD}_{n,t}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\big(-\text{PS}(q_{n,i}, k_{t,i})/\tau\big)}{\sum_{j=1}^{N}\exp\big(-\text{PS}(q_{n,i}, k_{t,j})/\tau\big)}, \tag{5}$$

where τ is the temperature and $\text{PS}(\cdot,\cdot)$ is a similarity function computing the similarity between two *probability distributions*. We use symmetric loss: $\mathcal{L}_{\text{MOD}_{n,t}} + \mathcal{L}_{\text{MOD}_{t,n}}$.

Non-text Modality Alignment. Where possible (Fig. 1), we enforced consistency across *non-text* modality pairs alignment, termed **non-text modality**. $\mathcal{L}_{\text{MOD}_{n,n'}}$ represents this alignment for $n \neq n'$ using the same equation as Eq. (5) by replacing t to n'. We trained PROBMED on ECG-CXR non-text modality (see §3.3).

Probabilistic Similarity Function. We adopt $1-H$, where H is the Hellinger distance [44], to measure the similarity between probabilistic embeddings because it is symmetric and bounded, making it well-suited for contrastive learning. The squared Hellinger distance between two multivariate Gaussian distributions [44], q_n and k_t, follows from Eq. (1), where the covariance matrices are $\text{diag}(\sigma_n^2) = \Sigma_n$ and $\text{diag}(\sigma_t^2) = \Sigma_t$. Then:

$$H^2(q_n, k_t) = 1 - \frac{\det(\boldsymbol{\Sigma}_n)^{\frac{1}{4}}\det(\boldsymbol{\Sigma}_t)^{\frac{1}{4}}}{\det\Big(\frac{\boldsymbol{\Sigma}_n+\boldsymbol{\Sigma}_t}{2}\Big)^{\frac{1}{2}}} \times \exp\Big(\big(-\frac{1}{8}(\mu_n-\mu_t)^\top\Big(\frac{\boldsymbol{\Sigma}_n+\boldsymbol{\Sigma}_t}{2}\Big)^{-1}(\mu_n-\mu_t)\Big). \tag{6}$$

We can then simplify the squared Hellinger distance as a product over the D dimensions:

$$H^2(q_n, k_t) = 1 - \prod_{o=1}^{D}\left[\left(\frac{2\sigma_{n,o}\sigma_{t,o}}{\sigma_{n,o}^2+\sigma_{t,o}^2}\right)^{\frac{1}{2}}\exp\left(-\frac{(\mu_{n,o}-\mu_{t,o})^2}{4(\sigma_{n,o}^2+\sigma_{t,o}^2)}\right)\right]. \tag{7}$$

Following this, we define the similarity measure as $\text{PS}(q_n, k_t) = 1 - \sqrt{H^2(q_n, k_t)}$, converting the squared Hellinger distance into a similarity metric. The details of calculating and computing Hellinger distance are described in the Supplementary §B.1. Unlike alternatives such as the PCME++ closed-form distance (CSD) and the Bhattacharyya distance, the Hellinger distance is symmetric and **bounded**, stabilizing training by mitigating excessive gradient magnitudes, an essential feature when handling noisy and uncertain nature of medical data. Furthermore, its sensitivity to differences in the means and variances of distributions enables it to capture subtle discrepancies between modalities. Empirically, our findings indicate that the Hellinger similarity facilitates superior convergence and creates a more discriminative joint embedding space.

3.4. Within-Modality Probabilistic Embeddings

To make the probability distributions of each modality robust we further propose **Synthetic Instance Sampling (SIS) Loss**. SIS loss encourages learning meaningful distributions by maximizing the similarity between two sampled instances from the same latent distributions. We adopt the reparameterization trick for multivariate Gaussian with a diagonal covariance structure from [29]. So, for an input with latent distribution (Eq. (2)) a sample is obtained:

$$z_m^l = \mu_m + \text{diag}(\sigma_m)\epsilon^l, \quad \epsilon^l \sim \mathcal{N}\big(0, \mathbf{I}\big), \quad l = \{1, \ldots, N_s\}, \tag{8}$$

where N_s is the number of sampled instances, $\mathbf{I}$ is a $D \times D$ identity matrix, and $\epsilon^l \in \mathbb{R}^D$. Note: σ_m is the modality, m, standard deviation. We reformulate InfoNCE [41] to operate on these sampled embedding instances. For each instance z_m^l and $z_m^{l'}$, where $l \neq l'$. Negative keys are drawn from other instances in the mini-batch. Thus, **SIS loss**:

$$\mathcal{L}_{\text{SIS}_m} = -\sum_{i=1}^{N}\log\frac{\exp\big(-\text{CS}(z_{m,i}^l, z_{m,i}^{l'})/\tau\big)}{\sum_{j=1,j\neq i}^{2N}\exp\big(-\text{CS}(z_{m,i}^l, z_{m,j}^l)/\tau\big)}, \tag{9}$$

where $\text{CS}(\cdot,\cdot)$ denotes the cosine similarity function. We present $\mathcal{L}_{\text{SIS}}$ for $N_s = 2$, which is analogous to SimCLR [6]. Thus, these two samples act as distinct instantiated *views* of the same distribution, reinforcing the distributions to maintain the meaningful variance. $\mathcal{L}_{\text{SIS}}$ encourages the model to learn distributions by enforcing consistency between samples drawn from the same underlying distribution. We used $\mathcal{L}_{\text{SIS}}$ on **all** modalities during pretraining.

3.5. Variational Information Bottleneck

To prevent the collapse of the variance, we use a Variational Information Bottleneck (VIB) loss [40] similar to [10]. VIB loss is computed as KL-divergence between learned latent distribution Z_m, Eq. (3) and a standard normal prior:

$$\mathcal{L}_{\text{VIB}_m} = -\frac{1}{N}\sum_{i=1}^{N}\text{KL}\Big(Z_{m,i}\,\Big\|\,\mathcal{N}(\mathbf{0}, \mathbf{I})\Big). \tag{10}$$

Empirically, VIB loss prevents variance collapse.

3.6. Final Loss Function

The overall loss is a weighted sum of multiple components. We randomly sample a modality pair from the available pairs for each gradient update. $\mathcal{L}_{\text{SIS}}$ and $\mathcal{L}_{\text{VIB}}$ are always included. So for $\{(m1, m2)\} \in B$, where B is the set of available modality pairs (defined in §3.3), then:

$$\mathcal{L}_{m1,m2} = \alpha(\mathcal{L}_{\text{MOD}_{m1,m2}} + \mathcal{L}_{\text{MOD}_{m2,m1}}) + \beta(\mathcal{L}_{\text{SIS}_{m1}} + \mathcal{L}_{\text{SIS}_{m2}}) + \gamma\,(\mathcal{L}_{\text{VIB}_{m1}} + \mathcal{L}_{\text{VIB}_{m2}}), \tag{11}$$

The weights of each loss and τ scaling were set empirically and are presented and are presented in Supplementary §D.

(a) **TEXT-to-CXR retrieval.**

	Prob?	Similarity	MIMIC-CXR R@1	MIMIC-CXR R@5	OpenI R@1	OpenI R@5	Chexpert5x200 R@1	Chexpert5x200 R@5	RSUM
MedCLIP [56]	✗	Cosine	1.0	4.3	0.6	2.8	2.6	3.0	14.3
CXR-CLIP [61]	✗	Cosine	47.3	70.4	**12.7**	25.2	8.5	23.0	187.1
BiomedCLIP [62]	✗	Cosine	36.2	59.9	9.0	19.9	6.4	19.8	151.2
CheXzero [50]	✗	Cosine	26.7	50.0	5.8	15.1	3.5	17.8	118.9
MEDBind [16]	✗	Cosine	40.8	67.5	11.6	25.5	7.9	21.4	174.7
BioVil-T [1]	✗	Cosine	28.4	58.2	8.1	18.9	4.9	17.1	135.6
SAT [32]	✗	Cosine	40.3	69.2	6.7	14.7	**9.1**	26.7	166.7
PCME++ [10]	✓	CSD	32.1	50.6	10.8	**28.3**	4.0	16.2	142.0
PROBMED (Ours)	✓	Hellinger	**47.9**	**71.4**	**12.7**	27.5	8.9	**28.4**	**196.8**

(b) **TEXT-to-ECG retrieval.**

	Similarity	MIMIC-ECG R@1	MIMIC-ECG R@5	PTB-XL R@1	PTB-XL R@5	RSUM
ECG-CLIP [16]	Cosine	40.8	76.7	2.3	9.8	129.6
MEDBind [16]	Cosine	44.1	78.2	**3.1**	12.1	137.5
PCME++ [10]	CSD	40.9	54.8	1.3	11.3	108.3
PROBMED (Ours)	Hellinger	**48.3**	**87.0**	2.8	**12.2**	**150.3**

(c) **TEXT-to-ECHO retrieval.**

	Similarity	MIMIC-ECHO R@1	MIMIC-ECHO R@5	RSUM
EchoCLIP [8]	Cosine	1.1	6.4	7.5
PCME++ [10]	CSD	1.0	5.0	6.0
PROBMED (Ours)	Hellinger	**2.4**	**7.8**	**10.2**

Table 1. Cross-modal retrieval performance, Recall@K, for (a) TEXT-to-CXR, (b) TEXT-to-ECG, and (c) TEXT-to-ECHO retrieval tasks. CSD stands for closed-form distance from [10]. The best performance is in **bold**, while the second-best is underlined.

Dataset	Emerg.	Task	#Cls	#test
MIMIC-CXR [26]*	✗	Retrieval/Multimodal	-	24,799
OpenI [13]	✗	Retrieval	-	2,864
CheXpert5x200 [24]	✗	Retrieval	5	1,000
RSNA [48]	✗	Classification	2	5,338
COVID Kaggle [7]	✗	Classification	2	2,780
Montgomery [4]	✗	Classification	2	106
CheXchoNet [2]	✓	Classification	2	3,667
MIMIC-ECG [19]*	✗	Retrieval/Multimodal	-	24,644
PTB-XL [55]	✗	Retrieval/Classification	71	2,198
ICBEB [33]	✗	Classification	9	1,376
MUSIC [36]	✓	Classification	2	125
MIMIC-ECHO [18]*	✗	Retrieval	-	1,957
EchoNet-Dynamic [42]	✗	Classification	2	1,264

Table 2. Datasets for CXR, ECG, and ECHO modalities. For each dataset, we report the task (i.e., classification, retrieval, and multimodal), number of classes (#Cls), and number of test samples (#test). CheXchoNet and MUSIC are both **emergent** (Emerg.) and external datasets. PROBMED was never trained on CXR-ECHO or ECG-ECHO pairs. *These datasets were used for pertaining, but test set was reserved (Supplementary §A)

4. Experiments and Results

This section introduces our experiments to evaluate PROBMED, compared to other Med-VLPMs. We trained PROBMED exclusively on MIMIC datasets. We also trained a PCME++ model for comparison. We evaluated on the 3 MIMIC and 10 external datasets. All datasets used in testing are in Tab. 2, and detailed in Supplementary §A.

Text-to-Modality Retrieval: In text-based retrieval experiments on free-form clinical notes, deterministic embeddings often miss linguistic nuances. Our probabilistic framework represents each text as a distribution, enabling similarity metrics that account for central tendency and uncertainty through probabilistic similarity comparisons.

Zero-shot and few-shot classification: We tested PROBMED on traditional zero-shot (ZS) and few-shot (FS) datasets. In addition, we explored emergent ZS and FS classification, which refers to a model's ability to align modality pairs that were not explicitly trained during pretraining (e.g., ECG and ECHO pairs were unseen). Inspired by [17], we observed that training this unified model, PROBMED, facilitated these emergent alignments. We illustrated this phenomenon by evaluating classification tasks that map ECG or CXR inputs to *ECHO labels*, even though these specific pairs were never observed together during training.

Multimodal Classification: Medical decision-making relies on multiple sources of evidence (e.g., combining information from both CXR and ECG for more accurate prognoses [15]). As discussed in the introduction, a central motivation behind PROBMED is its capacity to encode and integrate information from various modalities. We demonstrated this by evaluating the ZS and FS performance of PROBMED using embeddings from multiple modalities and comparing its performance to that of traditional approaches with a single modality.

Pushing the Boundaries of Probabilistic Models: While prior approaches [61, 62] emphasize improved ZS/FS performance using deterministic embeddings, PROBMED leverages probabilistic embeddings to capture each modality's intrinsic uncertainty and distributional characteristics. In addition to the conventional way of doing ZS and FS, we underlined use-cases that leverage the *learned distributions* to improve ZS and FS performance. **1) ZS**: We used

(a) **CXR-based zero and few-shot**

	COVID Kaggle			RSNA			Montgomery			CheXchoNet ⋆			Overall Rank	
	ZS	4S	16S	ZS	4S	16S	ZS	4S	16S	ZS	4S	16S	ZS	FS
MedCLIP [56]	75.3	85.5	90.8	75.7	58.0	65.4	88.3	87.3	88.5	61.6	55.8	63.9	4	8
CXR-CLIP [61]	76.9	**86.7**	91.6	72.7	64.1	70.9	81.1	85.8	91.6	61.1	53.2	59.7	6	7
BiomedCLIP [62]	84.4	86.0	89.4	81.6	80.3	84.0	84.5	87.0	92.2	62.1	59.8	61.8	3	2
CheXzero [50]	79.5	82.8	88.4	47.9	75.0	82.7	71.6	88.5	92.9	67.9	60.3	66.1	8	3
MEDBind [16]	**86.4**	86.2	**92.0**	80.0	67.3	73.4	86.8	89.9	91.8	62.0	57.6	65.4	2	4
BioViL-T [1]	69.9	67.7	76.6	71.5	**82.2**	**85.9**	**88.6**	88.6	91.8	63.3	56.6	64.7	5	5
SAT [32]	72.1	81.8	87.4	73.5	56.0	61.2	75.9	78.9	82.3	60.3	56.0	65.9	7	10
MedKLIP† [57]	68.8	85.4	90.2	82.4	79.3	82.8	51.7	79.5	84.2	63.8	56.0	63.6	9	5
PCME++ [10]	48.6	79.6	85.9	45.2	73.2	79.2	50.4	75.7	81.8	59.9	56.8	62.7	10	9
ProbMED (Ours)	**86.4**	86.5	91.8	**82.5**	**82.2**	84.7	84.3	**93.1**	**93.7**	**69.5**	**63.3**	**68.5**	1	1

(b) **ECG-based zero and few-shot**

	PTB-XL			ICBEB			MUSIC ⋆			Overall Rank	
	ZS	4S	16S	ZS	4S	16S	ZS	4S	16S	ZS	FS
ECG-CLIP [16]	52.3	67.1	71.2	61.9	69.1	74.1	60.4	48.6	51.4	4	5
MEDBind [16]	55.3	71.1	81.8	65.7	81.2	87.8	61.3	51.5	54.6	3	2
ECG-FM [37]	-	69.1	71.6	-	69.3	71.8	-	50.0	53.1	-	4
PCME++ [10]	61.2	75.4	79.9	71.3	74.1	80.5	67.2	46.7	48.6	2	3
ProbMED (Ours)	**64.5**	**82.6**	**87.6**	**75.3**	**84.8**	**90.1**	**70.5**	**53.8**	**59.1**	1	1

(c) **ECHO-based zero and few-shot**

	EchoNet-Dynamic			Overall Rank	
	ZS	4S	16S	ZS	FS
EchoCLIP [8]	75.1	**88.3**	95.0	2	2
PCME++ [10]	73.6	87.5	94.1	3	3
ProbMED (Ours)	**82.6**	87.7	**96.2**	1	1

Table 3. Performance comparison of ProbMED with state-of-the-art Med-VLPMs on many different modalities, few-shot tasks across (a) CXR-based, (b) ECG-based, and (c) ECHO-based datasets. We report AUROC scores under zero-shot (ZS), 4-shot (4S), and 16-shot (16S). For probabilistic models (i.e., PCME++ and ProbMED), only μ embeddings were used in few-shot learning for fair comparisons. ⋆**CheXchoNet** is an *emergent* dataset using CXR as an input to predict an ECHO label, composite of severe left ventricular hypertrophy (SLVH) and dilated left ventricle (DLV). ⋆**MUSIC** is an *emergent* dataset using ECG (input) to ECHO label, composite of SLVH and DLV. †MedKLIP performance was with the 77 classes in disease book (i.e., COVID and CheXchoNet were added to the original disease book).

an uncertainty-based prompt filtering mechanism proposed in [10], which filters out prompts with high uncertainty, assuming more ambiguous prompts are therefore less helpful as classifier references. **2) FS**: the probabilistic nature of ProbMED is leveraged to generate synthetic samples that provide more training samples. For example, by sampling from the latent distribution, a *single* image can generate *multiple* effective training examples for FS training.

4.1. Text-to-Modality Retrieval

We benchmarked ProbMED across several text-to-modality retrieval tasks against state-of-the-art Med-VLPMs. Our evaluation leveraged a probabilistic framework that measures central tendency and uncertainty using tailored similarity metrics. For all models, we used the most appropriate similarity for the respective model (i.e., cosine similarity was used for deterministic methods, while PCME++ and ProbMED used probabilistic similarities). Tab. 1a, Tab. 1b, and Tab. 1c summarizes the top-k performance (Recall@K) across three retrieval tasks: TEXT-to-CXR, TEXT-to-ECG, and TEXT-to-ECHO. Compared to competing models, ProbMED achieved the highest overall recall performance in these tasks, as indicated by the recall total sum (RSUM). These results underscore the feasibility of incorporating probabilistic modeling and multimodal binding into classical tasks. Using our proposed method and training on multiple modality pairs, we saw that ProbMED improves the modality-text alignment.

4.2. Zero-shot and few-shot classification

We assessed ProbMED using *traditional* ZS and FS protocols [17, 46] for a fair comparison between probabilistic and deterministic models. ZS performance was calculated using the similarity metrics between text and non-text modality embeddings according to [17]. Like recall, we applied the most appropriate similarity to measure the relationship between modality-text pairs. We employed linear probing for FS learning following [17]. Since ProbMED uses probabilistic embeddings, we only used μ embeddings to enable direct comparison to the output embeddings from the deterministic encoders. See §4.4 for leveraging the complete probabilistic distribution.

Tab. 3a, Tab. 3b, and Tab. 3c summarizes the area under the receiver operating characteristic curve (AUROC) of ZS and FS results for multiple datasets. ProbMED had competitive results and often outperformed state-of-the-art models across the evaluated modalities. In particular, our method significantly enhances CXR and ECG tasks while marginally surpassing ECHO experiments. In the **emergent** tasks (CheXchoNet and MUSIC), we noted substantial performance gains with ProbMED, where training on only ECHO-text pairing significantly improves performance.

4.3. Multimodal Classification

We evaluated ProbMED for chronic kidney disease (CKD) and chronic heart disease (CHD) classification under ZS

(a) **CXR and ECG concatenated to improve prognostication.**

(b) **AUROC performance (in %) on two prognostic tasks.**

	CKD			CHD			SUM
	ZS	4S	16S	ZS	4S	16S	
State-of-the-art best performance from respective VLPMs							
CXR	73.6[50]	66.9[50]	71.2[62]	77.1[50]	68.4[61]	75.2[61]	432.4
ECG	61.2[16]	66.4[16]	67.8[16]	65.7[16]	73.5[16]	74.1[16]	408.8
CXR+ECG*	71.5[16]	68.4[16]	69.8[16]	75.3[16]	71.7[16]	78.6[16]	435.3
PCME++							
CXR	51.0	68.6	68.8	51.1	72.7	76.7	388.9
ECG	31.7	66.7	70.3	40.3	74.7	76.7	360.4
CXR+ECG	46.8	69.4	71.6	52.4	**76.9**	78.7	395.8
PROBMED *(Ours)*							
CXR	75.0	70.6	76.5	77.0	71.9	79.8	450.8
ECG	68.5	67.4	71.1	70.3	72.2	76.7	426.2
CXR+ECG	**78.1**	**71.5**	**76.8**	**78.4**	73.0	**80.8**	**458.6**

Table 4. Multimodal classification results for chronic kidney disease (CKD) and chronic heart disease (CHD). (a) A visual example of multimodal decision-making with two non-text modalities. (b) Classification results under ZS, 4S, and 16S settings. SUM represents the total performance across the ZS and FS settings (in %). The best Med-VLPM performance is reported in (b) and cited. *Only MEDBind[16] processes both ECG and CXR.

and FS settings in MIMIC datasets, particularly on a subset of patients with both CXR and ECG (**MIMIC-CONNECT**, described in more detail in Supplementary §A.4). In these experiments, we explored using CXR-only, ECG-only, and *concatenation* of CXR and ECG (i.e., multimodal).

The integration process is visualized in Tab. 4a, showing how CXR and ECG can be used *together* for complex diseases. Tab. 4b presents AUROC comparisons among state-of-the-art Med-VLPMs, PCME++, and our PROBMED approach across single-modality (CXR *or* ECG), and multimodality (CXR *and* ECG). Overall, PROBMED outperformed competing methods. CXR+ECG with PROBMED produced gains in tasks, improving single-modality baselines by **the largest gain of 7.8%**–in ZS and FS scenarios. These results emphasize the value of leveraging multiple data sources in medical decision-making and highlight PROBMED's effectiveness in integrating multimodal information for improved prognosis.

4.4. Pushing the Boundaries of Probabilistic Models

This section outlines some experimental assessments for the utility of probabilistic modeling.

Uncertainty-based Prompt Filtering: We showcased this task as a proof-of-concept for ZS classification in CXR datasets, first proposed in [10]. Although ZS requires a well-curated set of prompts for prediction, the common way is taking the *average* of multiple text prompts to make a robust prototype for each class. However, optimizing the best set of prompts is not trivial. We perform uncertainty-based prompt filtering, where we choose the prompts based on the learned variance, σ^2, assuming the prompts with low variance are robust. For comparison, we conducted ZS experiments using: (1) a single prompt `"Chest X-ray of {·}"` where · denotes the name of a target disease for each dataset, and (2) 80 prompts describing the disease for the CXR datasets. For the uncertainty-based filtering, we chose k prompts with the lowest variance, which worked the best for each dataset [10]. Our results in Tab. 5 demonstrate that using prompt filtering on the learned variance consistently improved the ZS performance across all CXR datasets. Thus, PROBMED captures meaningful probability distributions of the given prompts.

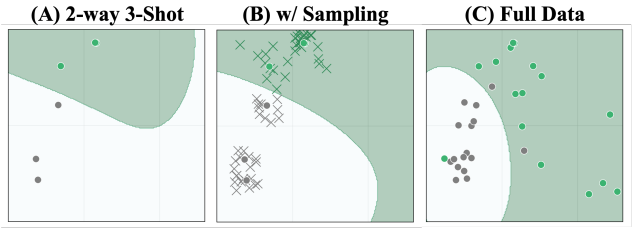

Figure 2. Qualitative comparison of decision boundaries on Kaggle COVID. Each plot shows data projected on the same PCA with an SVM (radial kernel) decision boundary. (A) The boundary is less reliable in a 2-way 3-shot few-shot setting. (B) PROBMED's probabilistic sampling (**X-marks** indicate sampled points) improves the boundary with more samples. (C) Using the full dataset. *Note*: two green points are overlapped in the 3-shot scenario.

Prompts	CX	RN	CV	MG	SUM
`"Chest X-ray of {·}"`	70.2	63.4	78.8	86.8	299.2
All 80 prompts	70.8	83.0	87.2	85.6	326.6
Uncertainty-based filtering	**71.0**	**85.6**	**87.3**	**87.7**	**331.6**

Table 5. Zero-shot performance using uncertainty-based prompt-filtering. **CX**: CheXchoNet, **RN**: RSNA, **CV**: COVID, and **MG**: Montgomery, **SUM**: the sum of performance across datasets.

Sampling to Improve Few-shot: We explore a novel probabilistic FS learning approach that harnesses feature sampling to enrich representation in FS regimes–enabled through probabilistic modeling. Unlike methods that rely solely on deterministic embeddings, PROBMED models each input as a distribution (i.e., refer to §3.3, Eq. (2). Let $\mathcal{P}(z)$ denote the learned embedding distribution for an input modality. Using the reparameterization trick, we drew n distinct feature samples from $\mathcal{P}(z)$ to create additional data $\{z'_j\}_{j=1}^{n}$, effectively capturing the latent uncertainty of the data. When applied to our FS datasets with $n = 16$, we improved over *all our FS* results (i.e., §4.2) with this probabilistic sampling approach, as shown in Fig. 3. This sampling approach is also visualized in Fig. 2, where, in a 2-way 3-shot scenario, the decision boundary for traditional methods is not enough to approximate the true decision boundary. With sampling, the decision boundary is much closer.

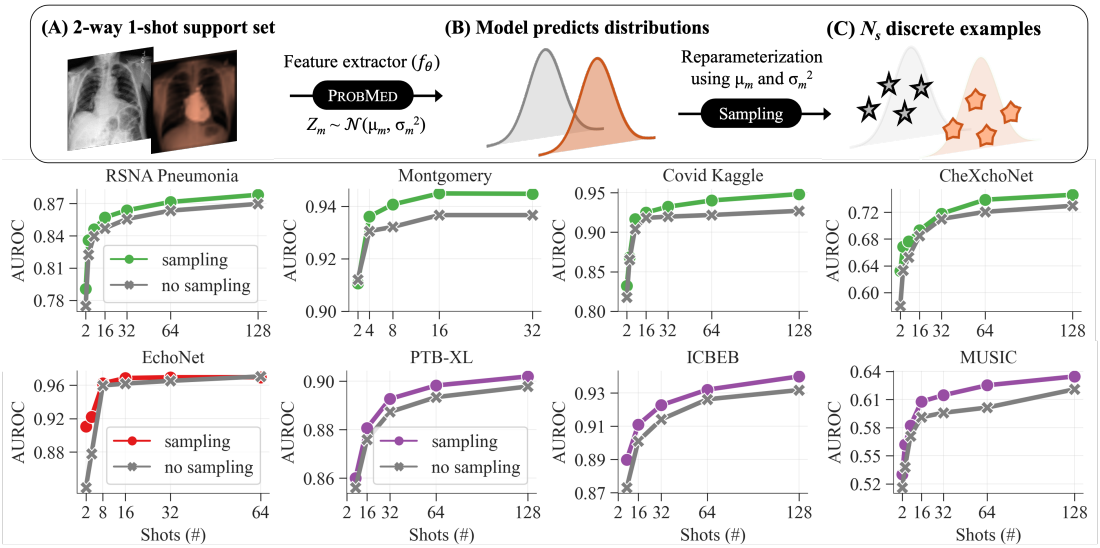

Figure 3. Probabilistic sampling in ProbMED improves few-shot performance. **(A–C)** Our probabilistic feature sampling strategy. ProbMED extracts a probabilistic distribution $Z \sim \mathcal{N}(\mu, \sigma^2)$ for inputs. We leverage the reparameterization trick to sample 16 distinct feature embeddings, capturing latent uncertainty. **(Bottom)** Few-shot AUROC comparison across 8 datasets. Sampling-based embeddings (*colored lines*) against using only the μ embedding (*grey lines*). For reference, *grey lines* is ProbMED results in Tab. 3a, 3b, and 3c.

Fig. 2 and Fig. 3 illustrate our approach, which we believe could be useful for limited data availability.

5. Ablation

We performed ablations to evaluate the impact of our ProbMED design choices, concentrating on the similarity metric applied for probabilistic embeddings and additional enhancements such as our SIS loss and BN. Tab. 6a compares four similarity metrics: cosine, CSD, Bhattacharyya, and Hellinger, on MIMIC-CXR, MIMIC-ECG, and MIMIC-ECHO (that is, all possible pairs of modality-text). Cosine represents the **deterministic** training approach on these modality pairs and performs well in some recall tests; however, Hellinger consistently achieved the highest Recall@1 and Recall@5 scores across datasets, confirming its effectiveness. Hellinger offers a more expressive similarity measure by accounting for the overlap between probability distributions. This enables the model to capture more nuanced cross-modality relationships. The low performance of CSD in Tab. 6a arises from the limitations of its metric, not probabilistic training, and shows that a robust probabilistic distance (e.g., Hellinger) is needed to outperform deterministic baselines.

Tab. 6b examines two enhancements: our SIS loss integrated into pretraining and BN applied before feature aggregation. Using the SIS loss on top of the Hellinger baseline improved performance in MIMIC-CXR, while BN further boosted scores in MIMIC-ECHO. Combined, we saw the highest Recall@1 and Recall@5 across datasets, underscoring the complementary benefits of modeling through sampling (SIS loss) and stabilizing feature distributions (BN).

(a) **Ablation training probabilistic similarity metric.**

Similarities	MIMIC-CXR		MIMIC-ECG		MIMIC-ECHO		RSUM
	R@1	R@5	R@1	R@5	R@1	R@5	
Cosine*	42.4	64.5	**48.9**	87.3	**2.0**	6.3	251.4
CSD	32.1	50.6	40.9	54.8	1.0	5.0	184.4
Bhattacharyya	43.4	63.7	41.9	87.9	1.0	3.2	241.1
Hellinger	**44.5**	**67.7**	48.1	**91.1**	**2.0**	**6.9**	**260.3**

(b) **Ablation of additional loss and batch normalization.**

Method	MIMIC-CXR		MIMIC-ECG		MIMIC-ECHO		RSUM
	R@1	R@5	R@1	R@5	R@1	R@5	
Hellinger	44.5	67.7	48.1	**91.1**	2.0	6.9	260.3
+ SIS Loss	46.2	**71.5**	48.1	87.0	2.1	6.4	261.3
+ BatchNorm	47.2	70.7	47.9	86.7	2.3	7.5	262.3
+ Both (Ours)	**47.9**	71.4	**48.3**	87.0	**2.4**	**7.8**	**264.8**

Table 6. (a) Similarity metric performance across MIMIC-CXR, MIMIC-ECG, and MIMIC-ECHO. (b) Results show the impact of the new loss and batch normalization on performance.* μ embedding was used to calculate the cosine similarity.

6. Conclusion

ProbMED is a probabilistic framework for binding multimodal medical data—modeling each modality as a distribution to capture many-to-many relationships—and outperforms existing Med-VLPMs in retrieval, ZS, and FS across 13 medical datasets. Future works include integrating additional modalities, utilizing label efficient fine-tuning, and exploring more probabilistic use cases. In summary, our probabilistic approach to multimodal Med-VLPM introduce fresh perspectives in the field.

Acknowledgments

Study was funded by NSERC RGPIN-2022-05117. CM holds the Chair in Medical Imaging at the Joint Department of Medical Imaging at University Health Network and University of Toronto (UofT). YG holds CIHR Canada Graduate Scholarship - Doctoral. SK is funded by the doctoral fellowship from Data Sciences Institute at UofT.

References

[1] Shruthi Bannur et al. Learning to exploit temporal structure for biomedical vision-language processing. In *CVPR*, 2023. 5, 6

[2] Shreyas Bhave, Victor Rodriguez, Timothy Poterucha, Simukayi Mutasa, Dwight Aberle, Kathleen M Capaccione, Yibo Chen, Belinda Dsouza, Shifali Dumeer, Jonathan Goldstein, et al. Deep learning to detect left ventricular structural abnormalities in chest x-rays. *European heart journal*, 45 (22):2002–2012, 2024. 5, 1, 2

[3] Tim GW Boers, Kiki N Fockens, Joost A van der Putten, Tim JM Jaspers, Carolus HJ Kusters, Jelmer B Jukema, Martijn R Jong, Maarten R Struyvenberg, Jeroen de Groof, Jacques J Bergman, et al. Foundation models in gastrointestinal endoscopic ai: Impact of architecture, pre-training approach and data efficiency. *Medical Image Analysis*, 98: 103298, 2024. 2

[4] Sema Candemir, Stefan Jaeger, Kannappan Palaniappan, Jonathan P Musco, Rahul K Singh, Zhiyun Xue, Alexandros Karargyris, Sameer Antani, George Thoma, and Clement J McDonald. Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration. *IEEE transactions on medical imaging*, 33(2):577–590, 2013. 5, 1, 2

[5] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15789–15798, 2021. 3, 4

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 4

[7] 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. Can ai help in screening viral and covid-19 pneumonia? *Ieee Access*, 8:132665–132676, 2020. 5, 1, 2

[8] Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5): 1481–1488, 2024. 1, 2, 3, 5, 6, 4

[9] Ivaylo Christov, Tatyana Neycheva, Ramun Schmid, Todor Stoyanov, and Roger Abächerli. Pseudo-real-time low-pass filter in ecg, self-adjustable to the frequency spectra of the waves. *Medical & biological engineering & computing*, 55: 1579–1588, 2017. 1

[10] Sanghyuk Chun. Improved probabilistic image-text representations. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 4, 5, 6, 7

[11] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8415–8424, 2021. 1, 2

[12] Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio De Rezende, Yannis Kalantidis, and Diane Larlus. Probabilistic embeddings for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8415–8424, 2021. 2

[13] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016. 5, 1, 2

[14] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[15] Cândida Fonseca, Teresa Mota, Humberto Morais, Fernando Matias, Catarina Costa, António G Oliveira, Fátima Ceia, and EPICA Investigators. The value of the electrocardiogram and chest x-ray for confirming or refuting a suspected diagnosis of heart failure in the community. *European journal of heart failure*, 6(6):807–812, 2004. 5

[16] Yuan Gao, Sangwook Kim, David E Austin, and Chris McIntosh. Medbind: Unifying language and multimodal medical data embeddings. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 218–228. Springer, 2024. 1, 2, 5, 6, 7

[17] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 5, 6

[18] B Gow, T Pollard, N Greenbaum, B Moody, A Johnson, E Herbst, et al. Mimic-iv-echo: Echocardiogram matched subset (version 0.1). *PhysioNet*, 2023. 5, 1, 2

[19] Brian Gow, Tom Pollard, Larry A Nathanson, Alistair Johnson, Benjamin Moody, Chrystinne Fernandes, Nathaniel Greenbaum, Seth Berkowitz, Dana Moukheiber, Parastou Eslami, et al. Mimic-iv-ecg-diagnostic electrocardiogram matched subset. 2023. 5, 1, 2

[20] Ibrahim Ethem Hamamci, Sezgin Er, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Muhammed Furkan Dasdelen, Omer Faruk Durugol, Bastian Wittmann, Tamaz Amiranashvili, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024. 2

[21] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 558–567, 2019. 5

[22] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. 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)*, pages 3942–3951, 2021. 1

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 3

[24] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 5, 1, 2

[25] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013. 1, 2

[26] Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019. 5, 1, 2

[27] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023. 2

[28] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, page 2, 2019. 2

[29] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 4, 3

[30] Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *International Conference on Machine Learning*, pages 17085–17104. PMLR, 2023. 1, 2

[31] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. 3, 4, 5

[32] Bo Liu et al. Improving medical vision-language contrastive pretraining with semantics-aware triage. *TMI*, 2023. 5, 6

[33] Feifei Liu, Chengyu Liu, Lina Zhao, Xiangyu Zhang, Xiaoling Wu, Xiaoyan Xu, Yulin Liu, Caiyun Ma, Shoushui Wei, Zhiqiang He, 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. 5, 1, 2

[34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF ICCV*, 2021. 3, 4, 5

[35] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 4

[36] Alba Martın-Yebra, Antonio Bayés de Luna, Rafael Vázquez, Pere Caminal, Pablo Laguna, and Juan Pablo Martınez. The music database: Sudden cardiac death in heart failure patients. *Computing in Cardiology*, 2024. 5, 1, 2

[37] Kaden McKeen, Laura Oliva, Sameer Masood, Augustin Toma, Barry Rubin, and Bo Wang. Ecg-fm: An open electrocardiogram foundation model. *arXiv preprint arXiv:2408.05178*, 2024. 2, 6

[38] Microsoft and Hugging Face. Swin-tiny patch4 window7 224. https://huggingface.co/microsoft/swin-tiny-patch4-window7-224, 2023. Accessed: 2025-03-05. 3, 4

[39] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 2

[40] Seong Joon Oh, Kevin P Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew C Gallagher. Modeling uncertainty with hedged instance embeddings. In *International Conference on Learning Representations*, 2019. 4

[41] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[42] David Ouyang, Bryan He, Amirata Ghorbani, Matt P Lungren, Euan A Ashley, David H Liang, and James Y Zou. Echonet-dynamic: a large new cardiac motion video data resource for medical machine learning. In *NeurIPS ML4H Workshop: Vancouver, BC, Canada*, 2019. 5, 2

[43] Daniel Pan, Pierpaolo Pellicori, Karen Dobbs, Jeanne Bulemfu, Ioanna Sokoreli, Alessia Urbinati, Oliver Brown, Shirley Sze, Alan S Rigby, Syed Kazmi, et al. Prognostic value of the chest x-ray in patients hospitalised for heart failure. *Clinical Research in Cardiology*, pages 1–14, 2021. 2, 3

[44] Leandro Pardo. *Statistical inference based on divergence measures*. Chapman and Hall/CRC, 2018. 4, 2

[45] Hoifung Poon. Multimodal generative ai for precision health. *NEJM AI Sponsored*, 2023. 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 6, 5

[47] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother,

Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024. 2

[48] George Shih et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 5, 1, 2

[49] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE journal of biomedical and health informatics*, 25(5):1519–1528, 2020. 3

[50] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature biomedical engineering*, 6(12): 1399–1406, 2022. 1, 2, 5, 6, 7

[51] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024. 1, 2

[52] Uddeshya Upadhyay, Shyamgopal Karthik, Massimiliano Mancini, and Zeynep Akata. Probvlm: Probabilistic adapter for frozen vison-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1899–1910, 2023. 2

[53] Ramachandran S Vasan, Vanessa Xanthakis, Asya Lyass, Charlotte Andersson, Connie Tsao, Susan Cheng, Jayashri Aragam, Emelia J Benjamin, and Martin G Larson. Epidemiology of left ventricular systolic dysfunction and heart failure in the framingham study: an echocardiographic study over 3 decades. *JACC: Cardiovascular Imaging*, 11(1):1–11, 2018. 2

[54] Jaroslav Vondrak and Marek Penhakert. Statistical evaluation of transformation methods accuracy on derived pathological vectorcardiographic leads. *IEEE Journal of Translational Engineering in Health and Medicine*, 10:1–8, 2022. 2

[55] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific data*, 7(1):154, 2020. 5, 1, 2

[56] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, page 3876, 2022. 2, 5, 6, 4

[57] Chaoyi Wu et al. Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In *ICCV*, pages 21372–21383, 2023. 6

[58] Jinxi Xiang, Xiyue Wang, Xiaoming Zhang, Yinghua Xi, Feyisope Eweje, Yijiang Chen, Yuchen Li, Colin Bergstrom, Matthew Gopaulchan, Ted Kim, et al. A vision–language foundation model for precision oncology. *Nature*, pages 1–10, 2025. 2

[59] Peng Ye, Weiqing Bai, Yuchen Ren, Wenran Li, Lifeng Qiao, Chaoqi Liang, Linxiao Wang, Yuchen Cai, Jianle Sun, Zejun Yang, et al. Genomics-fm: Universal foundation model for versatile and data-efficient functional genomic analysis. *bioRxiv*, pages 2024–07, 2024. 2

[60] Jianzhong You, Yuan Gao, Sangwook Kim, and Chris McIntosh. X2ct-clip: Enable multi-abnormality detection in computed tomography from chest radiography via tri-modal contrastive learning. *arXiv preprint arXiv:2503.02162*, 2025. 2

[61] Kihyun You, Jawook Gu, Jiyeon Ham, Beomhee Park, Jiho Kim, Eun K Hong, Woonhyuk Baek, and Byungseok Roh. 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, 2023. 1, 2, 5, 6, 7

[62] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. 1, 5, 6, 7

[63] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine learning for healthcare conference*, pages 2–25. PMLR, 2022. 2