Hyperbolic Relational Prompts for Intersectional Fairness in Medical VLMs

Supplementary Material

1. Theoretical Analysis: Information Gain and Geometric Tightness

In this section, we provide a theoretical analysis to better understand the advantages of the fairness-aware relational prompting (FRP) framework. Our analysis suggests that the proposed approach alleviates representational limitations commonly encountered in Euclidean baselines. In particular, we show that (1) relational aggregation can reduce uncertainty in latent demographic structure, and (2) hyperbolic geometry provides a principled mechanism for mitigating the representational collapse of fine-grained intersectional subgroups, which is important for effective fairness enforcement.

1.1. Preliminaries and Problem Setup

Let the input space be $\mathcal{X}$ and the sensitive attribute space be $\mathcal{S}$. We consider the dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, \mathbf{s}_i)\}_{i=1}^N$, where $\mathbf{s}_i$ represents hierarchical intersectional identities (e.g., specific demographic combinations).

Assumption 1 (Latent δ-Hyperbolic Structure)**.** *The data generation process is governed by a latent metric space $(\mathcal{H}, d_{\mathcal{H}})$ which is δ-hyperbolic (tree-like) [7, 9].*

1. ***Hierarchical Identity:*** *The sensitive attributes $\mathbf{s}_i$ correspond to nodes in a discrete hierarchy embedded in $\mathcal{H}$ with branching factor b and depth L. The total number of intersectional subgroups is $M \approx b^L$.*
2. ***Geometric SBM:*** *The connectivity graph A follows a Geometric Stochastic Block Model, where the edge probability between nodes i and j depends on their latent distance: $P(A_{ij} = 1) \propto \exp(-d_{\mathcal{H}}(\mathbf{h}_i, \mathbf{h}_j)/\tau)$, where $\mathbf{h}_i \in \mathcal{H}$ is the true latent position [7].*
3. ***Noisy Observation:*** *The node feature $\mathbf{x}_i$ is a noisy observation of $\mathbf{h}_i$, such that $I(\mathbf{x}_i; \mathbf{h}_i) < H(\mathbf{h}_i)$.*

Assumption 2 (Bounded Representation Space)**.** *For numerical stability and regularization, the learned representation $\mathbf{z}$ is constrained to a compact ball of radius R in the embedding manifold $\mathcal{M}^d$ (either $\mathbb{R}^d$ or $\mathbb{H}^d$). We require the decoder $q_\phi(y|z)$ to be K-Lipschitz continuous w.r.t. the manifold metric $d_{\mathcal{M}}$.*

1.2. Relational Information Gain

We first formalize why the graph structure A is critical for recovering the latent identity $\mathbf{h}_i$, which dictates both the target label and fairness constraints.

Lemma 1 (Variance Reduction via Aggregation)**.** *Let $\hat{\mathbf{h}}_i(\mathbf{x}_i)$ be the estimator of the latent position based solely on node features, and $\hat{\mathbf{h}}_i^{agg}(\mathbf{x}_i, \mathcal{N}(i))$ be the estimator using the ego-graph. Under Assumption 1, the relational aggregation strictly increases the mutual information with the target Y conditioned on the sensitive group S:*

$$I(Z_{agg}; Y|S) \geq I(Z_{node}; Y|S) + \Delta_{SNR}, \quad \Delta_{SNR} > 0. \tag{1}$$

Proof. Consider the neighbors $\mathcal{N}(i)$ as auxiliary views of the latent variable $\mathbf{h}_i$. Since edges follow the Geometric SBM, neighbors are sampled from a local neighborhood of $\mathbf{h}_i$. Let $\mathbf{x}_j = \phi(\mathbf{h}_i) + \epsilon_j$ for $j \in \mathcal{N}(i) \cup \{i\}$, where ϵ_j is independent noise. The single-node estimator has variance proportional to $\text{Var}(\epsilon)$. The aggregated estimator, averaging over degree $k = |\mathcal{N}(i)|$, yields a variance proportional to $\frac{1}{k+1}\text{Var}(\epsilon)$. Since the mutual information $I(Z; \mathbf{h})$ is monotonically decreasing with the estimator's variance (for Gaussian channels, $I \propto \log(1 + \text{SNR})$), aggregation effectively denoises the position $\mathbf{h}_i$. By the Data Processing Inequality on the chain $Y \leftarrow \mathbf{h} \rightarrow Z$, a tighter estimate of $\mathbf{h}$ implies a tighter lower bound on $I(Z; Y)$. □

1.3. Geometric Tightness and Capacity Gap

Our theoretical analysis further suggests that, even with sufficient information, Euclidean embeddings may struggle to separate the fine-grained subgroups required for fairness without substantial overlap.

Theorem 1.1 (The Euclidean Bottleneck)**.** *Let $N_{\mathcal{M}}(\epsilon, R)$ denote the packing number (maximum number of disjoint ϵ-balls) within a ball of radius R in manifold $\mathcal{M}^d$. For a hierarchy of depth L and branching factor b, if the complexity satisfies*

$b^L \gg (R/\epsilon)^d$, *then for any Euclidean embedding* $f_{\mathbb{E}} : \mathcal{X} \to \mathbb{R}^d$, *there exists a non-zero probability of* ***Intersectional Collapse****:*

$$P(\exists i, j : \mathbf{s}_i \neq \mathbf{s}_j \wedge d_{\mathbb{E}}(\mathbf{z}_i, \mathbf{z}_j) < \epsilon/K) \to 1. \tag{2}$$

Conversely, for a Hyperbolic embedding $f_{\mathbb{H}}$ *with curvature* $c < 0$*, the collision probability approaches 0 provided* R *scales linearly with* L*.*

Proof. **Step 1: Capacity Requirement.** To satisfy the Lipschitz constraint (Assumption 2) and distinguish $M = b^L$ intersectional groups, the embedding space must support a packing of M balls of radius $r = \epsilon/K$. Thus, we require $N_{\mathcal{M}}(r, R) \geq b^L$.

Step 2: Euclidean Limitation. In Euclidean space $\mathbb{R}^d$, the packing number scales polynomially with the radius:

$$N_{\mathbb{R}^d}(r, R) \leq \left(\frac{R}{r}\right)^d. \tag{3}$$

As the hierarchy depth L increases, the required capacity b^L grows exponentially. When $b^L > (R/r)^d$, by the Pigeonhole Principle, distinct groups must be mapped within distance r. Due to Lipschitz continuity, the decoder output $q(y|z)$ becomes indistinguishable for these groups.

Step 3: Hyperbolic Sufficiency. In Hyperbolic space $\mathbb{H}^d$, the volume of a ball grows exponentially with radius: $\text{Vol}(R) \sim e^{(d-1)R}$ [12]. The packing number satisfies:

$$N_{\mathbb{H}^d}(r, R) \geq C \cdot \exp\left((d-1)R\right). \tag{4}$$

This exponential growth matches the complexity of the hierarchy. By setting $R \approx \frac{L \ln b}{d-1}$, $\mathbb{H}^d$ can isometrically embed the leaves of the tree with arbitrary precision ϵ [12, 13].

Step 4: Impact on Risk. Using Fano's Inequality, the lower bound on the classification error P_e for intersectional groups is:

$$P_e \geq 1 - \frac{\max I(Z; S)}{\log M} \approx 1 - \frac{d \log(R/r)}{L \log b}. \tag{5}$$

In Euclidean space, as $L \to \infty$, $P_e \to 1$. In Hyperbolic space, capacity matches demand, allowing $P_e \to 0$. □

Corollary 1 (Intersectional Blindness and Fairness)**.** *The "Intersectional Collapse" in Euclidean space implies that the model treats distinct sensitive subgroups (e.g.,* S_a *= Black-Female and* S_b *= Black-Male) as geometrically identical. Under such collapse, any fairness regularizer* $\mathcal{L}_{fair}$ *becomes ill-posed because the gradients* $\nabla_z \mathcal{L}_{fair}$ *for these groups cancel out or conflict. Hyperbolic space preserves the latent separability, providing the necessary geometric support for the Fairness Modulator to function effectively.*

1.4. Discussion

Our analysis reveals that the superiority of FRP stems from solving a **dual optimization problem**: the graph layer (Lemma 1) maximizes the mutual information lower bound via denoising, while the hyperbolic layer (Theorem 1.1) expands the geometric capacity to accommodate the exponential complexity of intersectional fairness constraints. This aligns with previous findings on the superiority of hyperbolic spaces for hierarchical data representation [4, 10]. □

2. Mathematical Preliminaries: Hyperbolic Geometry

In this section, we provide the detailed mathematical formulations for the Poincaré ball model of hyperbolic geometry, which serves as the foundation for our structural prior A_{base} and the hyperbolic graph layer (HGL) described in Section 3.2 and 3.3 of the main paper. We operate in a hyperbolic space with constant negative curvature $-c$ ($c > 0$).

2.1. The Poincaré Ball Model

The Poincaré ball model $(\mathbb{B}_c^d, g_x^{\mathbb{B}})$ is defined on the open ball $\mathbb{B}_c^d = \{x \in \mathbb{R}^d : \|x\| < 1/\sqrt{c}\}$. The Riemannian metric tensor $g_x^{\mathbb{B}}$ at a point $x \in \mathbb{B}_c^d$ is conformal to the Euclidean metric $g^{\mathbb{E}}$ [4]:

$$g_x^{\mathbb{B}} = \lambda_x^2 g^{\mathbb{E}}, \quad \text{where } \lambda_x = \frac{2}{1 - c\|x\|^2} \tag{6}$$

is the conformal factor. The induced geodesic distance $d_c(\cdot, \cdot)$ between two points $x, y \in \mathbb{B}_c^d$ is given by:

$$d_c(x, y) = \frac{2}{\sqrt{c}} \text{arctanh}\left(\sqrt{c}\| - x \oplus_c y\|\right) \tag{7}$$

This distance function serves as the core metric for computing the structural similarity in our base adjacency matrix A_{base}.

2.2. Möbius Addition

Standard Euclidean vector addition is not closed in the Poincaré ball. Instead, we use the Möbius addition $\oplus_c$, which is essential for aggregating features in the HGL (Eq. 13 in the main paper) [3, 4]:

$$x \oplus_c y = \frac{(1 + 2c\langle x, y\rangle + c\|y\|^2)x + (1 - c\|x\|^2)y}{1 + 2c\langle x, y\rangle + c^2\|x\|^2\|y\|^2} \tag{8}$$

Note that when $c \to 0$, this operation recovers the standard Euclidean addition.

2.3. Mappings between Tangent and Hyperbolic Spaces

To perform operations like linear transformations or attention mechanisms (which are well-defined in vector spaces), we map features between the hyperbolic manifold $\mathbb{B}_c^d$ and the tangent space $\mathcal{T}_x\mathbb{B}_c^d$ at a reference point x. For simplicity, we often perform operations in the tangent space at the origin $\mathcal{T}_0\mathbb{B}_c^d \cong \mathbb{R}^d$ [3].

Logarithmic Map ($\log_0^c$): Maps points from the hyperbolic manifold to the tangent space at the origin. This is used in Eq. 10 ($Z_{tan} = \log_0^c(Z^{\mathbb{H}})$) and Eq. 14 of the main paper:

$$\log_0^c(x) = \frac{2}{\sqrt{c}}\text{arctanh}(\sqrt{c}\|x\|)\frac{x}{\|x\|} \tag{9}$$

Exponential Map ($\exp_0^c$): Maps tangent vectors back to the hyperbolic manifold. This is used to project the initial features (Section 3.3) and the updated features (Eq. 13) back to the ball:

$$\exp_0^c(v) = \tanh\left(\frac{\sqrt{c}}{2}\|v\|\right)\frac{v}{\sqrt{c}\|v\|} \tag{10}$$

For numerical stability, we clip the norm of vectors to ensure they remain strictly within the ball boundary (i.e., $\|x\| \leq 1/\sqrt{c} - \epsilon$).

2.4. Hyperbolic Graph Layer Operations

Based on the above definitions, our HGL performs feature aggregation as follows, inspired by Hyperbolic GCNs [3]:

1. **Lift:** Input Euclidean features x_E are mapped to hyperbolic space: $x_H = \exp_0^c(x_E)$.
2. **Aggregating in Tangent Space:** Since direct weighted averaging is computationally expensive in hyperbolic space (requiring Fréchet mean), we approximate it by mapping to the tangent space, aggregating using the unified adjacency matrix A, and then mapping back.
3. **Linears and Activation:** Linear transformations $M \cdot x + b$ in hyperbolic space are implemented as:

$$f_{linear}(x) = \exp_0^c\left(M \cdot \log_0^c(x) + b\right) \tag{11}$$

This formulation ensures that our operations respect the hyperbolic geometry while maintaining computational efficiency compatible with standard deep learning frameworks.

3. Algorithm Pseudocode

The detailed training procedure of our proposed FRP framework is summarized in **Algorithm 1**.

4. Comprehensive Experimental Analysis

In this section, we provide a comprehensive analysis of the dataset distribution, detailed implementation settings, and additional experimental results that were omitted from the main text due to space constraints. These experiments further validate the robustness and scalability of our proposed FRP framework.

4.1. Data Analysis: Systematic Bias and Evaluation Rigor

To objectively assess the source of algorithmic bias, we analyze the distributional properties of the FairVLMed [**?**] and Harvard-GF [8] datasets. Figure 1 visualizes the sample distribution decomposed by **Split** (Training vs. Testing) and **Intersectional Attributes** (Race $\times$ Gender $\times$ Diagnosis).

Algorithm 1 Fairness-Aware Relational Vision-Language Pretraining (FRP)

Require: Mini-batch $\mathcal{B} = \{(\mathbf{x}_{Ii}, \mathbf{x}_{Ti}, y_i, \mathbf{s}_i)\}_{i=1}^{B}$, Pre-trained Encoders f_I, f_T, Hyperparams λ, c
Ensure: Optimized model parameters θ

Stage 1: Graph Construction & Initialization

$\mathbf{z}_{Ii}, \mathbf{z}_{Ti} \leftarrow f_I(\mathbf{x}_{Ii}), f_T(\mathbf{x}_{Ti})$
$\mathbf{z}_i \leftarrow \text{Concat}(\mathbf{z}_{Ii}, \mathbf{z}_{Ti})$
$\mathbf{z}_i^{\mathbb{H}} \leftarrow \exp_0^c(\mathbf{z}_i)$ ▷ Project to Poincaré Ball
// Compute Unified Adjacency Matrix A
$D_{ij} \leftarrow d_c(\mathbf{z}_i^{\mathbb{H}}, \mathbf{z}_j^{\mathbb{H}})$
$\mathbf{A}_{\text{base}} \leftarrow \text{Softmax}(-D_{ij})$ ▷ Structural Prior
$w_{ij}^{\text{fair}} \leftarrow \sigma(\text{MLP}([\mathbf{s}_i; \mathbf{s}_j]))$ ▷ Fairness Weights
$\mathbf{A} \leftarrow (1-\mu)\mathbf{A}_{\text{base}} + \mu \cdot (\mathbf{A}_{\text{base}} \odot \mathbf{W}^{\text{fair}})$

Stage 2: Hyperbolic Message Passing (HGL)

$\mathbf{Z}^{(0)\mathbb{H}} \leftarrow [\mathbf{z}_1^{\mathbb{H}}, \ldots, \mathbf{z}_B^{\mathbb{H}}]^\top$
for $l = 1$ to L **do**
  $\mathbf{Z}_{\text{tan}} \leftarrow \log_0^c(\mathbf{Z}^{(l-1)\mathbb{H}})$
  $\mathbf{H}_{\text{agg}} \leftarrow \mathbf{A} \cdot \mathbf{Z}_{\text{tan}}$
  $\mathbf{H}_{\text{trans}} \leftarrow \text{Dropout}(\text{LN}(\text{Linear}(\mathbf{H}_{\text{agg}})))$
  $\mathbf{Z}^{(l)\mathbb{H}} \leftarrow \mathbf{Z}^{(l-1)\mathbb{H}} \oplus_c \exp_0^c(\mathbf{H}_{\text{trans}})$ ▷ Möbius Addition
end for
$\mathbf{Z}_{\text{final}} \leftarrow \log_0^c(\mathbf{Z}^{(L)\mathbb{H}})$

Stage 3: Prompt Synthesis & Optimization

$\mathbf{P}_{\text{img}} \leftarrow \text{Project}(\text{Mean}(\mathbf{z}_I))$; $\mathbf{P}_{\text{text}} \leftarrow \text{Project}(\text{Fusion}(\mathbf{Z}_{\text{final}}))$
$\mathbf{C}_{\text{final}} \leftarrow \mathbf{C}_{\text{base}} + (\mathbf{P}_{\text{img}} + \mathbf{P}_{\text{text}})/2$
Compute $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \mathcal{L}_{\text{fair}}$
Update θ via gradient descent

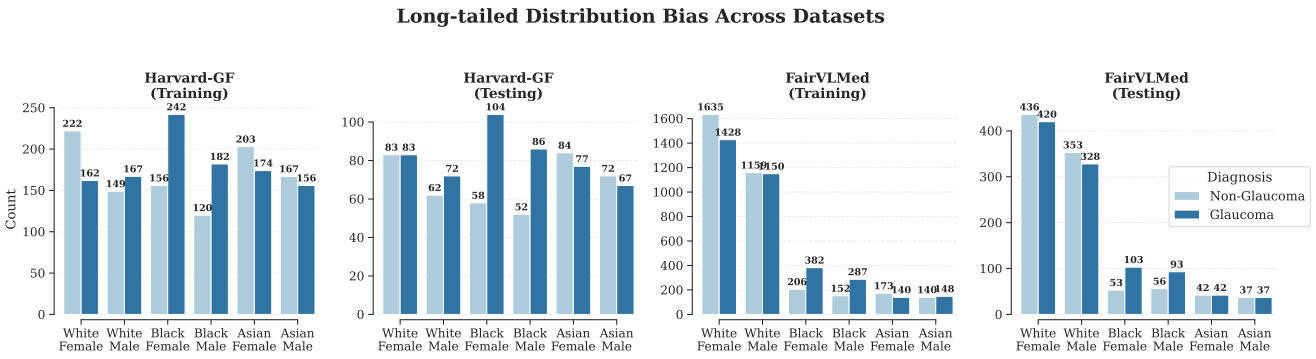


Figure 1. **Distributional Analysis across Datasets and Splits.** The visualization highlights two distinct challenges: (1) The extreme long-tailed distribution in FairVLMed, and (2) The persistent conditional label skewness in both datasets (e.g., varying Positive/Negative ratios across race-gender groups). The consistency between training and testing splits ensures rigorous evaluation.

Observation 1: Systemic Distributional Consistency. As shown in Fig. 1, the testing splits largely preserve the distributional characteristics of the training sets. This consistency ensures that our evaluation metrics reflect model performance under realistic and challenging conditions, rather than on artificially balanced subsets.

Observation 2: Distinct Bias Patterns. The two datasets exhibit different bias patterns that highlight the challenges of fairness evaluation. In **FairVLMed**, the data distribution is highly long-tailed, with extreme sparsity in certain minority groups (e.g., Asian Male). Moreover, the label distributions differ substantially across subgroups: for example, *Black Females* in the training set show a higher prevalence of glaucoma (Positive $>$ Negative), whereas *White Females* exhibit the opposite trend. Such discrepancies may encourage models to rely on demographic shortcuts during learning [**?**]. In contrast, **Harvard-GF** presents relatively balanced sample counts across groups, but still exhibits notable conditional label

skewness. Specifically, the conditional probability $P(Y|S)$ varies across demographic groups (e.g., Black Male prevalence > White Female prevalence), indicating that bias may arise from learned spurious correlations rather than purely from sample imbalance [8].

4.2. Implementation Details and Evaluation Metrics

Implementation details. We implemented the proposed FRP framework using PyTorch, with all models trained on a single NVIDIA A100 GPU. Consistent with the main paper, we employ the pre-trained CLIP (ViT-B/16) [11] as the frozen backbone. Input images are preprocessed by resizing to 224×224, normalizing with standard ImageNet statistics, and applying data augmentation techniques including random resized cropping and horizontal flipping. For optimization, we utilize the SGD optimizer with a batch size of 32. The learning rate is managed by a cosine decay scheduler set to a maximum of 0.002, preceded by a one-epoch linear warmup starting from 1×10^{-5}. The total training duration is set to 50 epochs. To ensure the reliability of our results, all reported metrics represent the average of three independent runs with different random seeds.

Model-specific configuration. For the FRP architecture, we initialize the learnable prompt tokens with a context length of $N = 32$. The proposed HGL is configured with a depth of $L = 2$. Regarding the geometric hyperparameters, the curvature c of the Poincaré ball is fixed at 1.0, and the fairness trade-off coefficient λ is set to 0.1.

Performance metrics. We assess the diagnostic utility using the area under the ROC curve (AUC). To analyze fine-grained performance, we also report the group-wise AUC, calculated independently for each demographic subgroup. Furthermore, to capture the trade-off between utility and equity, we employ the equity-scaled AUC (ES-AUC), defined as:

$$\text{ES-AUC} = \frac{\text{AUC}}{1 + \sum_{a \in \mathcal{A}} |\text{AUC} - \text{AUC}_a|}. \tag{12}$$

Fairness metrics. We quantify demographic disparities using two standard intersectional fairness metrics [1, 2]. First, the demographic parity difference (DPD) measures the maximum divergence in positive prediction rates between any two subgroups $a, b \in \mathcal{A}$:

$$\text{DPD} = \max_{a,b \in \mathcal{A}} \left| \mathbb{P}(\hat{y} = 1 \mid a) - \mathbb{P}(\hat{y} = 1 \mid b) \right|. \tag{13}$$

Second, the equalized odds difference (DEOdds) provides a stricter measure by accounting for disparities in both true positive rates (TPR) and false positive rates (FPR):

$$\begin{aligned} \text{DEOdds} = \max_{a,b \in \mathcal{A}} \Big(& |P(\hat{y} = 1 \mid y = 1, a) - P(\hat{y} = 1 \mid y = 1, b)| \\ & + |P(\hat{y} = 1 \mid y = 0, a) - P(\hat{y} = 1 \mid y = 0, b)| \Big). \end{aligned} \tag{14}$$

For interpretation, higher values indicate better performance for AUC-based metrics (↑), while lower values indicate greater fairness for DPD and DEOdds (↓).

4.3. Additional Experiments: Backbone Scalability

To validate that the efficacy of our method is not limited to specific architectures, we extended our experiments to the larger **ViT-L/14** backbone. We compare FRP against representative baselines: vanilla CLIP, FairCLIP, and CoOp. As shown in Tables 1 and 2, scaling up the backbone generally improves diagnostic performance (AUC). However, our experiments reveal that standard prompt learning methods like CoOp still exhibit significant fairness gaps (e.g., DEOdds of 7.31% on FairVLMed Gender), which FRP effectively reduces.

4.4. Extended Ablation Studies

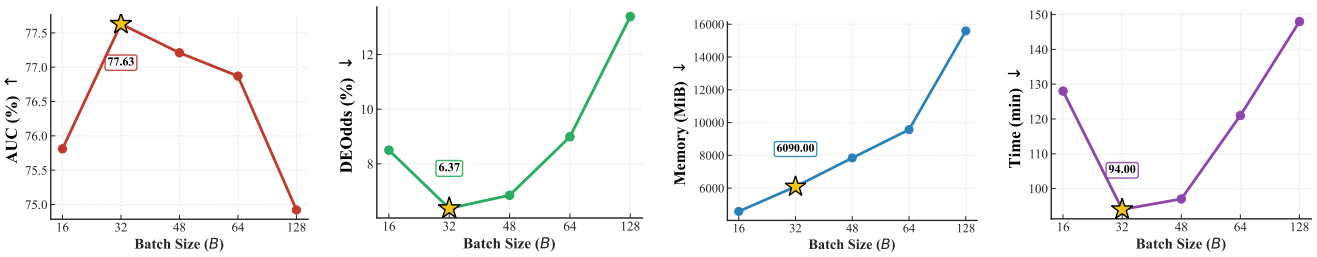
Impact of Batch Size. Since our FRP framework relies on constructing a dynamic relational graph within each mini-batch to approximate the global demographic manifold, the batch size B serves as a critical hyperparameter governing the density of information propagation. We conduct a comprehensive ablation study varying $B \in \{16, 32, 48, 64, 128\}$, with results summarized in Figure 2. We observe that in the small batch regime ($B = 16$), the relational graph becomes overly sparse. As shown in Figure 2(a)-(b), this sparsity leads to insufficient neighbor aggregation, preventing the Fairness Modulator from

Table 1. **Scalability Analysis on FairVLMed [?] (ViT-L/14).** Performance comparison using the larger ViT-L/14 backbone. All experiments are repeated three times and mean ± std are reported (%). The best result is in **bold**, and the second-best is underlined.

Attribute	Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
						Asian	Black	White
Race	CLIP [11]	10.10 ± 9.44	10.79 ± 10.41	67.83 ± 2.92	63.53 ± 1.83	70.65 ± 4.58	70.12 ± 3.39	66.22 ± 2.97
	FairCLIP [?]	17.79 ± 4.86	18.30 ± 2.07	69.88 ± 2.00	66.54 ± 1.73	71.78 ± 2.18	71.79 ± 2.13	68.67 ± 1.99
	CoOp [15]	4.95 ± 1.24	6.53 ± 2.10	76.97 ± 0.87	73.42 ± 1.56	77.48 ± 2.47	73.32 ± 1.88	77.64 ± 0.72
	FRP (Ours)	**3.82** ± 0.64	**5.29** ± 0.88	**79.55** ± 0.47	**76.73** ± 0.58	**82.15** ± 1.02	**79.68** ± 0.73	**78.84** ± 0.39
						Female	**Male**	
Gender	CLIP [11]	2.93 ± 3.17	4.29 ± 4.05	67.83 ± 2.92	63.86 ± 2.36	65.13 ± 2.60	71.31 ± 3.24	
	FairCLIP [?]	5.82 ± 1.22	8.14 ± 2.62	69.74 ± 0.95	66.00 ± 1.55	67.29 ± 1.38	72.99 ± 0.83	
	CoOp [15]	2.96 ± 0.95	7.31 ± 1.34	76.97 ± 0.87	72.63 ± 1.12	74.26 ± 1.65	80.24 ± 0.78	
	FRP (Ours)	**0.41** ± 0.17	**1.88** ± 0.52	**79.55** ± 0.47	**78.71** ± 0.53	**79.72** ± 0.61	**80.45** ± 0.68	

Table 2. **Scalability Analysis on Harvard-GF [8] (ViT-L/14).** Performance comparison using the larger ViT-L/14 backbone. All experiments are repeated three times and mean ± std are reported (%). The best result is in **bold**, and the second-best is underlined.

Attribute	Model	DPD ↓	DEOdds ↓	AUC ↑	ES-AUC ↑	Group-wise AUC ↑		
						Asian	Black	White
Race	CLIP [11]	9.23 ± 1.35	15.41 ± 2.35	80.87 ± 0.82	76.15 ± 1.24	82.95 ± 1.68	76.34 ± 2.95	81.88 ± 0.76
	CoOp [15]	6.00 ± 1.45	7.27 ± 1.92	83.48 ± 0.68	78.12 ± 1.04	85.59 ± 2.15	79.98 ± 1.76	84.72 ± 0.88
	FRP (Ours)	**2.18** ± 0.54	**5.85** ± 0.72	**85.42** ± 1.54	**82.04** ± 0.81	**86.16** ± 1.12	**83.10** ± 0.48	**84.37** ± 0.55
						Female	**Male**	
Gender	CLIP [11]	7.62 ± 1.58	11.05 ± 2.72	80.87 ± 0.82	78.94 ± 1.15	79.85 ± 1.32	81.92 ± 0.88	
	CoOp [15]	6.57 ± 1.33	7.33 ± 1.50	83.48 ± 0.68	82.90 ± 0.92	83.27 ± 1.18	83.96 ± 0.75	
	FRP (Ours)	**1.98** ± 0.42	**5.25** ± 0.63	**85.42** ± 1.54	**84.61** ± 0.75	**85.21** ± 0.52	**86.07** ± 0.63	



(a) Utility (AUC) (b) Fairness (DEOdds) (c) Memory Usage (d) Training Time

Figure 2. **Impact of Batch Size (B).** We evaluate the trade-off between utility (AUC), fairness (DEOdds), and computational efficiency (Memory & Time). $B = 32$ (highlighted with ⋆) emerges as the optimal operating point, offering the best balance of high AUC, low intersectional disparity, and minimal training cost. Note that for DEOdds, Memory, and Time, lower values are better (↓).

effectively capturing the local manifold structure, causing diagnostic performance to drop to 75.81% AUC and fairness to significantly degrade (8.50% DEOdds). Conversely, in the large batch regime ($B \geq 64$), contrary to the intuition that larger batches provide better global context, we observe a performance degradation at $B = 128$ (AUC drops to 74.92% and DEOdds spikes to 13.39%). We attribute this to the over-smoothing phenomenon in graph neural networks, where forcing aggregation over a dense graph in hyperbolic space introduces noisy connections between semantically distant nodes, diluting the fine-grained intersectional signals required for fairness. Furthermore, larger batches incur quadratic memory costs ($O(B^2)$) and

longer convergence times. Ultimately, we identify $B = 32$ as the optimal operating point, which strikes the ideal balance by providing sufficient demographic diversity to construct a meaningful A_{fair} while maintaining the structural integrity of the hyperbolic hierarchy. This configuration achieves the highest peak performance (77.63% AUC) and the most equitable outcomes (6.37% DEOdds) while minimizing training time (94 min), and is thus adopted for all main experiments.

4.5. Computational Overhead Analysis

We analyze the computational footprint of our framework in Table 3, comparing GPU memory usage, prompt context length (N_{ctx}), and total convergence time. All measurements were conducted on a single NVIDIA A100 (80GB) GPU.

Table 3. **Computational Overhead Analysis.** We compare GPU memory, N_{ctx}, and the total time required for convergence. While FRP incurs higher training costs due to dynamic graph construction, it maintains **identical inference speed** to the baseline, ensuring deployment efficiency. ↓: Lower is better.

Model	Batch Size	N_{ctx}	Epochs	Memory (MiB) ↓	Train Time (min) ↓
CoOp [15]	32	16	50	4,034	**15.5**
CoCoOp [14]	2	16	10	1,725	62.7
VPT [5]	4	2	10	1,528	<u>18.6</u>
MaPLe [6]	4	2	10	1,183	22.7
FRP (Ours)	32	32	50	6,090	92.5

Resource Efficiency and Deployment Feasibility. As detailed in Table 3, the proposed FRP framework incurs a moderate computational overhead during the training phase, with GPU memory usage increasing to 6,090 MiB and convergence time extending to approximately 92.5 minutes. This increase is attributable to the complexity of dynamic relational graph construction ($O(B^2)$) and the associated Riemannian optimization within the hyperbolic space. However, this overhead presents no barrier to practical deployment. First, the absolute memory footprint remains exceptionally low, fitting comfortably within standard consumer-grade hardware (e.g., RTX 2080) rather than requiring specialized data-center GPUs. Second, and most critically, the computational complexity is strictly confined to the offline training stage. During inference, the relational graph module is detached, allowing FRP to achieve an inference latency identical to the simple CoOp baseline [15]. This decoupling ensures that while the model benefits from complex structural reasoning during learning, it introduces zero additional latency for real-time clinical workflows.

5. Limitations and Future Work

While FRP demonstrates a strong accuracy–fairness trade-off on *Race* and *Gender*, our analysis reveals challenges when scaling to higher-order intersectional spaces and specific architectural constraints.

5.1. Challenges in Data Fragmentation and Batch Dependencies

High-Order Intersectional Fragmentation. The primary limitation arises from the combinatorial nature of expanding the sensitive attribute set. While FRP demonstrates robust performance on dual-attribute intersections (*Race* & *Gender*), scaling to higher-order demographics (e.g., *Race* × *Gender* × *Language* × *Ethnicity*) presents a **data fragmentation challenge**. As illustrated in Figure 3, increasing the attribute order partitions the dataset into exponentially more fine-grained subgroups. This results in **significantly diminished sample density** within specific intersectional cells, which inflates the statistical variance of the relational graph neighbors. The performance degradation observed in high-order settings is thus primarily driven by the **low signal-to-noise ratio** in these sparser connections, which hinders the stability of the fairness message passing, rather than a fundamental lack of geometric capacity in the hyperbolic embedding itself.

Dependency on Batch Diversity. Furthermore, our framework constructs a dynamic relational graph within each mini-batch to approximate the global demographic manifold. A theoretical limitation arises in the extreme case of a **homogeneous mini-batch**, where all sampled instances belong to the same demographic or diagnostic group. In such degenerate scenarios, the fairness modulator (A_{fair}) cannot compute valid pairwise disparities to guide the optimization, effectively reducing the fairness regularization for that specific iteration. While our use of random shuffling makes the probability of such occurrences

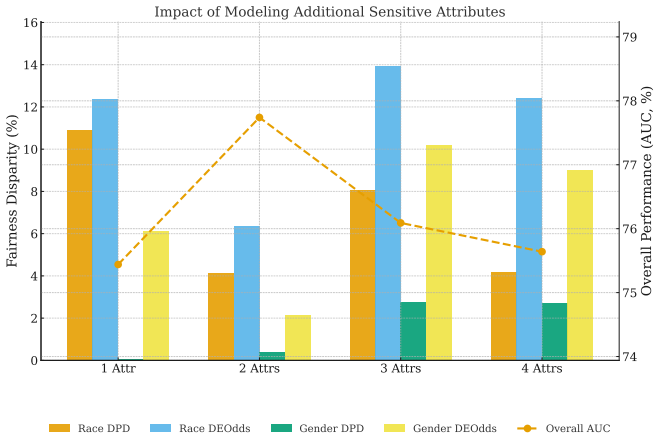


Figure 3. **Effect of Expanding the Sensitive-Attribute Set.** Bars: fairness gaps (lower is better); dashed line: overall AUC (higher is better). Moving from two attributes to four (adding Language, Ethnicity) increases disparities and reduces utility, highlighting the challenge of high-order intersectionality.

statistically negligible, this dependency highlights a **"local-to-global" context gap** inherent to batch-wise graph learning, which we aim to address via global prototypes in future work.

5.2. Future Directions

To overcome these hurdles, future work will focus on three complementary strategies. First, to stabilize optimization in high-dimensional spaces, we plan to implement curriculum optimization, introducing attributes progressively to guide the model through a smoother learning trajectory. Second, to resolve geometric bottlenecks, we aim to introduce adaptive geometric capacity by incorporating learnable curvature and deeper layers, allowing the embedding space to dynamically adjust to complex hierarchies. Finally, we will address batch limitations by developing global prototype graphs. By maintaining a memory bank of subgroup prototypes rather than relying solely on batch statistics, we can ensure robust relational modeling even for sparse samples, paving the way for FRP to scale to richer, clinically meaningful environments.

References

[1] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: an extensible toolkit for detecting. *Understanding, and Mitigating Unwanted Algorithmic Bias*, 2, 2018. 5

[2] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. *Microsoft, Tech. Rep. MSR-TR-2020-32*, 2020. 5

[3] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[4] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, pages 534–543, 2018. 2, 3

[5] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *Computer Vision – ECCV 2022*, pages 709–727, Cham, 2022. Springer Nature Switzerland. 7

[6] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 7

[7] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguñá. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010. 1

[8] Yan Luo, Yu Tian, Min Shi, Louis R Pasquale, Lucy Q Shen, Nazlee Zebardast, Tobias Elze, and Mengyu Wang. Harvard glaucoma fairness: a retinal nerve disease dataset for fairness learning and fair identity normalization. *IEEE Transactions on Medical Imaging*, 2024. 3, 5, 6

[9] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, pages 6341–6350. Curran Associates, Inc., 2017. 1

[10] Hongbin Pei, Bingzhe Wei, Kevin Chen-Chuan Chang, Yu Lei, and Bo Yang. Geom-gcn: Geometric graph convolutional networks. In *International Conference on Learning Representations*, 2020. 2

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

[12] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4460–4469. PMLR, 2018. 2

[13] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pages 355–366. Springer, 2011. 2

[14] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4992–5001, 2022. 7

[15] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 6, 7