Hyperbolic Concept Embedding Models for Interpretable Medical Image Diagnosis

Qihao Xu[1], Yadong Liu[1], Chengliang Liu[2], Yong Xu[1,*], Xiaoling Luo[3,*]
[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[2]PAMI Research Group, Department of Artificial Intelligence, University of Macau
[3]College of Computer Science and Software Engineering, Shenzhen University
xqh51199597@outlook.com, liuyadong221010@163.com, liucl1996@163.com
laterfall@hit.edu.cn, xiaolingluoo@outlook.com

Abstract

Although deep neural networks (DNNs) have shown strong performance in medical classification, their opaque reasoning limits interpretability and undermines the trustworthiness of clinical diagnoses. Concept bottleneck models (CBMs) introduce an intermediate concept layer, decomposing black-box prediction into an explicit reasoning path: image $\rightarrow$ concept $\rightarrow$ disease. However, concepts and their corresponding diagnostic results in medical image diagnosis are often hierarchical, inclusive, and semantically unevenly distributed. Most concept-based methods represent concepts in Euclidean space and treat them as isolated entities, hindering the modeling of their structural relationships and reducing diagnostic transparency. To address this, we introduce hyperbolic concept embedding models (HCEMs) that map medical concepts into a hyperbolic space better suited for hierarchical representation, enabling explicit modeling of complex concept relationships and their semantic associations with diseases. In these models, we regularize the hyperbolic concept embedding space using a positive–negative concept contrastive loss and a concept entailment cone loss. Furthermore, we introduce an intervention-aware regularization strategy for hyperbolic disease–concept embeddings to capture their dynamic relationships. It avoids rigid rule priors while improving diagnostic consistency and flexibility under concept interventions. Extensive experiments on four datasets demonstrate that our HCEMs achieve high accuracy in both concept and disease classification while enhancing interpretability and intervenability.

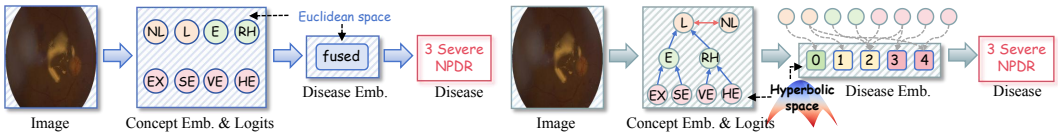
1 Introduction

Deep neural networks (DNNs) [1, 2, 3] have achieved remarkable success in medical image diagnosis, such as diabetic retinopathy (DR) grading [4, 5, 6] and skin disease classification [7, 8]. However, their black-box nature makes the decision process difficult to interpret, which limits clinical trust and practical deployment. It highlights the need to introduce Explainable Artificial Intelligence (XAI) into medical image analysis. Researchers have explored several post-hoc interpretability methods in medical imaging. For instance, Grad-CAM [9] visualizes the regions most relevant to the result, while LIME [10], SHAP [11], and counterfactual methods [12, 13] explain the contribution of input features. Nevertheless, these methods provide auxiliary or local explanations after the prediction, rather than offering inherently transparent reasoning.

To achieve this, concept-based models have emerged as a promising direction in interpretable medical image diagnosis. Representative models include Concept Bottleneck Models (CBMs) [14], which first predict human-understandable concepts (e.g., parts, attributes, relations) and then derive the final

*Corresponding authors: Yong Xu (laterfall@hit.edu.cn) and Xiaoling Luo (xiaolingluoo@outlook.com).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Concept Embedding Models **(b) Hyperbolic Concept Embedding Models**

Figure 1: Illustration of CEMs and HCEMs for interpretable medical disease diagnosis.

results from these concepts. Compared with other inherently interpretable methods [15, 16], CBMs not only provide a transparent reasoning pipeline but also support test-time concept interventions to rectify predictions. However, low-dimensional concepts contain limited information, resulting in lower task accuracy in CBMs. Thus, Concept Embedding Models (CEMs) [17] introduce positive and negative concept embeddings to increase concept information. Researchers have explored concept-based models from several perspectives, including predictive accuracy [18], interpretability [19, 20], intervention ability [21, 22], and no concept annotations [23, 24]. In medical imaging, some CBM variants have been applied to interpretable DR grading and skin disease analysis [25, 26, 27], with concepts typically considered as lesions or medical attributes. These findings underscore the broad research potential and clinical utility of the concept reasoning paradigm.

However, existing methods typically model concepts as independent semantic units, thereby overlooking the inherent hierarchical organization and semantic constraints of medical concepts and diseases. Taking DR grading as an example, retinal hemorrhage (RE) can be viewed as a high-level concept that includes finer-grained lesions such as microaneurysms (MA), hemorrhages (HE), and vitreous hemorrhage (VH), reflecting a coarse-to-fine hierarchy. In addition, the presence of RE suggests at least mild NPDR rather than a healthy state, reflecting the inclusion and exclusion relationships. These observations highlight the necessity of hierarchical concept modeling and explicit constraints.

In this paper, we introduce hyperbolic geometry to enhance concept representations and interpretability. Unlike Euclidean space, where volume grows polynomially with radius, hyperbolic space exhibits exponential volume growth, making it particularly suitable for representing hierarchical, tree-like structures. This geometric property naturally aligns with the hierarchical organization of medical concepts and diseases. Specifically, we propose **H**yperbolic **C**oncept **E**mbedding **M**odels (**HCEMs**) that embed concepts and diseases into a hyperbolic space (**Lorentz model**) for reasoning. Our HCEMs employ a geometric regularization strategy to model both concept–concept and disease–concept relationships. First, to capture complex relationships among medical concepts, such as inclusion and exclusion, we incorporate a concept knowledge graph and introduce two complementary objectives: positive–negative concept contrastive learning and concept entailment learning. The former improves concept discriminability by encouraging greater separation between positive and negative concept embeddings, while the latter preserves hierarchical concept relationships through hyperbolic entailment cones. Second, disease–concept relationships are inherently many-to-many. Individual diseases exhibit diverse manifestations, while the same medical concepts can be associated with multiple diseases. To capture these heterogeneous associations and improve concept intervention capability, we introduce an intervention-aware disease–concept entailment objective, optimized using a dynamic disease–concept entailment cone loss and a radius regularization loss.

In summary, our main contributions are as follows:

- We propose HCEMs, a hyperbolic concept embedding framework that leverages hyperbolic geometry to capture the hierarchical structures of medical concepts and diseases, enabling interpretable concept-to-disease reasoning.
- We develop a hyperbolic geometric regularization strategy that disentangles positive and negative concept embeddings, preserves hierarchical concept relations, and enhances intervenability through dynamic disease-concept entailment learning.
- Extensive experiments on four public medical datasets demonstrate that our HCEMs achieve high accuracy in both concept and disease classification, along with superior interpretability and intervenability.

2 Related Work

2.1 Concept-based Models

Concept-based models provide an interpretable framework for image classification by incorporating human-understandable concepts as intermediate representations. Concept Bottleneck Models (CBMs), introduced by Koh et al. [14], establish the foundation for this paradigm by explicitly separating concept prediction from downstream task inference. To balance predictive accuracy and interpretability, Espinosa et al. [17] proposed Concept Embedding Models (CEMs), which employ positive and negative concept embeddings to preserve richer semantic information. Mert et al. developed PCBM-h models that incorporate residual fitting to improve predictive performance. Kim et al. [28] and Xu et al. [18] introduced ProbCBMs and ECBMs through probabilistic and energy-based concept modeling. Galliamov et al. [29] further developed information-theoretic CBMs (CIBMs) to improve concept representation learning. To enhance concept intervenability, Sheth et al. [21] proposed Cooperative-CBMs (Coop-CBMs) with auxiliary concept supervision, while Zhang et al. [22] introduced DCBMs based on mutual information optimization. Other studies have explored logic-based [19, 20] and graph-based [27] concept reasoning to improve interpretability and capture relationships among concepts. Despite these advances, existing concept-based models rely on Euclidean representations, which do not inherently encode the hierarchical and asymmetric relationships among medical concepts. In contrast, our HCEMs leverage negatively curved hyperbolic geometry to capture hierarchical, inclusion-aware, and non-uniform semantic structures of medical concepts and diseases. By explicitly modeling concept–concept and disease–concept relationships in hyperbolic space, HCEMs provide a geometrically structured foundation for interpretable medical diagnosis and reliable concept intervention.

2.2 Explainable AI in Medical Image Diagnosis

Interpretability is a crucial research direction in medical image analysis, and we broadly categorize it into post-hoc and intrinsic interpretability. Early work mainly focused on post-hoc explanations, including saliency maps [4] and Shapley-value methods [30]. However, current diagnostic practice increasingly demands clinician-interpretable reasoning rather than isolated local explanations. Building on CBM, Gao et al. [25] developed an uncertainty-aware concept model (Evi-CEM), and Gao et al. [31, 32] extended this line with a logic-based framework for skin lesion diagnosis. In DR grading, Wen et al. [26] proposed CLAT, a language-enhanced, lesion-aware explainable Transformer. Xu et al. [27] proposed the concept graph. There are several efforts [20, 24, 33] for other medical domains.

2.3 Hyperbolic Representation Learning

Hyperbolic space has constant negative curvature, making it well-suited for modeling hierarchical and tree-like structures. Two widely used representations are the Poincaré ball model [34] and the Lorentz model [35]. Hyperbolic representations have been applied to various computer vision tasks, including image classification [36, 37], incremental learning [38], and few-shot learning [39]. In multimodal learning, CLIP-based methods [40, 41] employ hyperbolic contrastive learning to capture hierarchical image–text relationships. Recent studies have also extended hyperbolic geometry to fully connected layers [42] and convolutional neural networks [43], expanding its applicability to deep learning architectures. Motivated by these advances, we introduce hyperbolic geometry into concept-based medical diagnosis to model hierarchical concept–concept relations and flexible disease–concept associations.

3 Hyperbolic Concept Embedding Model

3.1 Hyperbolic Geometry Preliminary

This section briefly introduces the preliminaries of hyperbolic geometry, focusing on the Lorentz model. The Lorentz model provides a convenient realization of n-dimensional hyperbolic space by embedding it into $(n+1)$-dimensional Minkowski space and restricting the points to the upper branch of a two-sheeted hyperboloid. Let $\mathbf{u} = (u_0, \tilde{\mathbf{u}}) \in \mathbb{R}^{n+1}$, where u_0 denotes the time-like coordinate and $\tilde{\mathbf{u}} \in \mathbb{R}^n$ contains the spatial components. Given a positive constant $\kappa > 0$, corresponding to

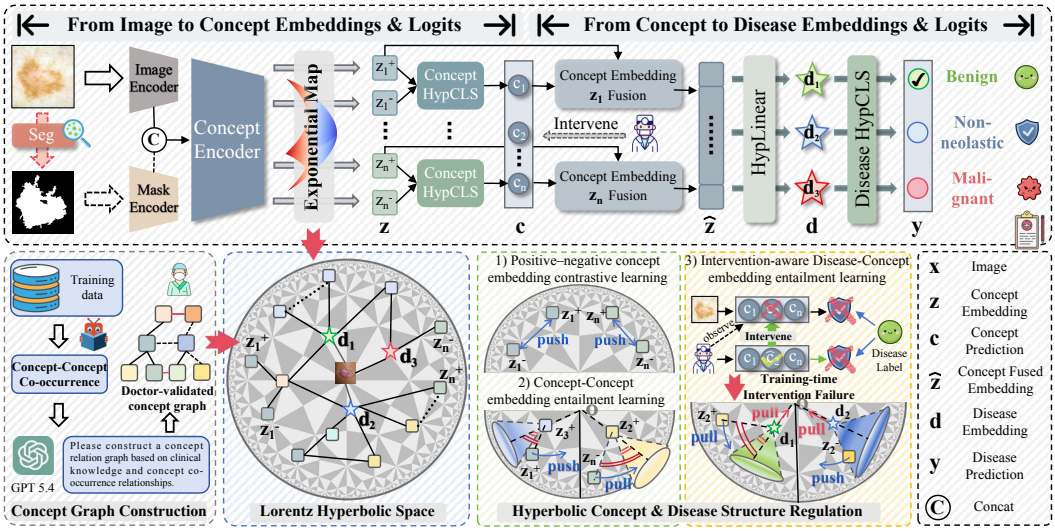


Figure 2: Overview of the HCEMs framework. It mainly consists of an interpretable reasoning process from inputs to concept embeddings and then to disease predictions, together with the proposed hyperbolic concept & disease structure regularization.

curvature $-\kappa$, the Lorentz manifold can be written as:

$$\mathbb{L}_\kappa^n = \left\{ \mathbf{u} \in \mathbb{R}^{n+1} \;\middle|\; \langle \mathbf{u}, \mathbf{u} \rangle_{\mathbb{L}} = -\frac{1}{\kappa},\ u_0 = \sqrt{\frac{1}{\kappa} + \|\tilde{\mathbf{u}}\|^2} \right\}. \tag{1}$$

Here, the Lorentzian inner product between two points $\mathbf{u} = (u_0, \tilde{\mathbf{u}})$ and $\mathbf{v} = (v_0, \tilde{\mathbf{v}})$ is defined by $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{L}} = -u_0 v_0 + \langle \tilde{\mathbf{u}}, \tilde{\mathbf{v}} \rangle_{\mathbb{E}}$, where $\langle \cdot, \cdot \rangle_{\mathbb{E}}$ denotes the standard Euclidean inner product. Based on this bilinear form, the geodesic distance on $\mathbb{L}_\kappa^n$ admits the closed-form expression

$$d_{\mathbb{L}}(\mathbf{u}, \mathbf{v}) = \frac{1}{\sqrt{\kappa}} \operatorname{arcosh}(-\kappa \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{L}}), \qquad \mathbf{u}, \mathbf{v} \in \mathbb{L}_\kappa^n. \tag{2}$$

For $\mathbf{u} \in \mathbb{L}_\kappa^n$, the tangent space $T_{\mathbf{u}}\mathbb{L}_\kappa^n$ is the local linear approximation of the manifold at $\mathbf{u}$. To satisfy the hyperboloid constraint, a tangent vector $\boldsymbol{\xi}$ must be Lorentz-orthogonal to $\mathbf{u}$, i.e., $T_{\mathbf{u}}\mathbb{L}_\kappa^n = \left\{ \boldsymbol{\xi} \in \mathbb{R}^{n+1} \;\middle|\; \langle \mathbf{u}, \boldsymbol{\xi} \rangle_{\mathbb{L}} = 0 \right\}$. The tangent space, equipped with the induced Riemannian metric, is isometric to an n-dimensional Euclidean space.

For $\mathbf{u} \in \mathbb{L}_\kappa^n$, the exponential map projects a tangent vector $\boldsymbol{\xi} \in T_{\mathbf{u}}\mathbb{L}_\kappa^n$ onto the manifold, while the logarithmic map transforms a point $\mathbf{v} \in \mathbb{L}_\kappa^n$ back to the tangent space at $\mathbf{u}$. The exponential and logarithmic maps establish a bijective correspondence between the tangent space and the hyperbolic manifold. They are defined as:

$$\exp_{\mathbf{u}}^{\kappa}(\boldsymbol{\xi}) = \cosh\left(\sqrt{\kappa}\, \|\boldsymbol{\xi}\|_{\mathbb{L}}\right) \mathbf{u} + \frac{\sinh(\sqrt{\kappa}\, \|\boldsymbol{\xi}\|_{\mathbb{L}})}{\sqrt{\kappa}\, \|\boldsymbol{\xi}\|_{\mathbb{L}}} \boldsymbol{\xi}, \tag{3}$$

$$\log_{\mathbf{u}}^{\kappa}(\mathbf{v}) = \frac{\operatorname{arcosh}(-\kappa \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{L}})}{\sqrt{(-\kappa \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{L}})^2 - 1}} \left(\mathbf{v} + \kappa \langle \mathbf{u}, \mathbf{v} \rangle_{\mathbb{L}} \mathbf{u}\right), \tag{4}$$

where the Lorentzian norm is $\|\boldsymbol{\xi}\|_{\mathbb{L}} = \sqrt{\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_{\mathbb{L}}}$. To transform Euclidean features into hyperbolic space, we treat them as tangent vectors at the origin $\mathbf{o} = (\sqrt{1/\kappa}, 0, \ldots, 0)^\top$, map them onto the hyperboloid via $\exp_{\mathbf{o}}^{\kappa}(\cdot)$, and conversely project hyperbolic points back to the tangent space at the same origin using $\log_{\mathbf{o}}^{\kappa}(\cdot)$ [44].

3.2 Interpretable Reasoning Pipeline

The goal of interpretable medical image diagnosis is to provide disease results $\mathbf{y} \in \{0, 1\}^K$ and symptom (medical concept) predictions $\mathbf{c} \in \{0, 1\}^N$ from a medical image, together with a transparent and clinically understandable diagnostic process. N and K denote the numbers of disease classes

and concept classes. In this paper, we propose hyperbolic concept embedding models (HCEMs) to construct and constrain the complex hierarchical concept–concept and disease–concept relationships. The core components of the model are hyperbolic concept modeling and hyperbolic disease-concept embedding structure regulation. In this section, we present the interpretable reasoning process.

Given an input $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, our HCEMs employ a pretrained model to generate a key-region mask $\mathbf{s} \in \mathbb{R}^{H\times W\times 1}$, and then uses shared encoders F_I, F_S to extract global appearance features and regional prior information from the image and mask, followed by channel-wise concatenation for feature fusion. Here, the segmentation operation is optional. The fused features are mapped into Euclidean concept embeddings by the concept encoder F_C. Subsequently, the concept embeddings $\{\mathbf{z}_i\}_{i=1}^N \in \mathbb{L}_\kappa^n$ are mapped into Lorentz hyperbolic space via $\exp_{\mathbf{o}}^\kappa(\cdot)$, enabling better modeling of hierarchical relationships in this exponentially expanding space. Following [] we split the concept embeddings along the channel dimension into positive and negative concept embeddings, representing evidence for and against the presence of each concept, respectively. The above process can be formulated:

$$\{\mathbf{z}_i\}_{i=1}^N = \{[\mathbf{z}_i^+, \mathbf{z}_i^-]\}_{i=1}^N = \exp_{\mathbf{o}}^\kappa(F_C(\mathrm{Concat}(F_I(\mathbf{x}),\, F_S(\mathbf{s})))) \,. \tag{5}$$

Through the hyperbolic concept classifier, the model obtains concept predictions $\{\mathbf{c}_i\}_{i=1}^N = \{HCC([\mathbf{z}_i^+, \mathbf{z}_i^-])\}_{i=1}^N$. Once both semantic embeddings and concepts are obtained, we construct the recomposed hyperbolic concept embedding $\hat{\mathbf{z}}_i$ for i-th concept as a weighted mixture of $\mathbf{z}_i^+$ and $\mathbf{z}_i^-$:

$$\hat{\mathbf{z}}_i \triangleq \mathbf{c}_i \odot \mathbf{z}_i^+ \oplus (1-\mathbf{c}_i) \odot \mathbf{z}_i^- = \exp_{\mathbf{o}}^\kappa\Big(\mathbf{c}_i \log_{\mathbf{o}}^\kappa\big(\mathbf{z}_i^+\big) + (1-\mathbf{c}_i)\log_{\mathbf{o}}^\kappa\big(\mathbf{z}_i^-\big)\Big). \tag{6}$$

Where $\odot$ and $\oplus$ denote the Lorentz scalar multiplication and addition. Next, the recomposed concept embeddings are processed by disease-specific hyperbolic encoders, to obtain disentangled hyperbolic disease embeddings $\mathbf{d}_i$. Each encoder consists of multiple hyperbolic fully connected layers. Finally, the model obtains the disease prediction $\mathbf{y}$ through a hyperbolic disease classifier. In this way, HCEMs achieve an interpretable reasoning process $\mathbf{x} \to (\mathbf{z}, \mathbf{c}) \to \mathbf{d} \to \mathbf{y}$ that mimics clinicians by identifying symptoms and then making the final diagnosis.

Fully Hyperbolic Linear Layer In our HCEM, both the hyperbolic embedding transformation (e.g., $\hat{\mathbf{z}} \to \mathbf{d}$) and the concept/disease classifiers adopt fully hyperbolic linear layers. Taking the disease classifier $\mathcal{D}$ as an example, given the input $\mathbf{d}_i$, the output of the disease $\mathbf{y}_i$ (unnormalized) can be defined as:

$$\mathbf{y}_i = \mathcal{D}(\mathbf{d}_i) = \frac{1}{\sqrt{\kappa}}\,\mathrm{asinh}\left(\sqrt{\kappa}\cdot\frac{\langle \mathbf{d}_i, \mathbf{w}_k\rangle_{\mathbb{L}}}{\sqrt{\|\mathbf{w}_{ks}\|^2 - w_{k0}^2}}\right) + b_k. \tag{7}$$

Here, $\mathbf{w}_k = (w_{k0}, \mathbf{w}_{ks})$ represents the Lorentz hyperplane parameters, where w_{k0} and $\mathbf{w}_{ks}$ denote the time-like and space-like components, respectively. The normalization term $\beta_k = \sqrt{\|\mathbf{w}_{ks}\|^2 - w_{k0}^2}$ ensures that $\mathbf{w}_k$ defines a valid space-like hyperplane. b_k is a learnable bias term. The Lorentz hyperplane classifier directly operates on hyperbolic embeddings without tangent-space projection, preserving geometric information during prediction.

3.3 Hyperbolic Concept & Disease Structure Regularization

To model complex concept-concept relationships, such as inclusion and exclusion, in hyperbolic space, we first construct a concept knowledge graph. Based on this graph, we then constrain the distances between positive-negative concept embeddings and the concept–concept entailment cone angles, thereby regularizing hyperbolic concept structure. Moreover, to construct hierarchical disease-concept relationships and enhance concept intervention ability, we further propose an intervention-aware disease-concept entailment cone objective.

Concept Knowledge Graph Construction The concept knowledge graph construction is a hybrid pipeline. Specifically, we first compute the co-occurrence relationships among concept states, including both presence and absence, from the training data. Then, GPT-5.4 [45] jointly analyzes these co-occurrence statistics and clinical knowledge to generate the initial graph, which clinicians further validate. Formally, we define the concept knowledge graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each concept embedding $\mathbf{z}_i$ is split into a positive node $\mathbf{z}_i^+$ and a negative node $\mathbf{z}_i^-$, yielding $\mathcal{V} = \{\mathbf{z}_1^+, \mathbf{z}_1^-, \ldots, \mathbf{z}_N^+, \mathbf{z}_N^-\}, |\mathcal{V}| = 2N$. The edge set $\mathcal{E}$ encodes concept–concept relations, including

inclusion, exclusion, and no direct relevance. The graph is represented by an adjacency matrix $\mathbf{A}$:

$$\mathbf{A}_{ij} = \begin{cases} 1, & \text{if node } i \text{ includes node } j, \\ -1, & \text{if node } i \text{ and node } j \text{ are mutually exclusive}, \\ 0, & \text{if node } i \text{ and node } j \text{ have no direct relevance}. \end{cases} \tag{8}$$

The detailed graph generation results are provided in A.4.

Hyperbolic Concept Structure Regularization Concept relations in hyperbolic space are the repulsion between positive and negative embeddings of the same concept and the complex semantic relations among different concepts. To model the relationship between positive and negative concepts and enrich their semantic representations, we design a contrastive learning objective of positive-negative concept embeddings. The hyperbolic contrastive loss is as follows:

$$\mathcal{L}_{\text{PNC}} = \frac{1}{N} \sum_{i=1}^{N} \exp\left(-d_{\mathbb{L}}\left(\mathbf{z}_i^+, \mathbf{z}_i^-\right)/\tau\right). \tag{9}$$

Where $\tau = 2.0$ is the temperature constant. Under this setting, the positive and negative concept embeddings are moderately pushed apart in the hyperbolic space. To model hierarchical relationships among concepts, we introduce a hyperbolic concept–concept entailment learning objective. Given two concept nodes $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j \in \mathcal{V}$, $\mathbf{A}_{ij} = 1$ indicates that $\tilde{\mathbf{z}}_i$ entails $\tilde{\mathbf{z}}_j$, i.e., $\tilde{\mathbf{z}}_i \Rightarrow \tilde{\mathbf{z}}_j$, whereas $\mathbf{A}_{ij} = -1$ denotes that they are exclusive, i.e., $\tilde{\mathbf{z}}_i \perp \tilde{\mathbf{z}}_j$. For the Lorentz model, the half-aperture of the entailment cone [40] rooted at $\tilde{\mathbf{z}}_i$ is defined as:

$$\omega(\tilde{\mathbf{z}}_i) = \arcsin\left(\frac{2\gamma}{\sqrt{\kappa}\|\bar{\mathbf{z}}_i\|}\right), \tag{10}$$

where $\gamma = 0.1$ is a constant, and $\bar{\mathbf{z}}_i$ denotes the spatial component of $\tilde{\mathbf{z}}_i$. The exterior angle of $\tilde{\mathbf{z}}_j$ relative to the entailment cone rooted at $\tilde{\mathbf{z}}_i$ is computed as:

$$\phi(\tilde{\mathbf{z}}_j, \tilde{\mathbf{z}}_i) = \arccos\left(\frac{z_{j0} + z_{i0}\kappa\langle\tilde{\mathbf{z}}_j, \tilde{\mathbf{z}}_i\rangle_{\mathbb{L}}}{\|\bar{\mathbf{z}}_i\|\sqrt{\left(\kappa\langle\tilde{\mathbf{z}}_j, \tilde{\mathbf{z}}_i\rangle_{\mathbb{L}}\right)^2 - 1}}\right). \tag{11}$$

Here, z_{i0} and z_{j0} denote the time components of $\tilde{\mathbf{z}}_i$ and $\tilde{\mathbf{z}}_j$. A smaller $\phi(\tilde{\mathbf{z}}_j, \tilde{\mathbf{z}}_i)$ indicates that $\tilde{\mathbf{z}}_j$ is closer to the entailment cone of $\tilde{\mathbf{z}}_i$, and $\tilde{\mathbf{z}}_j$ is considered entailed when this angle is smaller than the cone aperture. Thus, the concept-concept entailment loss $\mathcal{L}_{\text{CCE}}$ is defined as:

$$\mathcal{L}_{\text{CCE}} = \frac{1}{|\mathcal{V}|} \sum_{\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}_j \in \mathcal{V}} \max\left(\mathbf{A}_{ij}\left(\phi(\tilde{\mathbf{z}}_j, \tilde{\mathbf{z}}_i) - \eta\omega(\tilde{\mathbf{z}}_i)\right), 0\right). \tag{12}$$

Here, $\eta = 0.7$ is a threshold for the half-aperture $\omega(\tilde{\mathbf{z}}_i)$. The angle constraint models directional entailment, while the sematic distance constraint separates positive and negative concept embeddings. Finally, we obtain the integrated hyperbolic concept regularization loss: $\mathcal{L}_{\text{HCR}} = \mathcal{L}_{\text{PNC}} + \mathcal{L}_{\text{CCE}}$.

Intervention-aware Hyperbolic Disease-Concept Regularization In our reasoning process $(\mathbf{z}, \mathbf{c}) \rightarrow \mathbf{d} \rightarrow \mathbf{y}$, the predicted concept $\mathbf{c}$ serves as a key neuron that significantly influences $\mathbf{y}$. Therefore, clinicians can correct the final diagnosis by revising the erroneous $\mathbf{c}$. Existing methods [31] often rely on fixed medical-rule reasoning constraints to regulate the diagnostic chain, which reduces the flexibility and accuracy of intervention. To model the hierarchical relationships between disease and concept embeddings and enhance intervention capability, we propose a flexible optimization objective, namely the intervention-aware disease–concept entailment loss. This loss pulls the corrected disease-relevant concepts into the entailment cone of the ground-truth disease embedding, thereby encouraging the model to respond correctly to concept-level interventions.

Specifically, during training, we substitute concept values $\mathbf{c}_j^{\text{int}} \sim \begin{cases} \mathcal{U}(0, 0.5), & \hat{\mathbf{c}}_j = 0 \\ \mathcal{U}(0.5, 1), & \hat{\mathbf{c}}_j = 1 \end{cases}$ into the incorrectly predicted concept entries and perform inference again, where $\mathcal{U}(a, b) = \text{Uniform}(a, b)$ and $\hat{\mathbf{c}}_j$ is binary concept label. We define the set of intervention-failed samples in this batch: $\mathcal{F} = \left\{n \mid \hat{\mathbf{y}}_n \neq \mathbf{y}_n,\ \mathbf{y}_n^{\text{int}} \neq \mathbf{y}_n\right\}$. Then, the latent disease-concept relation is derived from intervened labels, where $\mathbf{B}_{kj} = 1$ denotes entailment between disease node $\mathbf{d}_k$ and concept node $\tilde{\mathbf{z}}_j$, and

Table 1: Performance comparison of disease classification across three random seeds. The best results are shown in **bold**, and the second-best results are underlined. (mean±std: %)

Method (Venue)	DDR		MFIDDR		Derm7pt		SkinCon	
	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1	Accuracy	Macro F1
Multi-Task CVPR'19	81.28±0.82	71.44±0.20	73.74±0.37	55.23±2.10	84.56±1.65	80.80±1.72	76.53±0.93	56.78±2.35
CBM ICML'20	80.78±1.44	67.78±1.96	73.02±0.40	55.94±0.62	84.12±1.87	79.92±1.89	75.34±1.59	56.05±1.90
CEM NeurIPS'22	81.57±1.59	71.73±1.51	74.11±0.45	56.37±1.27	84.57±1.64	82.14±1.05	76.28±1.21	56.45±2.17
ProbCBM ICML'23	81.01±2.12	67.92±1.51	73.75±0.85	55.12±1.27	81.09±2.23	78.97±3.51	74.13±1.61	54.21±0.98
Coop-CBM NeurIPS'23	81.47±0.98	69.20±1.04	72.76±0.87	56.01±1.13	84.51±1.32	78.69±1.82	76.01±0.73	56.39±0.61
Evi-CEM MICCAI'24	82.23±1.26	71.88±1.34	74.21±1.12	56.02±1.20	85.79±1.46	83.10±1.68	78.51±1.52	57.23±1.29
DCBM TPAMI'25	81.97±1.89	72.09±1.45	74.70±0.79	56.32±1.10	86.04±1.99	83.24±1.56	76.02±1.78	56.53±1.61
VLM-GCR AAAI'26	83.10±0.74	73.58±0.86	75.47±0.61	56.78±1.51	85.24±1.43	82.05±1.68	76.41±1.35	56.28±0.97
CIBM ICLR'26	80.65±1.11	70.46±1.68	73.26±0.98	56.01±2.02	83.83±1.98	80.02±2.68	75.26±1.07	55.07±1.03
HCEMs No mask	84.30±0.64	74.82±0.59	75.89±0.45	58.09±0.72	87.21±0.73	84.29±0.92	78.49±0.73	57.96±0.50
HCEMs Ours	**84.52±0.79**	**75.01±0.92**	**76.02±0.58**	**58.34±0.67**	**87.52±0.89**	**84.76±0.89**	**79.09±0.81**	**58.32±1.30**

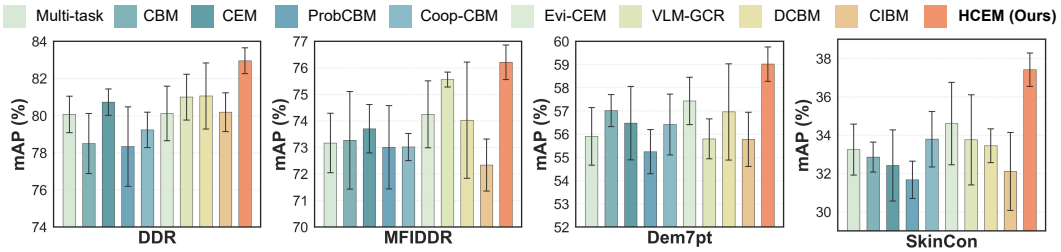


Figure 3: Performance comparison of mAP for concept recognition across three random seeds in four datasets. Error bars indicate the standard deviation. (mean±std: %)

$\mathbf{B}_{kj} = -1$ denotes exclusion. Finally, combining Equs. (10) and (11), the intervention-aware disease-concept entailment loss $\mathcal{L}_{\text{IDCE}}$ is defined as:

$$\mathcal{L}_{\text{IDCE}} = \frac{1}{|\mathcal{F}|} \sum_{n \in \mathcal{F}} \frac{1}{|\mathcal{V}_n^{dc}|} \sum_{(\mathbf{d}_k, \tilde{\mathbf{z}}_j) \in \mathcal{V}_n^{dc}} \max\left(\mathbf{B}_{kj}\left(\phi(\tilde{\mathbf{z}}_j, \mathbf{d}_k) - \eta\omega(\mathbf{d}_k)\right), 0\right). \tag{13}$$

where $\mathcal{V}_n^{dc} = \{\mathbf{z}_1^+, \mathbf{z}_1^-, \ldots, \mathbf{z}_N^+, \mathbf{z}_N^-, \mathbf{d}_1, \ldots, \mathbf{d}_K\}$ denotes the disease-concept embedding pairs for sample n. In addition, the distance to the origin in hyperbolic space represents the hierarchical radius. We impose a radial MAE loss on disease embeddings to control their distances to the hyperbolic origin: $\mathcal{L}_{\text{Rad}} = \frac{1}{K}\sum_{k=1}^{K} |d_{\mathbb{L}}(\mathbf{d}_k, \mathbf{o}) - \rho|$. Here, $\mathbf{o}$ is the center of the hyperbolic space, and $\rho = 0.9$ denotes the target radius of the disease embedding, which controls its hierarchical position. The disease-concept embedding regularization loss is: $\mathcal{L}_{\text{DCR}} = \mathcal{L}_{\text{IDCE}} + \mathcal{L}_{\text{Rad}}$.

3.4 Training objective

We integrate the task loss and the hyperbolic concept and disease structure regularization loss to obtain the final training objective:

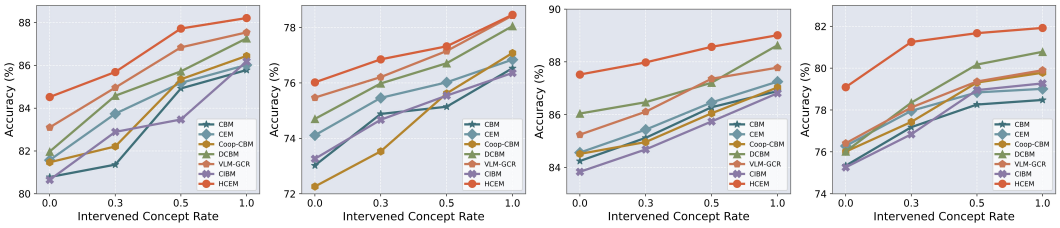
$$\mathcal{L} = \mathcal{L}_{\text{Task}} + \alpha\mathcal{L}_{\text{HCR}} + \beta\mathcal{L}_{\text{DCR}}, \tag{14}$$

where $\mathcal{L}_{\text{Task}} = \mathcal{L}_{\text{Concept}} + \mathcal{L}_{\text{Disease}}$ includes both the concept prediction loss and the disease prediction loss, and α, β are the balancing coefficients.

4 Experiments

4.1 Experimental Setup

Datasets We conduct experiments on four public medical datasets annotated with both disease and concept labels. Specifically, the DDR dataset [46] is a DR grading dataset (13,673 fundus images) acquired using various imaging devices and exhibiting diverse quality. We select 2,334 images with six lesion concept annotations for our experiments. The MFIDDR dataset [27, 47]



(a) DDR (b) MFIDDR (c) Dem7pt (d) Skincon

Figure 4: Performance comparison of disease diagnosis accuracy under different test-time concept intervention rates on four datasets.

is a large-scale DR grading dataset containing 34,452 images, with five grading categories and annotations for six lesion concepts. Derm7pt [48] is a melanoma diagnosis dataset containing 1,011 dermoscopic images annotated with the 7-point checklist, whereas SkinCon [49, 25] is a concept dataset of 3,230 clinical skin images densely annotated by dermatologists. They contain 2/26 and 3/22 disease-category/attribute-level labels, respectively. For MFIDDR, we split the official training set into training and validation sets in an 8:2 ratio. For the other three datasets, we randomly split the data into training, validation, and test sets with a 7:1:2 ratio. All experiments utilize the above setup and more details of these datasets are provided in A.3.

Comparison Methods We benchmark our method against nine representative baselines, consisting of one non-interpretable multi-task model [50] and eight concept-based methods, including CBM [14], CEM [17], ProbCBM [28], Coop-CBM [21], Evi-CEM [25], VLM-GCR [27], DCBM [22], and CIBM [29]. Among them, Evi-CEM and VLM-GCR are the best concept-based methods for skin lesion and DR diagnosis, respectively. Brief descriptions of these methods are given in Related Work.

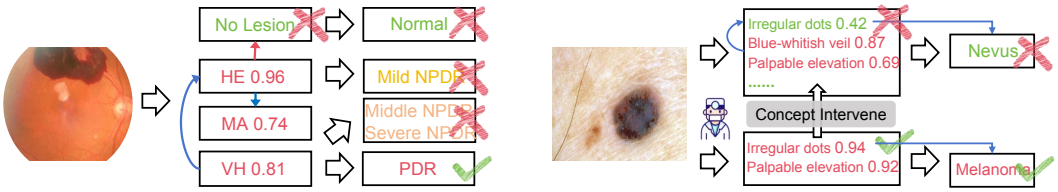
Evaluation Metrics Concept-based interpretable medical image diagnosis involves three aspects of evaluation: disease classification, concept recognition, and disease classification under concept test-time intervention [17]. For disease classification, we use Accuracy for overall evaluation and Macro-F1 to assess performance on imbalanced classification. For concept recognition, we adopt mean average precision (mAP) as the evaluation metric to treat each concept fairly.

Implementation Details All experiments run on a single NVIDIA RTX 4090 GPU. Each method is trained for 100 epochs using 224×224 medical images with a batch size of 16. We optimize all models using Adam, starting from a learning rate of 5×10^{-5} and applying cosine annealing throughout training. For a fair comparison, all methods share the same pretrained VGG16 backbone [51] and are evaluated under three random seeds. Following [17, 25, 27], we report the results of the checkpoint that attains the best disease diagnosis accuracy on the validation set.

4.2 Comparison to State-of-the-Arts

Analysis of Diagnostic Performance To evaluate the performance of the compared methods, we report the disease diagnosis and medical concept classification results in Table 1 and Fig. 3. We obtain the following observations. First, our HCEMs consistently outperform both the multi-task model and existing concept-based models in disease classification, regardless of whether mask priors are used. Specifically, HCEMs improve all metrics across the four datasets by 0.52%-1.56% compared with the second-best methods. Second, HCEMs improve the mAP of concept prediction by 1.02%-2.09% across four datasets. These results demonstrate the effectiveness of HCEMs in learning informative concept representations and capturing disease–concept relationships for accurate medical diagnosis.

Intervention and Interpretability Analysis Concept-based models can correct disease predictions by intervening in concept predictions. We evaluate intervention effects across seven methods in Fig. 4, where DCBM and Coop-CBM are the previous strongest intervenable models. Due to the introduction of hyperbolic concept modeling and intervention-aware hyperbolic disease-concept embedding constraints, our HCEMs consistently achieve the best disease accuracy under various concept intervention ratios. In Fig. 5(b), we observe that intervening on a specific concept can correct an erroneous disease prediction. Regarding interpretability, compared with CEM and its variants, our model captures the implicit concept–concept and disease–concept relationships. Taking Fig.



(a) Visualization of the Reasoning Process (b) Visualization of Test-Time Intervention

Figure 5: Visualization of our HCEMs reasoning and test-time concept intervention.

5(a) as an example, the presence of VH, together with the latent VH–HE and VH–PDR entailment relationships as diagnostic rules, forms a complete reasoning chain of disease–concept rule.

4.3 Ablation Study

To further investigate the effectiveness of each design component within our HCEM, we conduct ablation experiments in this section. We first analyze the roles of masks and hyperbolic concept modeling in interpretable medical image diagnosis, and then examine the effects of different regularization terms on diagnostic results. From Table 2, we observe that integrating masks and hyperbolic modeling improves both disease classification and concept recognition. On Dem7pt, disease accuracy increases by 0.7% and 0.8%, while concept mAP improves by 0.5% and 1.0%, respectively. The regularization terms for hyperbolic concept and disease embeddings are also indispensable. For the hyperbolic distance constraints on positive and negative concept embeddings, 'HCEMs *w/o* $\mathcal{L}_{\text{PNC}}$' shows that removing these constraints leads to approximately 2.0% and 1.5% drops in disease accuracy and concept mAP. To validate the effectiveness of the concept entailment cone loss, we report the results in 'HCEMs *w/o* $\mathcal{L}_{\text{CCE}}$'. Removing $\mathcal{L}_{\text{CCE}}$ leads to performance decreases of 1.8% and 1.5% in the corresponding metrics on Dem7pt. Intervention-aware disease-concept entailment loss $\mathcal{L}_{\text{IDCE}}$ is also shown to be important. Compared with statically constraining disease-concept relations '$\mathcal{L}_{\text{IDCE}}$ *w/o Interv.*', it improves all metrics on two datasets. Overall, these results demonstrate the indispensable role of hyperbolic modeling and regularization terms in both disease diagnosis and concept classification.

Table 2: Ablation study on MFIDDR and Dem7pt in three seeds. The baseline is CEM, while 'D.' and 'C.' denote disease and concept, respectively.

Method	MFIDDR		Dem7pt	
	D. Acc	C. mAP	D. Acc	C. mAP
Baseline	0.741	0.737	0.846	0.565
Baseline *w mask*	0.746	0.739	0.852	0.572
Baseline *w hyperbolic*	0.748	0.745	0.854	0.575
HCEM	**0.760**	**0.762**	**0.875**	**0.591**
HCEMs *w/o* $\mathcal{L}_{\text{PNC}}$	0.743	0.740	0.860	0.570
HCEMs *w/o* $\mathcal{L}_{\text{CCE}}$	0.754	0.748	0.857	0.576
HCEMs *w/o* $\mathcal{L}_{\text{IDCE}}$	0.748	0.743	0.855	0.569
HCEMs *w/o* $\mathcal{L}_{\text{Rad}}$	0.754	0.752	0.868	0.582
$\mathcal{L}_{\text{IDCE}}$ *w/o Interv.*	0.756	0.754	0.862	0.573

4.4 Hyperbolic Embedding Visualization

We visualize the learned hyperbolic space in lower dimensions by CO-SNE [52] to see if the disease and concept embeddings are distributed in a proper semantic hierarchy. In Fig. 6, we present the concept embeddings $\{\hat{\mathbf{z}}_i\}_{i=1}^{6}$ and the disease embeddings $\{\mathbf{d}_i\}_{i=1}^{5}$(corresponding to the predictions) in $\mathbb{L}_\kappa^2$ for 468 images from the DDR test set. The disease embeddings of the five DR grades are closer to the center than the concept embeddings, reflecting the hierarchical structure of disease-concept relationships. The concepts and their positive/negative counterparts are distinctly separated, and the underlying exclusion and entailment relations are also clear. For instance, the semantic closeness of VH to HE is clearly reflected, along with the exclusion relations between positive and negative concepts. The gray line denotes the latent disease-concept correspondence, e.g., no lesion corresponds to Normal, while VO and VH are associated with PDR. These results reflect the latent diagnostic rules for DR grading.

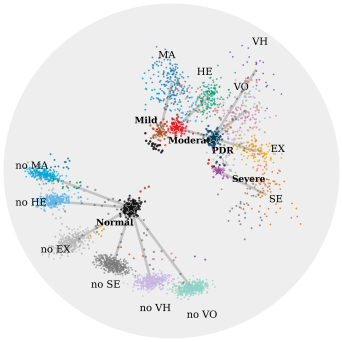


Figure 6: Hyperbolic disease & concept embedding visualization.

4.5 Hyperparameter Analysis

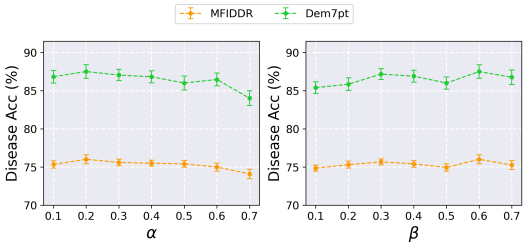


Figure 7: Experimental results of the loss balancing coefficients across three seeds. (mean±std: %)

We investigate the sensitivity of HCEMs to the regularization coefficients α and β on the MFIDDR and Derm7pt datasets, as shown in Fig. 7. These coefficients control the relative contributions of hyperbolic concept structure regularization and intervention-aware disease–concept regularization, respectively. The results show that both datasets achieve the highest disease diagnosis accuracy when $\alpha = 0.2$ and $\beta = 0.6$. Smaller regularization weights may provide insufficient constraints on hierarchical relationships, whereas excessively large weights may overemphasize structural consistency at the expense of discriminative learning. The consistent optimal configuration across both datasets suggests that balancing concept-level semantic relationships and disease–concept hierarchical constraints is important for effective diagnostic reasoning. Accordingly, we adopt $\alpha = 0.2$ and $\beta = 0.6$ throughout our experiments.

Table 3: Training HCEMs in different settings of curvature.

κ	MFIDDR		Derm7pt	
	D. Acc.	C. mAP	D. Acc.	C. mAP
Fixed param.				
1.0	75.3	75.1	86.4	58.2
0.7	75.0	74.9	85.3	57.7
0.3	74.7	75.8	86.5	58.6
0.1	75.8	75.4	87.1	58.8
Learnable param.				
1.0	**76.0**	**76.2**	**87.5**	**59.1**

We further investigate the influence of the Lorentz curvature parameter κ, which determines the geometric properties of hyperbolic space and its capacity to represent hierarchical relationships. As reported in Table 3, different fixed curvature values lead to noticeable variations in model performance on MFIDDR and Derm7pt, suggesting that a predefined curvature may not adequately capture the distinct hierarchical structures of different datasets. To address this limitation, we treat κ as a learnable parameter, initialized at 1.0 and constrained within $[0.1, 10.0]$, allowing the model to adapt its geometric structure during training. Compared with fixed-curvature settings, the learnable-curvature approach consistently achieves superior performance in both disease diagnosis and concept prediction across the two datasets. This improvement indicates that adaptive curvature provides greater flexibility in modeling complex concept dependencies and disease–concept hierarchies, thereby enhancing the discriminative capability of hyperbolic representations. These findings support the effectiveness of incorporating learnable curvature into HCEMs for interpretable medical image diagnosis.

5 Conclusion

In this paper, we propose HCEMs, hyperbolic concept embedding models for interpretable medical image diagnosis. HCEMs embed medical concepts and diseases into hyperbolic space and constrain concept-concept and disease-concept relationships through contrastive, entailment cone, and intervention-aware objectives. Experiments on four medical datasets show that HCEMs improve disease & concept recognition and intervention performance while providing transparent reasoning.

Acknowledgments

This work was supported by the Shenzhen Medical Research Fund [No. C2601012], the National Natural Science Foundation of China [No. 62502320], the Natural Science Foundation of Guangdong Province [No. 2025A1515010184], and the project of Shenzhen Science and Technology Innovation Committee [No. JCYJ20240813141424032].

References

[1] Chengliang Liu, Yuanxi Que, Wai Keung Wong, Yabo Liu, and Xiaoling Luo. Multi-view hilbert curve-based hierarchical information aggregation for incomplete multimodal alzheimer's disease diagnosis. *IEEE Transactions on Medical Imaging*, 2026.

[2] Yongting Hu, Haoran Li, Qihao Xu, Yuxin Lin, Jiahua Shi, Jun Shen, Yong Xu, and Xiaoyan Dou. A cnn-injected transformer network with lesion reconstruction for multi-view diabetic retinopathy grading. *Medical Image Analysis*, page 104178, 2026.

[3] Xiaoling Luo, Qihao Xu, Huisi Wu, Chengliang Liu, Zhihui Lai, and Linlin Shen. Like an ophthalmologist: Dynamic selection driven multi-view learning for diabetic retinopathy grading. In *AAAI*, volume 39, pages 19224–19232, 2025.

[4] Rui Sun, Yihao Li, Tianzhu Zhang, Zhendong Mao, Feng Wu, and Yongdong Zhang. Lesion-aware transformers for diabetic retinopathy grading. In *CVPR*, pages 10938–10947, 2021.

[5] Xiaoling Luo, Chengliang Liu, Waikeung Wong, Jie Wen, Xiaopeng Jin, and Yong Xu. Mvcinn: multi-view diabetic retinopathy detection using a deep cross-interaction neural network. In *AAAI*, volume 37, pages 8993–9001, 2023.

[6] Xiaoling Luo et al. A lesion-fusion neural network for multi-view diabetic retinopathy grading. *IEEE Journal of Biomedical and Health Informatics*, 29(5):3184–3193, 2025.

[7] Mohamed Akrout et al. Diffusion-based data augmentation for skin disease classification: Impact across original medical datasets to fully synthetic images. In *MICCAI*, pages 99–109. Springer, 2023.

[8] Matthew Groh et al. Deep learning-aided decision support for diagnosis of skin disease across skin tones. *Nature Medicine*, 30(2):573–583, 2024.

[9] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626, 2017.

[10] Zeren Tan, Yang Tian, and Jian Li. Glime: general, stable and local lime explanation. In *NeurIPS*, volume 36, pages 36250–36277, 2023.

[11] Hugh Chen, Ian C Covert, Scott M Lundberg, and Su-In Lee. Algorithms to estimate shapley value feature attributions. *Nature Machine Intelligence*, 5(6):590–601, 2023.

[12] Dylan Slack, Anna Hilgard, Himabindu Lakkaraju, and Sameer Singh. Counterfactual explanations can be manipulated. In *NeurIPS*, volume 34, pages 62–75, 2021.

[13] Valentyn Melnychuk, Dennis Frauen, and Stefan Feuerriegel. Causal transformer for estimating counterfactual outcomes. In *ICML*, pages 15293–15329. PMLR, 2022.

[14] Pang Wei Koh et al. Concept bottleneck models. In *ICML*, pages 5338–5348. PMLR, 2020.

[15] Dawid Rymarczyk et al. Interpretable image classification with differentiable prototypes assignment. In *ECCV*, pages 351–368. Springer, 2022.

[16] Meike Nauta, Jörg Schlötterer, Maurice Van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *CVPR*, pages 2744–2753, 2023.

[17] Espinosa Zarlenga et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *NeurIPS*, volume 35, pages 21400–21413, 2022.

[18] Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *ICLR*, 2024.

[19] Ciravegna et al. Logic explained networks. *Artificial Intelligence*, 314:103822, 2023.

[20] Shuting Xu, Junlin Hou, and Hao Chen. Explain any pathological concept: Discovering hierarchical explanations for pathology foundation models. In *MICCAI*, pages 216–225. Springer, 2025.

[21] Ivaxi Sheth and Samira Ebrahimi Kahou. Auxiliary losses for learning generalizable concept-based models. In *NeurIPS*, volume 36, pages 26966–26990, 2023.

[22] Rui Zhang, Xingbo Du, Junchi Yan, and Shihua Zhang. The decoupling concept bottleneck model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1250–1265, 2025.

[23] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *ICLR*, 2023.

[24] Lijie Hu et al. Towards multi-dimensional explanation alignment for medical classification. In *NeurIPS*, volume 37, pages 129640–129671, 2024.

[25] Yibo Gao et al. Evidential concept embedding models: Towards reliable concept explanations for skin disease diagnosis. In *MICCAI*, pages 308–317. Springer, 2024.

[26] Chi Wen, Mang Ye, He Li, Ting Chen, and Xuan Xiao. Concept-based lesion aware transformer for interpretable retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 44(1):57–68, 2025.

[27] Qihao Xu, Xiaoling Luo, Yuxin Lin, Chengliang Liu, Yongting Hu, Jinkai Li, Xinheng Lyu, and Yong Xu. Vision-language models guided graph concept reasoning for interpretable diabetic retinopathy diagnosis. In *AAAI*, volume 40, pages 27314–27322, 2026.

[28] Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *ICML*, pages 16521–16540. PMLR, 2023.

[29] Karim Galliamov, Syed M Ahsan Kazmi, Adil Khan, and Adín Ramírez Rivera. Concepts' information bottleneck models. In *ICLR*, 2026.

[30] Jiawen Zheng et al. Causality-inspired graph neural network for interpretable strabismus subtype classification. In *MICCAI*, pages 107 – 116. Springer, September 2025.

[31] Yibo Gao et al. Learning concept-driven logical rules for interpretable and generalizable medical image classification. In *MICCAI*, pages 291–300. Springer, 2025.

[32] Yibo Gao, Hangqi Zhou, Zheyao Gao, Bomin Wang, Shangqi Gao, Sihan Wang, and Xiahai Zhuang. Adaptive concept rule learning for interpretable and generalizable medical image classification. *Medical Image Analysis*, page 104299, 2026.

[33] Midhat Urooj, Ayan Banerjee, and Sandeep Gupta. Neuro-guard: Neuro-symbolic generalization and unbiased adaptive routing for diagnostics-explainable medical ai. In *CVPR*, pages 21–29, 2026.

[34] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *NeurIPS*, volume 30, 2017.

[35] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pages 3779–3788. PMLR, 2018.

[36] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, pages 6840–6849, 2023.

[37] Zhigang Tu, Zhengbo Zhang, Jia Gong, Junsong Yuan, and Bo Du. Informative sample selection model for skeleton-based action recognition with limited training samples. *IEEE Transactions on Image Processing*, 34:7335–7346, 2025.

[38] Tao Hu, Lan Li, Zhen-Hao Xie, and Da-Wei Zhou. Hierarchical semantic tree anchoring for clip-based class-incremental learning. *arXiv preprint arXiv:2511.15633*, 2025.

[39] Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. Curvature generation in curved spaces for few-shot learning. In *ICCV*, pages 8691–8700, 2021.

[40] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023.

[41] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *ICLR*, 2025.

[42] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on pattern analysis and machine intelligence*, 44(12):10023–10044, 2021.

[43] Neil He, Menglin Yang, and Rex Ying. Lorentzian residual neural networks. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pages 436–447, 2025.

[44] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pages 6418–6428, 2020.

[45] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

[46] Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501: 511–522, 2019.

[47] Xiaoling Luo, Qiaojian Zheng, Chengliang Liu, Yangneng Chen, Xiaoyan Dou, Tianyi Luo, Yingying Wen, Chao Huang, Yong Xu, and Jie Wen. A large-scale multi-view fundus images dataset and benchmark for diabetic retinopathy grading. *Information Fusion*, page 104363, 2026.

[48] Jeremy Kawahara, Sara Daneshvar, Giuseppe Argenziano, and Ghassan Hamarneh. Seven-point checklist and skin lesion classification using multitask multimodal neural nets. *IEEE Journal of Biomedical and Health Informatics*, 23(2):538–546, 2018.

[49] Matthew Groh et al. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *CVPR*, pages 1820–1828, 2021.

[50] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *CVPR*, pages 1871–1880, 2019.

[51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[52] Yunhui Guo, Haoran Guo, and Stella X Yu. Co-sne: Dimensionality reduction and visualization for hyperbolic data. In *CVPR*, pages 21–30, 2022.

A Appendix

A.1 Descriptions of Symbols

We summarize the meanings of the main symbols in Table 4.

A.2 Comparative Analysis of HCEMs Against Concept-based Interpretable Methods

CBM and its variants adopt a concept bottleneck structure, where human-defined medical concepts are first predicted and then used for disease diagnosis. It provides good interpretability and intervention ability. CEM and its variants further introduce concept embeddings, using positive and negative concept states to represent richer concept semantics. With almost no loss of interpretability, they achieve stronger representation capacity and diagnostic performance than conventional CBM. However, these methods usually model concepts independently, making it difficult to capture hierarchical, mutually exclusive, and entailment relations among concepts and between concepts and diseases. Moreover, their Euclidean-space modeling paradigm lacks explicit constraints on medical diagnostic hierarchies and clinical prior relations. In contrast, our HCEMs embed concepts and diseases into hyperbolic

Table 4: Main symbols used in this paper.

Symbol	Description	Symbol	Description
$\mathbf{x}$	Input medical image.	$\log_{\mathbf{o}}^{\kappa}(\cdot)$	Logarithmic map to the tangent space.
$\mathbf{s}$	Key-region mask or segmentation prior.	$\mathbf{z}_i$	Hyperbolic embedding of the i-th concept.
N	Number of medical concept classes.	$\mathbf{z}_i^+$	Positive embedding of the i-th concept.
K	Number of disease classes.	$\mathbf{z}_i^-$	Negative embedding of the i-th concept.
$\mathbf{c} = \{c_i\}_{i=1}^N$	Predicted concept logits or scores.	$\hat{\mathbf{z}}_i$	Recomposed hyperbolic concept embedding.
$\hat{\mathbf{c}} = \{\hat{c}_i\}_{i=1}^N$	Binary concept labels.	$\tilde{\mathbf{z}}_i$	Concept node in the knowledge graph.
$\mathbf{y}$	Disease prediction.	$\mathbf{d}_k$	Hyperbolic embedding of the k-th disease.
$\hat{\mathbf{y}}$	Ground-truth disease label.	$\mathcal{G} = (\mathcal{V}, \mathcal{E})$	Concept knowledge graph.
$\mathbb{L}_\kappa^n$	Lorentz hyperbolic space with curvature $-\kappa$.	$\mathcal{V}$	Node set of concept embeddings.
κ	Curvature parameter of the Lorentz model.	$\mathcal{E}$	Edge set of concept relations.
$\mathbf{o}$	Origin of the Lorentz manifold.	$\mathbf{A}$	Concept–concept relation matrix.
$\langle\cdot,\cdot\rangle_{\mathbb{L}}$	Lorentzian inner product.	$\mathbf{B}$	Disease–concept relation matrix.
$d_{\mathbb{L}}(\cdot,\cdot)$	Geodesic distance in Lorentz space.	$\omega(\tilde{\mathbf{z}}_i)$	Half-aperture of the entailment cone.
$\exp_{\mathbf{o}}^{\kappa}(\cdot)$	Exponential map to the Lorentz manifold.	$\phi(\tilde{\mathbf{z}}_j, \tilde{\mathbf{z}}_i)$	Exterior angle relative to the cone.

space, which is more suitable for representing the hierarchical diagnostic structure between medical concepts and diseases.

Notably, under the same input size and visual encoder, concept-based models usually have only minor differences in computational cost and parameter size. This is also why computational efficiency has been less emphasized in this line of interpretable research. Here, we report the number of parameters and FLOPs of all compared models in Table 5.

A.3 Dataset Visualization and Details

Fig. 8 shows representative samples from the four datasets, together with their binary concept annotations and disease labels. DDR and MFIDDR focus on diabetic retinopathy diagnosis, where six lesion concepts, including EX, HE, MA, SE, VH, and VO, are annotated and associated with five DR grades. Derm7pt and SkinCon provide richer dermatological concept annotations, containing 26 and 22 binary concepts, respectively, for skin disease diagnosis. As summarized in Table 6, these datasets cover both ophthalmic and dermatological scenarios, enabling a comprehensive evaluation of concept-based diagnosis.

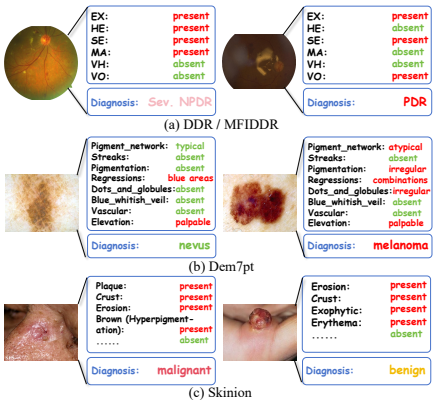


Figure 8: Sample visualization of the four datasets.

A.4 Concept Knowledge Graph

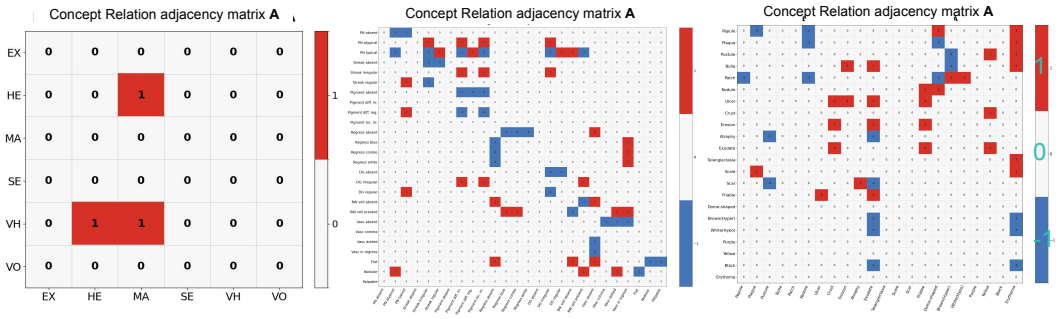
In Fig. 9, we illustrate the concept of relation adjacency matrices $\mathbf{A}$ on four datasets. Rows denote source concepts and columns denote target concepts. Red entries indicate positive entailment or co-occurrence relations, blue entries indicate negative or mutually exclusive relations, and white entries denote uncertain or unrelated concept pairs. For DDR and MFIDDR, the concept knowledge

Table 5: Model complexity comparison. Params and FLOPs are estimated using 224×224 input images. FLOPs are calculated as twice the number of MACs. "+Mask" denotes that both the original image and the mask prior are used as inputs.

Metric	Multi-task	CBM	CBM+Mask	CEM	CEM+Mask	ProbCBM	Coop-CBM	Evi-CEM	VLM-GCR	DCBM	HCEM(No Mask)	HCEM
Backbone	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16	VGG16
Input	Image	Image	Image+Mask	Image	Image+Mask	Image	Image	Image	Image	Image	Image	Image+Mask
Params (M)	14.76	14.72	19.67	14.56	18.72	14.39	14.89	15.72	18.64	16.71	16.15	19.77
FLOPs (G)	30.70	30.69	41.39	30.25	40.51	30.01	29.62	31.69	32.75	32.20	31.02	40.90

Table 6: Binary concepts and disease labels for the Derm7pt, DDR, MFIDDR, and SkinCon datasets.

Data	# Concept classes	Binary concepts	# Disease classes	Diseases
Derm7pt	26	**Pigment network:** absent, atypical, typical; **Streaks:** absent, irregular, regular; **Pigmentation:** absent, diffuse irregular, diffuse regular, localized irregular; **Regression structures:** absent, blue areas, white areas, combinations; **Dots and globules:** absent, irregular, regular; **Blue-whitish veil:** absent, present; **Vascular structures:** absent, comma, dotted, within regression; **Elevation:** flat, nodular, palpable	2	nevus, melanoma
SkinCon	22	**Primary lesions:** papule, plaque, pustule, bulla, patch, nodule, ulcer; **Secondary lesions:** crust, erosion, atrophy, exudate, telangiectasia, scale, scar, friable, dome-shaped; **Pigment and erythema:** brown, white, purple, yellow, black, erythema	3	non-neoplastic, benign, malignant
DDR	6	**Lesions:** exudates (EX), hemorrhages (HE), microaneurysms (MA), soft exudates (SE), vitreous hemorrhage (VH), vitreous opacity (VO)	5	no DR, mild NPDR, moderate NPDR, severe NPDR, PDR
MFIDDR	6	**Lesions:** exudates (EX), hemorrhages (HE), microaneurysms (MA), soft exudates (SE), vitreous hemorrhage (VH), vitreous opacity (VO)	5	no DR, mild NPDR, moderate NPDR, severe NPDR, PDR



(a) DDR/MFIDDR (b) Dem7pt (c) SkinCon

Figure 9: The concept of relation adjacency matrices in four datasets.

graph is relatively sparse and primarily captures progressive DR lesion dependencies, such as VH $\rightarrow$ HE, MA$\}$. For Dem7pt and SkinCon, the matrices contain richer dermatological relations, including both supportive dependencies among abnormal visual patterns and suppressive relations between mutually exclusive morphology, color, or lesion states. These directed matrices encode asymmetric medical priors rather than simple symmetric co-occurrence, providing structured knowledge for concept-level regularization and interpretable diagnosis.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

Justification: Refer to Introduction.

Guidelines:

- The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Refer to the limitations in Supplementary materials.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide the corresponding discussion.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.

- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The datasets used in this study are publicly available, and we will release the experimental code in the future.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will release the experimental code and data in the future.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.

- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Refer to experimental setting in main text.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes. We add the standard deviation.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Refer to implement details in Appendix, or as Supplemental material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Introduction.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This paper is not involved any high risk data or model.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: CC-BY-NC 4.0 license.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [No]

    Justification: The LLM is used only for writing and editing.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.