HYDEN: Hyperbolic Density Representations for Medical Images and Reports

Zhi Qiao[1], Linbin Han[1,2,3], Xiantong Zhen[1], Jiahong Gao[2,3,4], Zhen Qian[1],
[1]Institute of Intelligent Diagnostics,
Beijing United-Imaging Research Institute, Beijing, China
[2]Center for MRI Research, AAIS, Peking University, Beijing, China
[3]Beijing City Key Lab for Medical Physics and Engineering,
Institution of Heavy Ion Physics, School of Physics, Peking University, Beijing, China
[4]McGovern Institute for Brain Research, Peking University, Beijing, China
Correspondence: zhen.qian@cri-united-imaging.com

Abstract

In light of the inherent entailment relations between images and text, embedding point vectors in hyperbolic space has been employed to leverage its hierarchical modeling advantages for visual semantic representation learning. However, point vector embeddings struggle to address semantic uncertainty, where an image may have multiple interpretations, and text may correspond to different images—a challenge especially prevalent in the medical domain. Therefor, we propose **HYDEN**, a novel hyperbolic density embedding based image-text representation learning approach tailored specifically for medical domain data. This method integrates text-aware local features with global features from images, mapping image-text features to density features in the hyperbolic space using hyperbolic pseudo-Gaussian distributions. An encapsulation loss function is employed to model the partial order relations between image-text density distributions. Experimental results demonstrate the interpretability of our approach and its superior performance compared to the baseline methods across various zero-shot tasks and fine-tuning tasks on different datasets.

1 Introduction

In recent years, cross-modal text-image representation learning has made tremendous advancements and drawn widespread attention in many tasks such as zero-shot learning and image-text retrieval. This success is largely due to the use of large volumes of image-text pair data to enhance vision-language representation learning (Radford et al., 2021). In medical imaging, cross-modal representation learning tailored to this specific domain data, such as chest radiographs and their associated radiology reports, can yield robust and powerful foundation models in specialized areas (Zhang and Metaxas, 2023; Zhang et al., 2024; Stevens et al., 2024).

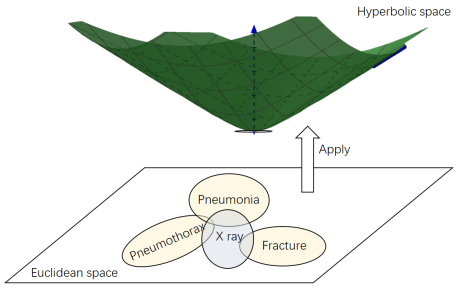


Figure 1: Representation of medical data embeddings transitioning from Euclidean to hyperbolic space to effectively capture and represent the density partial ordering, while maintaining the integrity of relative density relationships.

As the proverb goes, 'A picture is worth a thousand words.' This suggests that an image inherently contains more information than a textual description of it, which can be seen as merely a simplified abbreviation of the image. This relationship, where the text may serve as an entailment of the image, can be considered as visual-semantic hierarchy (Vendrov et al., 2016). Consequently, it is a plausible hypothesis that incorporating such inductive biases of visual semantic hierarchies into cross-modal alignment tasks could enhance the generalizability of representations and improve the interpretability of learned representations. Vendrov et al. (2016) introduced an order embedding strategy considering these hierarchical semantic during the text-image alignment process. However, numerous studies (Nickel and Kiela, 2017, 2018; Xu et al., 2022, 2023; Fu et al., 2023) have demonstrated that modeling data with inherent hierarchical features in non-Euclidean hyperbolic spaces can provide superior representations. By leveraging the advantages of hyperbolic space in modeling hierarchical structures and the generalization capabilities of cross-modal contrastive learning in zero-shot scenarios, (Desai et al., 2023) has proposed cross-modal hyperbolic representation learning. This approach employs the Lorentz manifold to map both image and text features into a hyperbolic space, utilizing

Proceedings of the 31st International Conference on Computational Linguistics, pages 6285–6297
January 19–24, 2025.

angular constraints based on entailment to learn the hierarchical order between text and images. However, representing image-text with point vectors has an intrinsic limitation: it cannot express semantic uncertainty (Vilnis and McCallum, 2014; Qian et al., 2021; Wu et al., 2023), meaning that a single image can generate different descriptions from various perspectives, and similarly, a single textual description can describe different but related images. This phenomenon is particularly evident in medical imaging and radiology reports. For instance, consider a patient with a rib fracture suspected of having right-sided pneumothorax. In the radiology report for this patient, the physician describes the imaging findings related to pneumothorax, highlighting the presence of a white line around the visceral pleural edge. Clinically, numerous pulmonary diseases, such as tuberculosis, cystic fibrosis, and pneumocystis jiroveci pneumonia, predispose individuals to pneumothorax. In domains such as document embedding (Zhu et al., 2023) and graph embedding (Gourru et al., 2022), the utilization of probability density embedding to represent objects as distributions within the target space effectively addresses this semantic uncertainty, resulting in significantly improved performance compared to point vector embedding.

Based on the motivation outlined above, which indicates on the hierarchical visual semantic features and inherent semantic uncertainty in medical imaging, we propose **HYDEN**, a hyperbolic density representation for medical images and reports. This approach leverages the advantages of the hyperbolic space for capturing visual-semantic hierarchy, while incorporating a probability density embedding strategy to model semantic uncertainty. The main contributions are as follows:

- To the best of our knowledge, this is the first work that introduces the hyperbolic space to cross-modal representation learning for medical image-text data.
- We introduce a text-aware image local feature extraction method that focuses on local regions, enhancing the granularity of analysis; moreover, we employ encapsulation constraints to model the density order between images and text, fostering a deeper semantic connection.
- Extensive experiments demonstrate the superior capabilities of our approach in achieving semantic alignment.

2 Related Work

Image-text representation learning has gained interest for its potential to improve visual representation. Traditional methods, like CLIP (Radford et al., 2021), primarily use contrastive learning in Euclidean space and have been applied across general domains. In the medical field, domain-specific challenges arise due to the complex prior knowledge in medical image-text data. Several studies have explored representation learning tailored to medical contexts (Wang et al., 2024; Müller et al., 2022a; Cheng et al., 2023a; Huang et al., 2021), but these methods still operate in Euclidean space. The hierarchical semantics in medical data suggest that hyperbolic space, with its ability to model hierarchies, could be more effective.

The MERU framework introduced hyperbolic image-text embeddings (Desai et al., 2023), departing from Euclidean methods. However, MERU and similar approaches still use point embeddings, which fail to capture *semantic uncertainty*—where one image maps to multiple descriptions and vice versa. To address this, we propose integrating **density embeddings in hyperbolic space** to capture semantic uncertainty. Unlike point embeddings, density embeddings represent objects as probability distributions, modeling semantic variation. While density embeddings have been used for uncertainty and entailment in Euclidean space (Vilnis and McCallum, 2014; Qian et al., 2021; Bojchevski and Günnemann), *our method is the first to extend this concept into hyperbolic space for image-text representation learning*.

3 Preliminaries

Hyperbolic Geometry Hyperbolic geometry is a non-Euclidean geometry with a constant negative curvature, and it can be visualized as the forward sheet of the two-sheeted hyperboloid. In this study, we will use the Lorentz model on the upper half of a two-sheeted hyperboloid, as claimed in (Nickel and Kiela, 2018), comes with a simpler closed form of the geodesics and does not suffer from the numerical instabilities in approximating the distance. Lorentz model $\mathbb{H}^n$ processing a constant curvature $-c$ can be represented as a set of points $z \in \mathbb{R}^{n+1}$. Lets $z, z' \in \mathbb{H}^n$, the Lorentzian product $\langle z, z' \rangle_{\mathcal{L}} = -z_0 z_0^{'} + \sum_{i=1}^{n} z_i z_i^{'}$. And, $\mathbb{H}^n = \{z \in \mathbb{R}^{n+1} : \langle z, z \rangle_{\mathcal{L}} = -1/c,\ c > 0\}$. The distance between z and z' is given by

$$d_\ell(z, z') = arccosh(-\langle z, z' \rangle_{\mathcal{L}}) \quad (1)$$

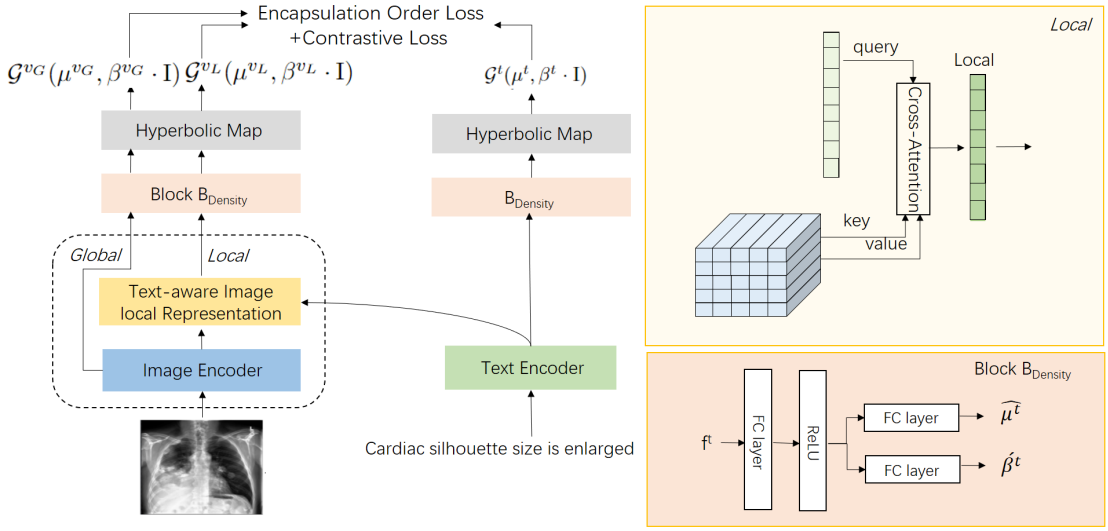


Figure 2: Framework of HYDEN: The contrastive loss function utilizes the negative Lorentzian distance as a metric for similarity. Additionally, an encapsulation loss is employed to enforce the density partial ordering of image and text embeddings within the representation space.

which is also the length of the geodesic that connects z and z'. We will refer to the one-hot vector $\mu_0 = [1/\sqrt{c}, 0, 0, 0...0] \in \mathbb{H}^n \subset \mathbb{R}^{n+1}$ as the origin of the hyperbolic space.

Tangent Space of Hyperbolic Space The tangent space at a point $\mu \in \mathbb{H}^n$ is a Euclidean space composed of vectors. Denoted by $T_\mu \mathbb{H}^n$, this tangent space represents the set of vectors in the same ambient space $\mathbb{R}^{n+1}$ where $\mathbb{H}^n$ is embedded. The vectors in $T_\mu \mathbb{H}^n$ satisfy an orthogonality condition relative to the Lorentzian product, defined as $T_\mu \mathbb{H}^n := \{u : \langle \mu, u \rangle_\mathcal{L} = 0\}$. This set can be visualized as the tangent space at the point μ on the forward hyperboloid sheet. Specifically, at the origin μ_0 of $\mathbb{H}^n$, the tangent space $T_{\mu_0} \mathbb{H}^n$ consists of vectors $v \in \mathbb{R}^{n+1}$. The norm $\|v\|_\mathcal{L}$, given by the Lorentzian inner product, simplifies to the Euclidean norm $\|v\|_2$, defined as $\|v\|_\mathcal{L} := \sqrt{\langle v, v \rangle_\mathcal{L}} = \|v\|_2$.

Exponential Map The exponential map provides a method for mapping a vector from a tangent space to its corresponding point on the surface of the hyperbolic space. For every $u \in T_\mu \mathbb{H}^n$, the exponential map $exp_\mu(u) : T_\mu \mathbb{H}^n \to \mathbb{H}^n$ allows us to project a vector u in $T_\mu \mathbb{H}^n$ onto $\mathbb{H}^n$ such that the distance from μ to the destination point of the map coincides with the Lorentzian norm $\|u\|_\mathcal{L}$ of u. In the context of hyperbolic space, the exponential map is given by the equation:

$$z = \exp_\mu(u) = \cosh(\|u\|_\mathcal{L})\mu + \sinh(\|u\|_\mathcal{L}) \frac{u}{\|u\|_\mathcal{L}} \quad (2)$$

In this paper, we specifically consider exponential maps where μ represents the origin of the hyperboloid ($\mathrm{O} = [\sqrt{1/c}, \mathbf{0}]$).

4 Method

In this section, we present a comprehensive introduction to the HYDEN model. Drawing on the foundation laid by the MERU model (Desai et al., 2023) and the widely acclaimed, user-friendly CLIP framework (Radford et al., 2021), our model adapts and extends these frameworks to address specific challenges in medical image-text representation learning. Figure 2 depicts the overall architecture of our model.

4.1 Image-Text Feature Embedding

In our model, the features $[\hat{f^v}, f^t]$ are derived from respective image and text encoders. For text data, we employ BioClinicalBERT (Alsentzer et al., 2019), a model that has been pre-trained on the MIMIC III dataset (Shen, 2016), to generate token-level embeddings. Consistent with practices outlined in (Cheng et al., 2023b), the output of the $[CLS]$ token is used as the medical text feature f^t, encapsulating the overall semantic content of the input text.

For image encoding, we utilize the widely-used Vision Transformer (ViT) architecture (Mu et al., 2022). We assume the output of the $[CLS]$ is considered as global feature $\hat{f}^v$, and the rest of the outputs from the image encoder is denotes as $\hat{f^v}$. Recognizing that pathological symptoms often occupy only a portion of a medical image, relying solely on global representations may not adequately capture

essential local semantic features. Thus, similar to approaches in (Huang et al., 2021; Cheng et al., 2023b; Müller et al., 2022b), we also extract local features besides global features. Specifically, we implement a Self-attention module (Vaswani et al., 2017), widely used in cross-modal feature extraction. In this setup, $\hat{f}^v$ acts as both the keys (K) and values (V), while the text embedding f^t functions as the query (Q). This configuration allows us to derive a text-aware image local representation, denoted as $\bar{\hat{f}}^v$.

4.2 Hyperbolic Density Embedding

Our objective is to transform image-text features into density representations within a hyperbolic space. Previous studies such as (Nagano et al.) and (Mathieu et al.) proposed methods like the pseudo-hyperbolic Gaussian distribution based on the Lorentz manifold. Due to the computational demands and numerical instabilities of the Poincaré-disk model, we opt for the more stable pseudo-hyperbolic Gaussian distribution for our hyperbolic density embedding. The tangent space $T_\mu\mathbb{H}^n$ of hyperbolic space $\mathbb{H}^n$ is a Euclidean space, and in $T_{\mu_0}\mathbb{H}^n$, vectors v satisfy $\mathrm{v} = \{v_0, v_1, ..., v_n\} \in \mathbb{R}^{n+1}$ where $v_0 = 0$, aligning with the dimensional properties.

To begin, we introduce separate deep nonlinear network blocks, $B_{density}$, for processing image and text features respectively. These blocks do not share parameters, ensuring distinct representations for each modality. As in Figure 2, for text features, $\hat{\mu^t}$ and $\hat{\beta^t}$ are the outputs of $B_{density}(f^t)$.

Instead of generating covariance matrices directly, which can introduce numerical instability, we use matrices based on diagonal or spherical assumptions. These are known for their computational efficiency and effectiveness in embedding tasks, particularly in the context of word distribution embedding where spherical covariance matrices have been shown to better model distributional partial order relationships (Vilnis and McCallum, 2014). We thus employ a covariance matrix based on the spherical assumption: $\Sigma^t = \hat{\beta^t} \cdot \mathrm{I} \in \mathbb{R}^{(n+1)\times(n+1)}$.

To ensure that our covariance matrix is positively definite, necessary for the stability of the pseudo-hyperbolic Gaussian distribution, we modify β^t using the expression $\beta^t = \exp(\hat{\beta^t})$ referring to solution in VAE(Kingma and Welling, 2019). This adjustment is crucial for maintaining the mathematical integrity of our model when dealing with real-world data. For the embedding vector $\hat{\mu^t} \in \mathbb{R}^n$, our aim is to project this vector onto hyperboloid space, which is achieved by mapping it through the exponential function as detailed in Equation 2.

The vector $\mu^t_{tan} = [0, \hat{\mu^t}]$ resides in $\mathbb{R}^{n+1}$ and belongs to the tangent space $T_{\mu_0}\mathbb{H}^n$ at the origin of the hyperboloid, $\mathcal{O}$. The norm $||\mu^t_{tan}||_{\mathcal{L}}$, which equals $||\hat{\mu^t}||_2$, ensures that the mapping preserves the distances inherent to the model's geometric structure. Upon applying the exponential map, we derive the expectation of the hyperbolic density representation:

$$\begin{aligned}\mu^t &= \exp_{\mu_0}(\mu^t_{tan}) \\ &= \left(\sqrt{1/c} \times \cosh(\sqrt{c}||\hat{\mu^t}||_2), \frac{\sinh(\sqrt{c}||\hat{\mu^t}||_2)}{\sqrt{c}||\hat{\mu^t}||_2}\hat{\mu^t}\right)\end{aligned} \quad (3)$$

The detailed procedure is shown in Appendix A.

This projection results in the hyperbolic density representation $\mathcal{G}^t(\mu^t, \beta^t \cdot \mathrm{I})$. Following a similar procedure, we also separately derive $\mathcal{G}^{v_L}(\mu^{v_L}, \beta^{v_L} \cdot \mathrm{I})$ and $\mathcal{G}^{v_G}(\mu^{v_G}, \beta^{v_G} \cdot \mathrm{I})$ for the local image features $\bar{\hat{f}}^v$ and global image features $\hat{\bar{f}}^v$, thereby ensuring a uniform approach to handling different modalities within our framework.

4.3 Loss Function Based on Density Embedding

Traditional point vector embedding often utilizes entailment angle constraints to define relationships between entities (Desai et al., 2023). However, when dealing with probability densities, the notion of partial order can be more complexly captured through the concept of encapsulation. Specifically, a density f is considered more specific than another density g if f is entirely encompassed by g, formally expressed as $f \preceq g \Leftrightarrow \{x : f(x) > \eta\} \subseteq \{x : g(x) > \eta\}$, for any $\eta \geq 0$, where η indicates the degree of encapsulation necessary for one distribution to entail another.

Imposing such partial order constraints on distributions brings up significant challenges. Drawing inspiration from (Athiwaratkun and Wilson), we employ asymmetric divergence measures between probability densities to address this. We introduce a simple penalty function, $d_\gamma(f, g) = \max(0, D(f \parallel g) - \gamma)$, which serves as a violation penalty rather than as a strict constraint of encapsulation. Here, $D(\|)$ represents the divergence measure used to quantify the extent of difference between distributions, and γ is a threshold defining the acceptable range of difference.

Among the choices for divergence measures, α-divergence provides a more flexible and generalized asymmetric measure (Renyi, 1961), allowing for adjustments in the zero-force penalty. This flexibility means that higher α values can enforce stricter encapsulation conditions $f \preceq g$. The general form of α-divergence, for $\alpha \neq 0, 1$, is given b $D_\alpha(f \parallel g) = \frac{1}{\alpha(\alpha-1)} \log\left(\int \frac{f(x)^\alpha}{g(x)^{\alpha-1}}\, dx\right)$.

This equation not only quantifies the differences between distributions but also facilitates a deeper understanding of the encapsulation relationships critical for effective density embedding.

We observe that as α approaches 0 or 1, it governs the degree of zero forcing, where minimizing $D_\alpha(f \parallel g)$ for high α values results in f becoming more concentrated in regions of g with high density. Conversely, for low α values, f tends to be mass-covering, encompassing regions of g even including those with low density. Notably, there exists a mathematical relationship between KL divergence and α-divergence, as indicated by: $\lim_{\alpha\to 1} D_\alpha(f \parallel g) = D_{KL}(f \parallel g)$ and $\lim_{\alpha\to 0} D_\alpha(f \parallel g) = D_{KL}(g \parallel f)$ (Pardo, 2006). Therefore, in our model, we opt for the more flexible and robust α-divergence as our metric.

For image-text embedded density $\mathcal{G}^{v_L}(\mu^{v_L}, \beta^{v_L} \cdot \mathrm{I})$ and $\mathcal{G}^t(\mu^t, \beta^t \cdot \mathrm{I})$, the encapsulation loss can be expressed as follows:

$$d_\gamma(\mathcal{G}^{v_L}, \mathcal{G}^t) = \max(0, -\frac{1}{2\alpha(\alpha-1)} \log\Big[\alpha(\frac{\beta^{v_L}}{\beta^t})^{\alpha(n+1)} + (1-\alpha)(\frac{\beta^t}{\beta^{v_L}})^{\alpha(n+1)}\Big] + \frac{(\mu^{v_L}-\mu^t)^T(\mu^{v_L}-\mu^t)}{\alpha(\beta^t)^{n+1} + (1-\alpha)(\beta^{v_L})^{n+1}} - \gamma) \tag{4}$$

The detailed procedure is shown in Appendix A. The loss function here only models the local representation density of the image, without imposing constraints on the global representation density. This is mainly because local representations are image features strongly correlated with text semantics, and they exhibit a visual-semantic hierarchy with text representations; however, global representations represent the features of the entire image, which may contain information not described in the current text due to the missed diagnoses or biases. Directly incorporating global features into the model may introduce semantic confusion. The constraints on global representations are involved in the contrastive loss function.

Given the batch sample $\mathbb{B} = \{\mathbb{B}^P, \mathbb{B}^N\}$, where $\mathbb{B}^P$ denotes the positive image-text sample set, and $\mathbb{B}^N$ represents the negative set, we define the encapsulation loss function as follows:

$$\mathfrak{L}_{order} = \sum_{(\mathcal{G}^t, \mathcal{G}^{v_L}) \in \mathbb{B}^P} d_\gamma(\mathcal{G}^t, \mathcal{G}^{v_L}) + \sum_{(\mathcal{G}^t, \mathcal{G}^{v_L}) \in \mathbb{B}^N} \max\{0, m - d_\gamma(\mathcal{G}^t, \mathcal{G}^{v_L})\} \tag{5}$$

For the positive samples, a definite partial order relationship exists, enabling the direct application of the density penalty $d_\gamma()$. For the negative samples, we enforce the penalty to exceed a margin m due to the absence of an order relationship.

Our goal is to enhance the similarity of semantic distributions between image-text pairs. Therefore, we also employ the classic CLIP contrastive solution (Radford et al., 2021) to compute the geodesic distance between the expectation values of image and text in hyperbolic densities as defined in Equation 1, applying Softmax normalization. We define $\mathfrak{L}^L_{con}$ as the contrastive loss between the image local representation and text, which is computed as an average of the contrastive losses from both image and text perspectives. Our goal is to ensure that the distributions of image and text can be roughly aligned within the same region. Therefor, for image global representation, we define $\mathfrak{L}^G_{con}$. The final loss function is $\mathfrak{L} = \tau\mathfrak{L}_{order} + 0.5 * (\mathfrak{L}^L_{con} + \mathfrak{L}^G_{con})$ where τ is the predefined variables.

5 Experiments

In this section, we aim to rigorously evaluate the performance of our algorithm. We first introduce the baseline model, followed by a description of the medical image-text data and training details used for model pre-training. Then, we discuss the advantages of our proposed model in medical image-text alignment from both quantitative and qualitative perspectives.

A key innovation of our algorithm lies in the use of density representations in hyperbolic space for image-text alignment. To validate the superiority of our approach, we compare it with three baseline methods: CLIP (Radford et al., 2021), which aligns image-text pairs in Euclidean space using point embeddings, Gloria (Huang et al., 2021) designing a global&local representation extraction module expanding CLIP to enhance the perception of local features also in Euclidean space, and MERU (Desai et al., 2023), which aligns image-text pairs in hyperbolic space using point embeddings. Gloria has released the pretrained weights[1]. For model training of the rest baselines, we primarily utilize the open-source code provided by the MERU project[2].

[1]https://github.com/marshuang80/gloria

[2]https://github.com/facebookresearch/meru

			Tuberculosis			RSNA			SIIM		
Methods	Dataset	Backbone	AUC	F1	ACC	AUC	F1	ACC	AUC	F1	ACC
Gloria	CheXpert	Resnet-50	0.701	0.393	0.608	0.714	0.490	0.729	0.534	0.382	0.405
CLIP	MIMIC-CXR	VIT-B	0.749	0.456	0.756	0.817	0.574	0.721	0.722	0.478	0.6306
MERU	MIMIC-CXR	VIT-B	0.765	0.442	0.675	0.792	0.545	0.671	0.725	0.477	0.649
HYDEN	MIMIC-CXR	Resnet-50	0.831	**0.592**	0.723	0.831	0.592	0.723	0.76	0.509	**0.701**
HYDEN	MIMIC-CXR	VIT-B	**0.854**	0.582	**0.843**	**0.856**	**0.629**	**0.787**	**0.786**	**0.529**	0.697

Table 1: Zero-shot image classification

Methods	Dataset	Backbone	Prec@3	Prec@5	Prec@10	NDCG@3	NDCG@5	NDCG@10
Gloria	CheXpert	Resnet-50	26.67	30.67	28	0.316	0.374	0.438
CLIP	MIMIC-CXR	VIT-B	23.81	18.57	12.14	0.408	0.414	0.439
MERU	MIMIC-CXR	VIT-B	33.33	30.0	27.86	0.417	0.452	0.465
HYDEN	MIMIC-CXR	Resnet-50	30.95	32.86	29.29	0.571	0.620	0.645
HYDEN	MIMIC-CXR	VIT-B	**38.1**	**37.14**	**35.71**	**0.624**	**0.656**	**0.677**

Table 2: Zero-shot image retrieval

While some variations of CLIP have been successfully applied in the medical image-text alignment domain, our primary focus is on comparing the differences between alignment in Euclidean space and hyperbolic space, as well as between point vector embeddings and distribution embeddings.

5.1 Training Details

Datasets: We train our alignment model using the MIMIC-CXR v2 dataset (Johnson et al., 2019), comprising over 227,000 studies of paired image-report data sourced from 65,379 patients undergoing various scans. Each study may contain one or two images, representing different scan views, resulting in a total of 377,110 images. During training, we perform random cropping, flipping, rotation, and other data augmentation techniques on the images, while also resizing them to a [224,224] dimension. Additionally, for the text data, we augment the reports by randomly adding medical entity prefixes to enhance semantic information, such as 'event_list: report'.

Settings: We employ ViT-B (Mu et al., 2022) with a patch size of 16 as the image encoder(Resnet-50 is also an alternative), as it has demonstrated competitive performance in hyperbolic space (Desai et al., 2023). Our initialization strategy for image/text encoders follows a similar style to MERU, with the exception of utilizing ClinicalBERT (Alsentzer et al.) as the pre-trained text encoder, which has been pre-trained on large-scale medical text data. For HYDEN, we initialize the learnable curvature parameter c to 1.0 and clamp it within the range of [0.1, 10.0] to prevent training instability. All experiments were conducted using three NVIDIA A40 GPU and the PyTorch framework.

Optimization: We adopt the AdamW optimizer with a weight decay of 0.2 and $(\beta_1, \beta_2) = (0.9, 0.98)$. Weight decay is disabled for all gains, biases, and learnable scalars. Models are trained for 32,000 iterations with a batch size of 384. The maximum learning rate is set to 1×10^{-5}, linearly increased for the first 500 iterations, followed by cosine decay to zero. We leverage mixed precision to expedite training, except when computing exponential maps and losses, where FP32 precision is used for numerical stability.

5.2 Quantitative Analysis

We evaluate all baselines and HYDEN on two categories of zero-shot downstream tasks(classification and text-image retrieval), and downstream fine-tuning classification task. We use five public datasets for the evaluation, where **Tuberculosis**(Rahman et al., 2020), **RSNA Pneumonia** (Shih et al., 2019), **SIIM-ACR Pneumothorax** (Kaggle, 2019) and **COVID-19**[3] are used for classification tasks, and **ChestXray14**(Wang et al., 2017a) is used for zero-shot text-image retrieval task. For the classification tasks, we report the Area Under the Curve (AUC), F1 score and Accuracy(ACC). For the retrieval task, Top-k Precision (abbreviated as Prec@k) and Tok-k Normalized Discounted Cumulative Gain (abbreviated as NDCG@k) are used to evaluate the retrieval performance. The details about our evaluation tasks and datasets are described in Appendix B.

Zero-shot Image Classification Table 1 presents the performance of the baselines and HYDEN across three classification datasets. The results indicate that HYDEN consistently demonstrates robust transfer classification performance in all classification tasks. Compared to CLIP, both MERU and HYDEN achieved improved accuracy. This suggests that using hyperbolic space

[3]https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

			RSNA			SIIM			COVID-19		
Methods	Dataset	Backbone	1%	10%	100%	1%	10%	100%	1%	10%	100%
ConVIRT	MIMIC-CXR	Resnet-50	0.839	0.856	0.876	0.83	0.856	0.876	0.868	0.954	0.973
Gloria	CheXpert	Resnet-50	0.859	0.866	0.884	0.859	0.866	0.884	0.906	0.938	0.972
MERU	MIMIC-CXR	VIT-B	0.855	0.874	0.886	0.849	0.879	0.906	0.852	0.968	0.995
HYDEN	MIMIC-CXR	Resnet-50	**0.886**	**0.899**	**0.907**	**0.866**	**0.901**	0.926	0.914	**0.989**	0.997
HYDEN	MIMIC-CXR	VIT-B	0.878	0.894	0.901	0.850	0.881	**0.927**	**0.924**	0.988	**0.999**

Table 3: Fine-tuning image classification (AUC metric)

	Text2Image@10		Tuberculosis		RSNA		SIIM	
	Prec	NDCG	AUC	F1	AUC	F1	AUC	F1
HYDEN	35.71	0.677	0.871	0.621	0.856	0.629	0.786	0.529
1. w KL Divergence	35	0.645	0.842	0.533	0.841	0.607	0.781	0.523
2. w/o encapsulation loss	27.86	0.626	0.846	0.576	0.816	0.567	0.756	0.501
3. w/o local representation	23.57	0.652	0.856	0.638	0.796	0.547	0.749	0.496

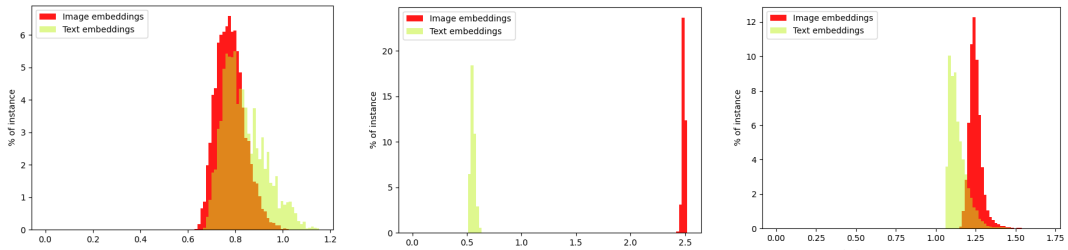
Table 4: Ablation Study of HYDEN: This table presents the results of ablating three key design choices within the HYDEN framework to evaluate their individual contributions.

for text-image representations, especially for medical data characterized by a visual semantic hierarchy, is more effective. Relative to MERU, HYDEN achieved the highest accuracy across almost all of metrics, highlighting the advantages of density embedding-based representation methods over point vector embedding, particularly in addressing the challenges of semantic uncertainty.

Zero-shot Retrieval Table 2 displays the performance of baseline models and HYDEN in "image-to-text" retrieval tasks. The results demonstrate that representation learning in hyperbolic space mostly outperforms that in Euclidean space; Among the methods, HYDEN exhibits the best retrieval performance. Furthermore, we observed a significant enhancement in the ranking quality of HYDEN's retrieval results compared to all baseline methods. We hypothesize that this improvement is linked to the method of density embedding. Similar to findings in the recommendation systems domain (Dos Santos et al., 2017), unlike point vector embeddings, density embeddings enable better handling of uncertainties, information sparsity, ambiguity, and even contradictions, which are common challenges in medical image-text data.

Fine-tuning Image Classification We further evaluate HYDEN on a fine-tuning image classification task where use different amounts of training data (1%, 10% or 100%) to evaluate the data-efficiency of the global image representations. The results are shown in Table 3. For the models pretrained in euclidean space, we train a linear classifier on top of the pretrained image encoder. Both MERU and HYDEN are pretrained in hyperbolic space, and they cannot be directly applied as a pre-trained model to downstream fine-tuning tasks(Desai et al., 2023). Hence, we introduce the multinomial logistic regression in the Lorentz model(Lorentz MLP)(Bdeir et al., 2024) as the linear classifier on top of density embedding of global image representation. The details are shown in Appendix A. Table 3 indicates that density representation based hyperbolic embedding can learn better representations for label-efficient classification.

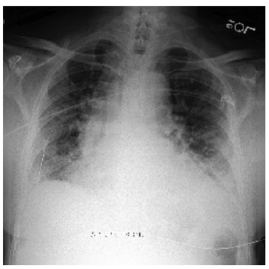
Ablation Studies In this section, we examine the impact of different design choices using HYDEN. Specifically, we trained three ablation models with default hyperparameters, and the results are presented in Table 4. From Table 4, we observe that: (1) Using α-divergence in the loss function instead of KL divergence better aligns with the encapsulation's partial order properties of text-image distribution embeddings. The experimental results also indicate that replacing α-divergence with KL divergence leads to performance degradation across all tasks. (2) Omitting the encapsulation loss, i.e., not using $\mathfrak{L}_{order}$ as defined in Equation 5 and relying primarily on $\mathfrak{L}_{con}$, results in performance degradation across all tasks. This is because not using encapsulation loss implies that the prior partial order of text and image cannot be utilized in hyperbolic space, thus losing the benefits introduced by hyperbolic geometry. (3) The model experiences a performance drop across all tasks when not using text-aware image representation. This is primarily due to the nature of medical image-text features. As discussed in the Introduction, most regions in medical images may differ in texture and morphology but not in clinical significance, while actual pathological changes are localized. The results also show that enhancing text-aware local features is meaningful for medical image-text alignment.



(a) CLIP (b) MERU (c) HYDEN

Figure 3: Distribution of embedding distances from [ROOT]: We embed 3858 testing images and text from the MIMIC-CXR v2 dataset using pre-trained CLIP, MERU, and HYDEN models.

Report: There is haziness of the hila with diffuse, but predominantly mid and lower lung heterogeneous opacities, consistent with moderate pulmonary edema, likely with both interstitial and alveolar components. The descending thoracic aorta is slightly tortuous. There may be small bilateral pleural effusions.



MERU: Texts retrieved from [IMAGE] -> [ROOT] traversal:

- Questions raise in the requisition to comment on presence of infiltrates versus edema cannot be answered in such detail on these four recent portable single view chest examinations.
- Any residual pneumothorax would be extremely small.
- Minimal left soft tissue air inclusions persist.

HYDEN: Texts retrieved from [IMAGE] -> [ROOT] traversal:

- Moderate bilateral layering pleural effusions, mild to moderate cardiac enlargement, opacification of the lower lobes and vascular congestion in the upper lobes reflect cardiogenic pulmonary edema.
- Moderate volume of bilateral pleural fluid is smaller, but there is collapse of both lower lobes and probably middle lobe as well.
- The extreme lung bases are not included on the image, given this limitation, small bilateral pleural effusions are seen, right greater than left, with bibasilar areas of consolidations that likely represent atelectasis.
- Although vascular injury cannot be definitively excluded by this exam, there appears to be slight interval decrease in the size of the mediastinum on the upright film, suggestive of a component of dependent left pleural effusion.
- Calcifications of the tracheobronchial tree as well as the aortic arch are noted.
- Chest pain, history of coronary artery disease.

Figure 4: Image traversals using MERU and HYDEN. We perform text retrieval at multiple steps while traversing from an image embedding to [ROOT] along the geodesic.

5.3 Qualitative Analysis

In this section, we explore the trained models to deduce the characteristics of the model in capturing the visual semantic hierarchy structure. The concept of 'Embedding distances from [ROOT]' was introduced by (Desai et al., 2023) to depict the generality differences between text and image embeddings in hyperbolic space. This concept highlights that in a representation space that effectively captures the visual semantic hierarchy, text embeddings are typically more general than image embeddings and, therefore, should be closer to the root node [ROOT].

Here, we visualize the differences in distance distributions between text and image embeddings. Given that our approach utilizes distribution embeddings, we specifically visualize the expectations of the distance distributions of text and image density embeddings. Figure 3 demonstrates that the distribution differences generated by our model lie between those produced by MERU and CLIP, with some overlapping distribution areas. This suggests that our model is capable of capturing the visual semantic hierarchy. Moreover, we perform text retrieval at multiple steps while traversing from an image embedding to the [ROOT]. The results are shown in Fig. 4. Compared with MERU, when we traverse from [ROOT] to the event, the retrieved texts remain relevant to the event but become progressively more specific and closely aligned with the context of the image. We speculate that, unlike the strong constraint imposed by entailment loss on the image-text norm values, the encapsulation loss between densities tends to make the density distributions of semantically similar image-text data closer, which explains why our method exhibits a wider span of norm distributions as in Figure 3.

6 Conclusion

In this paper, we propose a novel approach, HYDEN, for text-image representation learning based on hyperbolic density embeddings. It is a visual language representation learning method tailored specifically for medical data. Experimental results demonstrate the interpretability of our method and its superior performance compared to baseline methods across various zero-shot tasks and different datasets.

7 Limitations

Different from the entailment angle constraints based on point embeddings, we adopt the encapsulation order constraints between densities. In this paper, we use soft encapsulation loss via asymmetric divergence to measure between probability densities. For both the current Hyperbolic pseudo-Gaussian and Gaussian distributions in Euclidean space, introducing strict encapsulation losses remains a challenge and we will attempt to solve it in our future work.

Acknowledgments

This work is supported partially by the Beijing Natural Science Foundation (IS24051) and National Natural Science Foundation of China (62176068).

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. Publicly available clinical bert embeddings. In *ACL2019*.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. In *ACL2019*.

Ben Athiwaratkun and Andrew Gordon Wilson. Hierarchical density order embeddings. In *ICLR2018*.

Ahmad Bdeir, Kristian Schwethelm, and Niels Landwehr. 2024. Fully hyperbolic convolutional neural networks for computer vision. In *ICLR2024*.

Aleksandar Bojchevski and Stephan Günnemann. Deep gaussian embedding of graphs: Unsupervised inductive learning via ranking. In *ICLR2018*.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. 2023a. Prior: Prototype representation joint learning from medical images and reports. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21304–21314.

Pujin Cheng, Li Lin, Junyan Lyu, Yijin Huang, Wenhan Luo, and Xiaoying Tang. 2023b. Prior: Prototype representation joint learning from medical images and reports. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21361–21371.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. 2023. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23.

Ludovic Dos Santos, Benjamin Piwowarski, and Patrick Gallinari. 2017. Gaussian embeddings for collaborative filtering. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, page 1065–1068, New York, NY, USA. Association for Computing Machinery.

Xingcheng Fu, Yuecen Wei, Qingyun Sun, Haonan Yuan, Jia Wu, Hao Peng, and Jianxin Li. 2023. Hyperbolic geometric graph representation learning for hierarchy-imbalance node classification. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 460–468, New York, NY, USA. Association for Computing Machinery.

Antoine Gourru, Julien Velcin, Christophe Gravier, and Julien Jacques. 2022. Dynamic gaussian embedding of authors. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2109–2119. Association for Computing Machinery.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. 2021. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3922–3931.

Alistair E. W. 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. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *Preprint*, arXiv:1901.07042.

Kaggle. 2019. Society for imaging informatics in medicine: Siim-acr pneumothorax segmentation.

Diederik P. Kingma and Max Welling. 2019. An introduction to variational autoencoders. 12(4):307–392.

Emile Mathieu, Charline Le Lan, Chris J. Maddison, Ryota Tomioka, and Yee Whye Teh. Continuous hierarchical representations with poincaré variational auto-encoders. In *NIPS2019*.

Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, page 529–544, Berlin, Heidelberg. Springer-Verlag.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. 2022a. Joint learning of localized representations from medical images and reports. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, page 685–701. Springer-Verlag.

Philip Müller, Georgios Kaissis, Congyu Zou, and Daniel Rueckert. 2022b. Joint learning of localized representations from medical images and reports. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, page 685–701, Berlin, Heidelberg. Springer-Verlag.

Yoshihiro Nagano, Shoichiro Yamaguchi, Yasuhiro Fujita, and Masanori Koyama. A wrapped normal distribution on hyperbolic space for gradient-based learning. In *ICML2019*.

Maximilian Nickel and Douwe Kiela. 2017. Poincaré embeddings for learning hierarchical representations. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6341–6350, Red Hook, NY, USA. Curran Associates Inc.

Maximillian Nickel and Douwe Kiela. 2018. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3779–3788. PMLR.

Leandro Pardo. 2006. In *Statistical Inference Based on Divergence Measures*.

Chen Qian, Fuli Feng, Lijie Wen, and Tat seng Chua. 2021. Conceptualized and contextualized gaussian embedding. In *AAAI Conference on Artificial Intelligence*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *CVPR2021*.

Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F. Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, Mohamed Arselene Ayari, and Muhammad E. H. Chowdhury. 2020. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601.

Alfred Renyi. 1961. On measures of entropy and information. In *In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*.

Alistair Edward William; Pollard Tom Joseph; Lehman Li-Wei; Feng Mengling; Ghassemi Mohammad Mahdi; Moody Benjamin Edward; Szolovits Peter; Celi Leo Anthony G.; Mark Roger G Shen, Lu; Johnson. 2016. Mimic-iii, a freely accessible critical care database. In *scientific data*.

George Shih, Carol C. Wu, Safwan S. Halabi, Marc D. Kohli, Luciano M. Prevedello, Tessa S. Cook, Arjun Sharma, Judith K. Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, Ritu R. Gill, Myrna C.B. Godoy, Stephen Hobbs, Jean Jeudy, Archana Laroia, Palmi N. Shah, Dharshan Vummidi, Kavitha Yaddanapudi, and Anouk Stein. 2019. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1.

Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. 2024. Bioclip: A vision foundation model for the tree of life. In *acl2024*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *ICLR2016*.

Luke Vilnis and Andrew McCallum. 2014. Word representations via gaussian embedding. In *ICLR2014*.

Fuying Wang, Yuyin Zhou, Shujun Wang, Varut Vardhanabhuti, and Lequan Yu. 2024. Multi-granularity cross-modal alignment for generalized medical visual representation learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22. Curran Associates Inc.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017a. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017b. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471.

Qiyu Wu, Mengjie Zhao, Yutong He, Lang Huang, Junya Ono, Hiromi Wakaki, and Yuki Mitsufuji. 2023. Towards reporting bias in visual-language datasets: bimodal augmentation by decoupling object-attribute association. *Preprint*, arXiv:2310.01330.

Shu-Lin Xu, Yifan Sun, Faen Zhang, Anqi Xu, Xiu-Shen Wei, and Yi Yang. 2023. Hyperbolic space with hierarchical margin boosts fine-grained learning from coarse labels. In *Advances in Neural Information Processing Systems*, volume 36, pages 71263–71274. Curran Associates, Inc.

Yi.shi Xu, Dongsheng Wang, Bo Chen, Ruiying Lu, Zhibin Duan, and Mingyuan Zhou. 2022. Hyperminer: Topic taxonomy mining with hyperbolic embedding. In *Advances in Neural Information Processing Systems*, volume 35, pages 31557–31570. Curran Associates, Inc.

Shaoting Zhang and Dimitris Metaxas. 2023. On the challenges and perspectives of foundation models for medical image analysis. *Preprint*, arXiv:2306.05705.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. 2024. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *Preprint*, arXiv:2303.00915.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. 2022.

Contrastive learning of medical visual representations from paired images and text. *Preprint*, arXiv:2010.00747.

Xi Zhu, Xue Han, Shuyuan Peng, Shuo Lei, Chao Deng, and Junlan Feng. 2023. Beyond layout embedding: Layout attention with gaussian biases for structured document understanding. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

A Material

Hyberbolic Mapping The vector $\mu_{tan}^t = [0, \hat{\mu^t}]$ resides in $\mathbb{R}^{n+1}$ and belongs to the tangent space $T_{\mu_0}\mathbb{H}^n$ at the origin of the hyperboloid, $\mathcal{O}$. The norm $||\mu_{tan}^t||_{\mathcal{L}}$, which equals $||\hat{\mu^t}||_2$, ensures that the mapping preserves the distances inherent to the model's geometric structure. We apply the exponential map function (Eq. 2) to μ_{tan}^t, decomposing the transformation into two parts:

$$\begin{aligned}\cosh(\sqrt{c}||\mu_{tan}^t||_{\mathcal{L}})\mathcal{O} &= [\sqrt{1/c} \times \cosh(\sqrt{c}||\mu_{tan}^t||_{\mathcal{L}}), \mathbf{0}] \\ &= [\sqrt{1/c} \times \cosh(\sqrt{c}||\hat{\mu^t}||_2), \mathbf{0}]\end{aligned} \tag{6}$$

$$\begin{aligned}\frac{\sinh(\sqrt{c}||\mu_{tan}^t||_{\mathcal{L}})}{\sqrt{c}||\mu_{tan}^t||_{\mathcal{L}}} v_{tan} &= [0, \frac{\sinh(\sqrt{c}||\mu_{tan}^t||_{\mathcal{L}})}{\sqrt{c}||\mu_{tan}^t||_{\mathcal{L}}} \hat{\mu^t}] \\ &= [0, \frac{\sinh(\sqrt{c}||\hat{\mu^t}||_2)}{\sqrt{c}||\hat{\mu^t}||_2} \hat{\mu^t}]\end{aligned} \tag{7}$$

Combine the above two equations, we can get

$$\begin{aligned}&\mu^t = \exp_{\mu_0}(\mu_{tan}^t) \\ &= \left(\sqrt{1/c} \times \cosh(\sqrt{c}||\hat{\mu^t}||_2), \frac{\sinh(\sqrt{c}||\hat{\mu^t}||_2)}{\sqrt{c}||\hat{\mu^t}||_2} \hat{\mu^t}\right)\end{aligned} \tag{8}$$

Rényi α-Divergence is a general family of divergences that introduce varying degrees of zero-forcing penalty. The general form of the α-divergence for $\alpha \neq 0, 1$ is described as below,

$$D_\alpha(f||g) = \frac{1}{\alpha(\alpha-1)} \log\left(\int \frac{f(x)^\alpha}{g(x)^{\alpha-1}}, dx\right) \tag{9}$$

It is notable that as α approaches 0 or 1, the α-divergence converges to the KL divergence and the reverse KL divergence, respectively. For two multivariate Gaussians f and g, the Rényi α-Divergence can be expressed as:

$$\begin{aligned}D_\alpha(f||g) = &- \frac{1}{2\alpha(\alpha-1)} log \frac{det(\alpha\Sigma_g + (1-\alpha)\Sigma_f)}{(det(\Sigma_f)^{1-\alpha} \cdot det(\Sigma_g)^\alpha)} \\ &+ (\mu_f - \mu_g)^T(\alpha\Sigma_g + (1-\alpha)\Sigma_f)^{-1}(\mu_f - \mu_g)\end{aligned} \tag{10}$$

Here, the parameter α modulates the extent of zero forcing: minimizing $D_\alpha(f||g)$ for high α values results in f being concentrated towards the high-density regions of g. Conversely, for low α, f tends to have broader support, covering regions of g including those with low density.

MLR in the Lorentz model Given parameters $a_c \in R$ and $z_c \in \mathbb{R}^n$, the Lorentz MLR's output logit corresponding to class c and input $x = [x_t, x_s] \in \mathbb{H}^n$ is given by

$$\begin{aligned}&v_{z_c,a_c} = \frac{1}{\sqrt{-K}} sign(\alpha)\beta|sinh^{-1}(\sqrt{-K}\frac{\alpha}{\beta})|, \\ &where\ \alpha = cosh(\sqrt{-K}a_c) < z_c, x_s > -sinh(\sqrt{-K}a_c) \\ &and\ \beta = \sqrt{||cosh(\sqrt{-K}a_c)z_c||^2 - (sinh(\sqrt{-K}a_c||z_c||))^2}\end{aligned} \tag{11}$$

where K denotes the pre-defined varibale, and z_c, a_c represent classifier weight parameter. We can consider the expection value of image global desify embedding as x in above equation to derive the hyperbolic logit output of input image.

B Evaluation Tasks & Data

Image Classification: We evaluate the pre-trained model on three representative medical image classification tasks:

- **RSNA Pneumonia Dataset**(Shih et al., 2019): Comprising over 260,000 frontal chest radiographs collected by the Radiological Society of North America (RSNA). These images can be classified into a binary classification task: pneumonia vs. normal. For evaluation purposes, we randomly sample 4003 images for evaluation.
- **SIIM-ACR Pneumothorax Dataset**(Kaggle, 2019): Contains more than 12,000 frontal chest radiographs collected by the Society for Imaging Informatics in Medicine and the American College of Radiology (SIIM-ACR). Similar to the RSNA Pneumonia dataset, it is used for a binary classification task to determine the presence or absence of pneumothorax. We use all 10,675 images for evaluation.
- **Tuberculosis Chest X-ray Dataset**(Rahman et al., 2020): Tuberculosis is a chronic lung disease that occurs due to bacterial infection and is one of the top 10 leading causes of death. This dataset Comprise 700 infected and 3500 normal chest X-ray images. These images can be classified into a binary classification task: Tuberculosis vs. Normal. For evaluation purposes, we all of 4200 images for evaluation.
- **COVID-19 chest x-ray dataset**[4]: This dataset contains 21,165 images, and it consists of 3,616 COVID-19 positive cases, 10,192 normal cases, 6,012 Lung Opacity (Non-COVID lung infection) cases and 1,345 Viral Pneumonia images. , where 3,616. These images can be classified into a binary classification task: Covid vs. Non-Covid.
- **ChestXray14 Dataset**(Wang et al., 2017a): NIH ChestXray14 has 112,120 chest X-ray images with 14 disease labels from 30,805 unique patients. The official test set released by the NIH, comprising 22,433 images, are distinctively annotated by board certified radiologists. For multi-label evaluation, we only test on the official test set.

For the zero-shot classification task, we use all of data to evaluate the performance for each dataset. For the fine-tuning classification task, we split each datasets into 0.7/0.2/0.1 for train/valid/test.

Zero-shot Text-Image Retrieval: For pre-training methods akin to CLIP, text-image retrieval tests are standard practice. Following the practices of CLIP (Radford et al., 2021) and MERU (Desai et al., 2023), we also introduce downstream tasks for text-image retrieval. In medical imaging reports, the same diagnosis often has varied textual descriptions, making retrieval from image to text impractical. Thus, we do not use images to query text; instead, we use text to retrieve specific categories of images as described in (Zhang et al., 2022). For this purpose, we first construct a text-image retrieval evaluation dataset. As described in the multi-label classification task, ChestXray14 (Wang et al., 2017b) encompasses 14 different disease classes and one 'normal' class, totaling 15 categories. Based on these class labels, we randomly extract 100 images for each class (exclusive), forming the ChestXray14x100 dataset, which consists of 1,500 images. We then write representative text prompts for each of the 15 categories. During testing, for each query, we encode its text using the learned text encoder, then retrieve from the candidate images in a similar manner. This evaluation assesses not only the quality of the learned image representations but also the consistency between text and image representations.

Prompts Design:
To create the textual queries for each category on each evaluation task, we consulted a board-certified radiologist to draft at least five distinct sentences describing each abnormality as it would appear in radiology reports. Drawing inspiration from ConVIRT(Zhang et al., 2022), we established the following criteria: 1) The sentences must clearly describe the specific category without ambiguity and should not reference other categories. 2) The sentences must be varied and distinct from one another. 3) The sentences should avoid mentioning highly specific anatomical locations or rare clinical findings.

[4]https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database

Disease Category	Prompts for Text to Image Retrieval & Classification Tasks
Atelectasis	There is a linear opacity consistent with atelectasis.
Cardiomegaly	The cardiac silhouette is prominently enlarged, pointing to possible cardiomegaly.
Effusion	There is fluid accumulating in the pleural space indicative of pleural effusion.
Infiltration	The chest x-ray shows widespread ground-glass appearance.
Mass	A well-defined mass is present in the lung.
Nodule	A solitary pulmonary nodule is observed on the imaging.
Pneumonia	Airspace disease with lobar distribution points to possible pneumonia.
Pneumothorax	The presence of free air in the pleural space suggests a pneumothorax.
Consolidation	Dense pulmonary lesions consistent with consolidation.
Edema	The radiographic appearance is consistent with pulmonary edema.
Emphysema	The lungs appear overinflated, consistent with emphysema.
Fibrosis	The presence of traction bronchiectasis indicates lung fibrosis.
Pleural_Thickening	The pleura appears thickened, indicating pleural thickening.
Hernia	There is evidence of a diaphragmatic hernia.

Table 5: Example prompts for each categories in the evaluation tasks.

		ChestXray Multilabel		
Methods	Backbone	AUC	F1	ACC
Gloria	Resnet-50	0.655	**0.214**	0.681
CLIP	VIT-B	0.631	0.177	0.674
MERU	VIT-B	0.636	0.196	0.673
HYDEN	Resnet-50	0.661	0.194	0.704
HYDEN	VIT-B	**0.677**	0.208	**0.718**

Table 6: Zero-shot image classification

Finally, we aggregated the results and selected the best textual queries for each abnormality category. For reference, examples of the textual queries are presented in Table 5.

C Supplementary Experiments

In the main text, we present the overall performance of our proposed method and various baseline methods on binary classification tasks. Tab. 6 furthermore shows the performance of the models in the zero-shot multi-label classification task. Here, we use offcial test set of ChestXray14 Dataset comporising 22,433 images with labels. The results indicate that HYDEN consistently demon strates robust transfer classification performance as in binary classification evaluation.

In the main body of this paper, we have demonstrated image traversal results. Here, we add text traversal results which means the retrieval paths from input text to the root node using both MERU and HYDEN. We observe that, compared to MERU, our model retrieves more relevant terms, and moreover semantic evolves or drifts in a semantically continuous manner alongside traversal path.

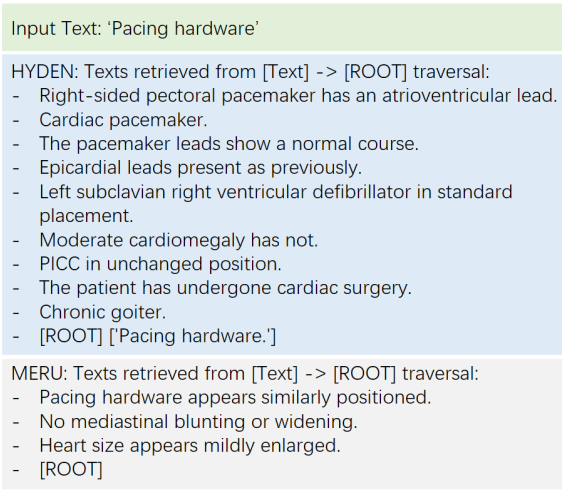


Figure 5: Text traversals using MERU and HYDEN. We perform text retrieval at multiple steps while traversing from input text embedding to [ROOT] along the geodesic.